Dourdan, France
Du 6 au 10 juin 2005

Tome 1

Conférences principales

Sous l'égide de

Association pour le Traitement Automatique des Langues (ATALA)

LIMSI-CNRS

Synapse Développement

CEA

UFR d'Orsay

Agence Universitaire de la Francophonie

France Télécom R & D

SINEQUA

Ministère délégué à la Recherche

IIE

ISBN 2-9524255-0-7

ISBN 2-9524255-0-7

EAN 9782952425506

Préambule

Cette année, la conférence TALN s'est installée au VVF de Dourdan dans la périphérie lointaine de Paris. Pour ne pas être tentés par les sirènes parisiennes, nous avons choisi une immersion totale dans un lieu convivial proche de la forêt de Rambouillet. Pour la petite histoire, le site que nous avions proposé initialement et qui avait été accepté par l'ATALA était celui de L'Ecole des Mines de Paris à Fontainebleau. Ce choix n'a pu être maintenu pour des raisons de logistique. Une trace est restée : les petites bouteilles qui servaient à la conservation du chasselas de Thomery et que l'ancien responsable du site de Fontainebleau, Philippe Vincent Jamet, a offert aux participants de la conférence.

C'est sous l'impulsion de Brigitte Grau qu'une partie de l'équipe LIR (Langues, Information et Représentations) du LIMSI a pris en charge l'organisation de TALN en 2005. Ce fut une aventure pleine d'imprévus, de rebondissements et de problèmes résolus au fur et à mesure, aventure très consommatrice de temps et d'énergie essentiellement humaine. Parmi les imprévus, nous pouvons citer le choix du lieu de la conférence, les discussions sur l'anonymisation des articles, les pannes de serveurs à des moments cruciaux comme le dernier jour de soumission des articles pour TALN puis pour les ateliers (merci à Olivier Galibert de nous avoir dépanner rapidement les deux fois). Pour les points durs prévisibles, l'installation par Guillaume Pitel du logiciel libre OpenConf qui a servi à gérer la réception et la relecture des articles, sa traduction (non automatique) par Anne Vilnat qui dut lutter pour l'édition des caractères accentués sous des formats très variés, sans oublier la budgétisation de la conférence dans les règles de l'administration publique (merci à Martine Charrue d'avoir réussi à dénouer des points financiers inextricables en particulier pour l'organisation de notre soirée de gala à Vaux-Le Vicomte). Comme vous avez pu le constater, notre site internet s'est enrichi au fur et à mesure de l'avancée de l'organisation grâce à Anne-Laure Ligozat qui a également participé avec Gaëlle Lortal à la conception graphique de l'affiche et à ses déclinaisons en étiquettes, couvertures des actes et du programme...

Une grande partie de nos efforts a été consacrée à la recherche de subventions et nous remercions tous ceux qui ont participé au mécénat de la conférence que ce soit par des apports financiers, de logiciels ou de livres. Du point de vue scientifique, TALN05 est un succès en terme de soumissions d'articles: 98Ąăarticles soumis. Nous remercions les relecteurs et les membres du comité de programme pour leur participation à la sélection des articles (36 présentations et 25 posters) et pour la qualité de leurs commentaires sur les articles soumis, cela d'autant plus que le nombre moyen d'articles relus par chacun s'est élevé à cinq.

La conférence RECITAL05 a, quant à elle, confirmé sa maturité et sa place dans la communauté par le nombre des soumissions en croissance par rapport aux années précédentes (35 articles) et la diversité géographique de leurs auteurs. Le choix de 11 articles en présentation orale et de 16 articles en poster a pu se faire sans accroc grâce à la participation et au sérieux des membres de son comité de programme et de son comité de lecture, que les co-présidents de RECITAL, Nicolas Hernandez et Guillaume Pitel, souhaitent ici remercier chaleureusement.

Michèle Jardino
Présidente de TALN 2005

Nicolas Hernandez & Guillaume Pitel
Présidents de RECITAL 2005

TALN 2005

Comité de programme

Salah Ait-Mokhtar	Xerox Research Centre Europe (XRCE)
Núria Bel	IULA - Universitat Pompeu Fabra
Philippe Blache	LPL
Christian Boitet	CLIPS-IMAG
Jean-Pierre Chevallet	CLIPS-IMAG
Béatrice Daille	LINA FRE CNRS 2729 - Université de Nantes
Laurence Danlos	Lattice
Olivier Ferret	CEA-LIST
Patrick Gallinari	LIP6 - UPMC
Claire Gardent	CNRS / LORIA
Brigitte Grau	LIMSI-CNRS
Michèle Jardino	LIMSI-CNRS
Daniel Kayser	Laboratoire d'Informatique de Paris-Nord, Université Paris-Nord
Philippe Langlais	RALI - Université de Montréal
Dominique Laurent	SYNAPSE Développement
Anne Nicolle	Université de Caen - GREYC - CNRS UMR 6072
Patrick Paroubek	LIMSI-CNRS
Marie-Paule Pery-Woodley	ERSS/Université de Toulouse-Le Mirail
Jean-Marie Pierrel	ATILF CNRS/Université Henri Poincaré Nancy
Martin Rajman	EPFL
Owen Rambow	Columbia University
Isabelle Robba	LIMSI-CNRS
Pascale Sébillot	IRISA
Gérard Sabah	LIMSI-CNRS
Anne Vilnat	LIMSI-CNRS
Michael Zock	LIMSI-CNRS
Pierre Zweigenbaum	AP-HP/INSERM/INaLCO

Comité de lecture

Jean-Yves Antoine	Laboratoire LI - Université Francois Rabelais de Tours
Delphine Battistelli	Université Paris-Sorbonne (Paris 4)
Patrice Bellot	LIA - Université d'Avignon / CNRS
Romaric Besançon	CEA - LIST
Pierre Beust	Université de Caen - GREYC - CNRS UMR 6072
Hervé Blanchon	CLIPS-IMAG
Malek Boualem	France Telecom - Recherche & Développement
Mohand Boughanem	IRIT
Gaël de Chalendar	CEA/LIST/LIC2M
Jean-Cédric Chappelier	EPFL
...	

TALN 2005

Comité de lecture (suite)

Laurent Charnay	France Télécom - Recherche & Développement
Christine Jacquin	LINA, université de Nantes
Stéphane Ferrari	GREYC - CNRS UMR 6072
Bertrand Gaiffe	Loria (Nancy)
Núria Gala Pavia	DELIC, Université d'Aix
Damien Genthial	Laboratoire CLIPS-IMAG, Grenoble
Kim Gerdes	ERSS, Université Bordeaux 3
Gregory Grefenstette	CEA
Emilie Guimier De Neef	France Télécom - Recherche & Développement
Caroline Hagège	Xerox Research Centre Europe (XRCE)
Nabil Hathout	ERSS, UMR5610 CNRS & Université de Toulouse Le Mirail
Agata Jackiewicz	Laboratoire LaLICC, Université de Paris IV Sorbonne
Evelyne Jacquey	ATILF-CNRS
Sylvain Kahane	Modyco, Université Paris 10
Mathieu Lafourcade	LIRMM
Guy Lapalme	RALI - Université de Montréal
Yves Lepage	ATR
Bernard Levrat	LERIA, Université d'Angers
Claude de Loupy	Sinequa & Université de Paris 10
Daniel Luzzati	LIUM
Aurélien Max	LIMSI-CNRS & Université Paris XI
Richard Moot	LaBRI
Emmanuel Morin	LINA - FRE CNRS 2729
Ghassan Mourad	LaLICC (Paris-Sorbonne)/ Université Libanaise
Adeline Nazarenko	Laboratoire d'Informatique de Paris-Nord (UMR 7030)
Guy Perrier	LORIA, Université Nancy 2
Thierry Poibeau	LIPN (CNRS et U. Paris 13)
Andrei Popescu-Belis	Université de Genève
Bruno Pouliquen	Centre Commun de Recherche de la Commission Européenne
Sophie Rosset	LIMSI-CNRS
Azim Roussanaly	LORIA / INRIA Lorraine
Patrick Ruch	Université de Genève/Hôpitaux Universitaires de Genève
Gilles Sérasset	GETA CLIPS IMAG
Susanne Salmon-Alt	ATILF-CNRS
Jacques Vergne	Université de Caen - GREYC - UMR 6072
Leo Wanner	ICREA et Université Pompeu Fabra
Francois Yvon	GET/ENST

RECITAL 2005

Comité de programme

Jean-Yves Antoine	LI - Université François Rabelais de Tours
Frédéric Bechet	LIA/CNRS - Université d'Avignon
Laurent Besacier	GEOD CRISP IMAG
Hervé Blanchon	CLIPS
Philippe Boula de Mareüil	LIMSI-CNRS
Estelle Campione	DELIC - Université de Provence
Gaël de Chalendar	CEA/LIST/LIC2M
Patrice Enjalbert	GREYC CNRS UMR 6072
Cécile Fabre	ERSS - Université Toulouse Le Mirail
Nathalie Friburger	LI - Université François Rabelais de Tours
Núria Gala Pavia	DELIC - Université de Provence
Thierry Hamon	LIPN - UMR CNRS 7030 - Université Paris Nord
Nicolas Hernandez	LIMSI-CNRS
Gabriel Illouz	LIMSI-CNRS
Philippe Langlais	RALI - Université de Montréal
Thomas Lebarbé	LIDILEM - Université Grenoble 3
Denis Maurel	Université François-Rabelais de Tours
Emmanuel Morin	LINA
Guillaume Pitel	LIMSI-CNRS / LORIA INRIA Lorraine
Laurent Romary	LORIA INRIA Lorraine
Laurent Roussarie	Université Paris 7
Susanne Salmon-Alt	ATILF-CNRS
Jean Véronis	DELIC - Université de Provence

Comité de lecture

Pierre Beust	GREYC CNRS UMR 6072
Jean-Cédric Chappelier	EPFL
Elisabeth Godbert	LIF - Université de la Méditerranée
Guy Perrier	LORIA INRIA Lorraine - Université Nancy 2
Romain Prudon	LIMSI-CNRS
Agata Savary	Université de Tours
Ludovic Tanguy	ERSS - Université de Toulouse le Mirail

TALN 2005 - RECITAL 2005

Comité d'organisation commun

Martine Charrue	LIMSI-CNRS
Brigitte Grau	LIMSI-CNRS / IIE
Nicolas Hernandez	LIMSI-CNRS / IIE
Gabriel Illouz	LIMSI-CNRS / Université Paris Sud
Michèle Jardino	LIMSI-CNRS
Anne-Laure Ligozat	LIMSI-CNRS
Gaëlle Lortal	Université Technologique de Troyes
Sophie Pageau-Maurice	LIMSI-CNRS
Patrick Paroubek	LIMSI-CNRS
Guillaume Pitel	LIMSI-CNRS / LORIA
Isabelle Robba	LIMSI-CNRS / IUT Vélizy
Anne Vilnat	LIMSI-CNRS / Université Paris Sud
Michaël Zock	LIMSI-CNRS

Déroulement de la conférence

Le **lundi 6 juin** est consacré à deux tutoriels :

- *Approches quantitatives des corpus de textes*, par André Salem, de l'*Université de la Sorbonne Nouvelle (Paris 3)* et Ludovic Lebart de l'*ENST* ;
- *Meta-données et ressources linguistiques*, par Laurent Romary du *LORIA*.

Le **mardi 7 juin**, après la conférence invitée *Formal Ontology and Natural Language Semantics* donnée par Nicola Guarino de l'*Istituto di Scienze e Tecnologie della Cognizione*, se sont déroulées les sessions de TALN :

- Grammaires ,
- Recherche d'information ,
- Sémantique et terminologie ,
- Analyse de phrase ,
- Analyse lexicale ,
- Représentations sémantiques ,

suivies de la session consacrée aux posters de TALN.

Le **mercredi 8 juin** se sont déroulées les présentations orales de RECITAL, suivies de l'atelier sur les évaluations EQueR et EASy, ainsi que des posters RECITAL.

Le **jeudi 9 juin**, après la conférence invitée *Opinion and Argument Extraction from Text* donnée par Eduard Hovy de l'*Information Sciences Institute of the University of Southern California*, se sont déroulées les sessions de TALN :

- Texte ;
- Traduction ;
- Dialogue ;
- Sémantique et corpus ;
- Grammaires ;
- Apprentissage.

Le **vendredi 10 juin** ont eu lieu les ateliers sur :

- Langues peu dotées ;
- Langues des signes ;
- Défi Fouille de textes ;

Sommaire général des actes de TALN et RECITAL 2005

Tome 1

Tome 2

Sommaire

Conférence principale TALN

Posters TALN

Conférence principale RECITAL

Posters RECITAL

TALN 2005

12ème conférence annuelle
sur le
Traitement Automatique des Langues Naturelles

CONFÉRENCE PRINCIPALE

Comment obtenir plus des Méta-Grammaires

François Thomasset, Eric Villemonte de la Clergerie
ATOLL - INRIA
Domaine de Voluceau
Rocquencourt, B.P. 105, 78153 Le Chesnay (France)
{Francois.Thomasset,Eric.De_La_Clergerie}@inria.fr

Mots-clefs : Méta-grammaires, Analyse Syntaxique, TAG, TIG

Keywords: Meta-grammars, Parsing, TAG, TIG

Résumé Cet article présente un environnement de développement pour les méta-grammaires (MG), utilisé pour concevoir rapidement une grammaire d'arbres adjoints (TAG) du français à large couverture et néanmoins très compacte, grâce à des factorisations d'arbres. Exploitant les fonctionnalités fournies par le système DYALOG, cette grammaire a permis de construire un analyseur syntaxique hybride TAG/TIG utilisé dans le cadre de la campagne d'évaluation syntaxique EASY.

Abstract This paper presents a development environment for Meta-Grammars (MG), used to design, in a short period, a wide coverage but still very compact Tree Adjoining Grammar (TAG) for French, thanks to tree factorizations. Exploiting the functionalities provided by DYALOG system, an hybrid TAG/TIG parser was compiled from the grammar and used for the EASY parsing evaluation campaign.

1 Introduction

Les méta-grammaires (MG) (Candito, 1999) renouvellent les méthodes de conception des grammaires, en introduisant un niveau plus abstrait de description à l'aide de contraintes élémentaires, regroupées en classes relativement simples, elles-mêmes insérées dans une hiérarchie multiple d'héritage. Une phase de compilation permet ensuite de croiser ces classes et d'utiliser les contraintes pour dériver des structures grammaticales pour un formalisme cible comme les grammaires d'arbres adjoints (TAG) ou les grammaires fonctionnelles lexicales (LFG) (Gaiffe *et al.*, 2003; Clément & Kinyon, 2003). Les descriptions deviennent plus modulaires et permettent la factorisation d'ensembles de contraintes communs à plusieurs phénomènes syntaxiques (comme des règles d'accord). L'héritage permet d'affiner progressivement la description d'un phénomène, par exemple pour la structure verbale. Il rend aussi raisonnable l'espoir qu'une partie de l'organisation en classes ainsi qu'une partie du contenu des classes puissent être conservées d'une langue à une autre et d'un formalisme cible à un autre.

Ces raisons nous ont conduit à choisir les méta-grammaires pour concevoir rapidement un analyseur syntaxique hybride TAG/TIG du français à large couverture, analyseur qui a finalement pu être déployé dans le cadre de la campagne EASY d'évaluation d'analyseurs syntaxiques. Néanmoins, nos premières tentatives ont montré certaines limites dans les capacités descriptives des MG mais ont également suggéré des possibilités pour obtenir à peu de frais des grammaires TAG beaucoup plus compactes. En effet, il est bien connu que les grammaires TAG à large couverture ont tendance à exploser en nombre d'arbres, avec plusieurs milliers ou dizaines de milliers de (schémas d') arbres (Abeillé, 2002), ce qui rend très difficile l'analyse, même en exploitant des techniques de filtrage par les mots de la chaîne d'entrée. Les alternatives proposées passent par des techniques d'analyse des arbres pour retrouver et factoriser leurs parties communes (Carroll *et al.*, 1998) ou par la description de schémas de parcours multiples dans les arbres (Harbusch & Woch, 2004). Les méta-grammaires s'appuyant sur des descriptions factorisées nous permettent d'aller plus facilement dans la direction de tels arbres *factorisés*.

Le système DYALOG que nous utilisons pour la construction d'analyseurs syntaxiques peut gérer de tels arbres factorisés (Section 2). En conséquence, en parallèle avec la conception d'une méta-grammaire du français, nous avons étendu les possibilités descriptives des MG et les possibilités génératives de notre compilateur de MG (Section 3). Nous avons également complété notre environnement de travail pour les MG. La section 4 fournit quelques éléments d'information sur notre méta-grammaire et sur la grammaire résultante, en particulier au niveau de la compacité. Enfin, la section 5 fournit quelques résultats préliminaires pour notre analyseur.

2 Analyseurs hybrides TAG/TIG avec le système DYALOG

Le système DYALOG (Villemonte de la Clergerie, 2002) fournit un environnement de compilation et d'exécution d'analyseurs syntaxiques tabulaires (à la Earley) offrant la puissance d'un langage de programmation en logique. Il couvre divers formalismes syntaxiques, dont ceux utilisés dans notre expérience, à savoir les Grammaires d'Arbres Adjoints (TAG) et les Grammaires d'Insertion d'Arbres (TIG).

Les TAG (Joshi, 1987) sont formées d'arbres partiels d'analyse combinables par substitution et adjonction. Un nœud feuille étiqueté par un non-terminal peut être substitué par un arbre initial. Une adjonction insère le contenu d'un arbre auxiliaire β au niveau d'un nœud N, le sous-arbre

de racine N étant rattaché au niveau du *pied* f_β de β. Dans les FTAG, les nœuds sont décorés par une paire d'attributs `top` et `bot`, généralement exprimés comme des structures de traits.

Les TIG (Schabes & Waters, 1995) sont une variante des TAG restreignant les arbres auxiliaires de sorte qu'ils ne puissent s'insérer qu'à droite ou à gauche du nœud d'adjonction. Cette condition implique en particulier que les arbres auxiliaires aient leur *dorsale* (c.a.d. le chemin de la racine au pied) comme frontière gauche ou droite. L'intérêt majeur des TIG provient du fait qu'elles sont analysables, comme les CFG, avec une complexité en $O(n^3)$ alors que les TAG le sont en $O(n^6)$ où n dénote la longueur de la chaîne d'entrée. De plus, la plupart des grammaires TAG sont essentiellement TIG et il est en fait possible de construire des analyseurs syntaxiques hybrides TAG/TIG (Alonso & Díaz, 2003). DYALOG peut analyser une grammaire TAG pour identifier les parties TIG afin de construire de tels analyseurs hybrides TAG/TIG[1].

Pour les différents formalismes syntaxiques qu'il couvre, le système DYALOG permet, à l'intérieur des structures grammaticales, l'usage d'*opérateurs réguliers* tels que la disjonction, l'étoile de Kleene et l'entrelacement, ce dernier permettant d'indiquer un ordre libre entre des séquences de constituants (Nederhof *et al.*, 2003). Ces opérateurs ne changent pas le formalisme sous-jacent car ils peuvent en théorie être expansés et éliminés en introduisant de nouvelles structures grammaticales (arbres ou productions) et/ou de nouveaux non-terminaux. Néanmoins, le taux d'expansion peut être exponentiel en le nombre d'occurrences de ces opérateurs. Leur utilisation permet donc d'obtenir des grammaires beaucoup plus compactes et plus efficaces, car ces opérateurs sont utilisés sans expansion. D'autre part, il est à noter que l'usage de ces opérateurs rend plus naturel les forêts de dérivations en évitant l'usage de non-terminaux artificiels.

3 Étendre les MetaGrammaires

```
1 class collect_real_subject_canonical {
2     <: collect_real_subject;
3     $arg.extracted = value(~cleft);
4     S >> VSubj;  VSubj < V;   V >> postsubj;  VMod < postsubj;
5     node postsubj: [cat:N2,id:subject, type:subst, top:[wh:-,sat:+]];
6     - postsubj::agreement; postsubj = postsubj::N;
7     postsubj =>
8       node(Infl).bot.inv = value(+),
9       $arg.extracted = value(-), $arg.real = value(N2),
10      desc.extraction = value(~-),
11      node(V).top.mode=value(~infinitive|imperative|gerundive|participle);
12    ~ postsubj => node(Infl).bot.inv = value(~+);
13 }
```

Listing 1 – Exemple de classe

Le listing 1 illustre une classe fille `collect_real_subject_canonical` héritant de la classe parente `collect_real_subject`. Cette dernière décrit l'ensemble des réalisations possibles du sujet et est utilisée comme modèle pour les diverses réalisations du sujet en position canonique ou en extraction clivée[2]. La classe fille complète la classe parente pour le cas cano-

[1]Il est à noter que cette analyse ne garantit pas toujours l'équivalence entre analyseurs TAG et analyseurs hybrides TAG/TIG suite aux décorations et à des gestions différentes de l'adjonction, à savoir adjonction «chaînée» (sur les racines des arbres auxiliaires) pour les TAG contre adjonction multiple pour les TIG.

[2]Type « C'est de travailler qui me fatigue ! ».

nique, en précisant la position du sujet (sous `S` et devant le noyau verbal `V`) et en introduisant la notion de sujet post-verbal uniquement réalisable par un groupe nominal (`N2`).

Plus formellement, les méta-grammaires permettent une description syntaxique éclatée à l'aide de contraintes élémentaires regroupées en classes. Une classe peut hériter des contraintes de plusieurs classes parentes (<:, ligne 2) et peut également fournir une ressource ($+r$) ou requérir une ressource ($-r$, l. 6).

Les contraintes peuvent porter sur les nœuds (l. 4 et 6) incluant l'égalité =, la précédence < ainsi que les dominances immédiates >> et indirectes >>+. Les contraintes peuvent aussi porter sur les décorations des nœuds (l. 5) ou de la classe elle-même (**desc**, l. 10). Les décorations sont exprimées comme des structures de traits (l. 5) avec possibilité d'utiliser des disjonctions | et négations ~ sur des valeurs atomiques (l. 11) ainsi que des variables ($arg). Les contraintes sur les décorations s'expriment soit directement soit au travers d'équations entre chemins de traits ancrés sur des nœuds (l. 8), sur la classe elle-même (**desc**, ligne 10) ou sur des variables (l. 9). Des macros peuvent être utilisées pour nommer des valeurs ou des chemins. Enfin, il est possible de faire porter des contraintes sur le père d'un nœud N avec la notation « father (N)».

L'objectif du compilateur de méta-grammaire[3] est alors de croiser, par point fixe, les classes terminales (c.a.d. sans descendants) de manière à obtenir des classes *neutres* pour lesquelles chaque ressource fournie est consommée et réciproquement. Les contraintes sont accumulées lors des croisements et seules sont conservées les classes dont les contraintes accumulées, prenant en compte leurs conséquences logiques, sont satisfiables[4] . Les contraintes des classes neutres survivantes sont ensuite exploitées pour produire les structures grammaticales minimales, en l'occurrence des arbres pour les TAG.

Dans la formalisation standard des MG (Candito, 1999), une ressource peut être neutralisée au plus une fois pour produire une classe neutre. Cette restriction amène à dupliquer certaines classes pour nommer différemment la même ressource. Ainsi, pour exprimer qu'une classe décrivant les verbes a besoin de 2 arguments verbaux, il faut dupliquer une partie importante de la hiérarchie des classes pour deux ressources similaires `-varg1` et `-varg2`. Pour lever cette limitation, nous avons introduit la notion d'espace de noms et rompu la symétrie entre fournisseurs et consommateurs : une ressource peut maintenant être demandée dans un certain espace de nom ns (ns = `postsub` dans −postsubj::agreement, l. 6) et lors d'un croisement avec une classe fournisseuse $\mathcal{C}$ (ici, fournissant +agreement), les nœuds, variables et besoins de $\mathcal{C}$ sont alors plongés dans l'espace de nom ns (ici postsubj). Les espaces de noms permettent un usage beaucoup plus intensif du mécanisme de ressources et une bien meilleure factorisation des méta-grammaires. Les MG sont alors moins redondantes et plus faciles à maintenir.

Les décorations portées par les nœuds et la classe sont libres mais certaines ont néanmoins un statut spécial par rapport à la génération des arbres TAG. Pour les nœuds, on peut citer les traits `cat` pour la catégorie syntaxique, `type` pour le type de nœud, `lex` pour une valeur lexicale, `adj` pour indiquer le statut du nœud pour l'adjonction, `top` et `bot` comme arguments. Pour les classes, le trait `ht` indique l'hypertag qui sera associé aux arbres pour permettre l'ancrage avec les entrées lexicales (voir Section 4).

La possibilité d'engendrer des arbres *factorisés* résulte de divers mécanismes. En premier lieu, à côté des types standards de nœuds, il existe les types spéciaux `alternative` et `sequence`. Le trait `optional` permet de rendre optionnel un nœud tandis que le trait `star` permet de

[3]Développé sous le système DYALOG.

[4]Par exemple, le compilateur vérifie qu'un noeud ne précède pas son père.

rendre un nœud répétable, correspondant à une étoile de Kleene[5]. Enfin, lors de l'énumération des arbres minimaux vérifiant un ensemble de contraintes, le compilateur utilise l'opérateur d'*entrelacement* (##) pour rendre compte de sous-spécification de précédence entre nœuds frères. Ainsi, les contraintes «N >> N_1; N>> N_2; N >> N_3; N_1 < N_2» produisent le fragment d'arbre $N((N_1, N_2)\#\#N_3)$ indiquant que N_3 se positionne librement (avant, au milieu, après) par rapport à la séquence N_1, N_2. Pour favoriser l'obtention d'arbres TIG, le compilateur évite, dans la mesure du possible, d'utiliser l'opérateur d'entrelacement quand il couvre un nœud pied comme dans $R_\beta(N\#\#F_\beta)$. Dans ce cas, les différentes possibilités d'ordonnancement des nœuds sont examinées pour produire des arbres que l'on espère être TIG. Il est également possible d'assigner un rang à un nœud avec le trait `rank` et les valeurs `first` et `last`.

L'optionnalité fournie par l'emploi du trait `optional` n'est pas assez fine en pratique. L'emploi de *gardes* permet d'imposer des conditions à l'existence d'un nœud (l. 7) ou à sa non-existence (l. 12). Ces gardes s'expriment comme des expressions booléennes sur des équations entre chemins. Le compilateur de MG vérifie la satisfiabilité de ces gardes, éliminant les alternatives conduisant à des échecs et les équations devenues tautologiquement vraies. Les gardes restantes sont alors émises dans les arbres TAG pour être évaluées pendant l'analyse.

Outre les extensions des méta-grammaires et du compilateur, le travail de description a été facilité par le déploiement d'un environnement de travail adapté pour pouvoir aisément visualiser et tester. En premier lieu, nous disposons d'un mode Emacs pour les MG interagissant avec un outil graphique de visualisation de la hiérarchie des classes. Par ailleurs, la chaîne de traitement allant des méta-grammaires aux analyseurs produit des représentations intermédiaires sous formats XML[6] pouvant être visualisées, en particulier sous forme HTML pour les arbres, décorations et gardes. Les forêts de dérivations produites par notre analyseur sont également convertibles en XML et visualisables sous différentes formes, en particulier sous forme de dépendances. L'utilisation d'un serveur d'analyseurs[7] couplé à divers scripts facilite la conduite de tests sur corpus, pour mesurer divers paramètres (temps d'analyse, taux d'ambiguïté, taux de couverture, ...) et indiquer les différences entre 2 séries de tests. Enfin, il est possible de désactiver des classes[8] pour déboguer ou, à terme, pour obtenir des grammaires spécialisées. Ces diverses possibilités permettent un suivi fin des performances de la grammaire engendrée.

4 Anatomie de la grammaire produite

Grâce aux résultats décrits précédemment, nous avons pu rapidement concevoir une méta-grammaire du français engendrant une grammaire très compacte, comme le montrent les diverses tables de la figure 1. Ainsi, la grammaire ne comporte que 133 arbres, incluant 7 arbres construits manuellement. Elle est essentiellement TIG avec seulement 12 arbres auxiliaires enveloppants principalement utilisés pour gérer les diverses formes de guillemets[9]. La grammaire n'est pas totalement lexicalisée, avec un nombre assez important d'arbres sans ancre (mais

[5] À terme, la valeur du trait sera exploitée pour pouvoir spécifier un intervalle de répétition.

[6] Ces formats XML s'appuient de plus sur les propositions de normalisation, à savoir TAGML pour les TAG et FSR pour les structures de traits.

[7] Accessible en ligne sur `http://atoll.inria.fr/parserdemo`.

[8] Il est en fait possible d'activer ou désactiver de manière plus fine, en exprimant un ensemble de contraintes invalidant une classe.

[9] Pour être plus précis, ces arbres sont uniquement utilisés pour les guillemets autour de groupes, ceux autour de mots simples sont gérés avant analyse syntaxique. Le traitement proposé est clairement une source d'inefficacité pouvant peut-être être géré autrement. Par ailleurs, il est à noter que le compilateur MG a produit plus

possédant éventuellement des nœuds lexicaux), essentiellement utilisés pour des adjonctions[10]. Les arbres ancrés le sont surtout par les verbes mais ils ne représentent qu'une infime fraction d'un ensemble équivalent d'arbres TAG non factorisés. On voit que 7 arbres suffisent à couvrir un ensemble conséquent de constructions verbales « canoniques ». Ces résultats découlent d'un usage intensif de la factorisation dans les arbres, en particulier contrôlée par des gardes. L'étoile de Kleene est uniquement utilisée pour gérer la coordination tandis que les entrelacements proviennent essentiellement d'un ordre libre entre arguments du verbe (incluant le sujet post-verbal). Les arbres factorisés obtenus peuvent être relativement conséquents (jusqu'à 46 nœuds) mais la figure 1(e) montre néanmoins que la plupart des arbres restent simples.

Classes	Arbres	Init.	Aux.	Aux. Env.	Aux. Gauches	Aux. Droits
191	133=126+7	44	89	12	29	48

(a) Distribution par types d'arbres

non ancrés	v	coo	adv	adj	csu	prep	aux	np	nc	det	pro
50	27	12	10	8	4	3	2	2	1	1	1

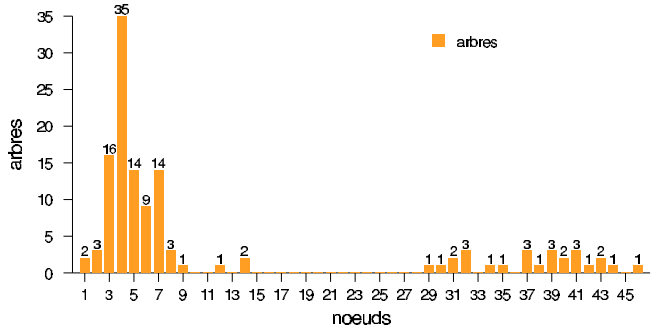
(b) Distribution par ancres

Canonique	Extr.	Actif	Passif	Quest.	Rel.	Clivées	Coord	Adv	Adj
7	19	19	6	4	4	11	12	14	11

(c) Distribution par phénomènes syntaxiques

Gardes	Disjonctions	Entrelacement	Étoiles de Kleene
820	92	26	13

(d) Distribution des factorisations



(e) Distribution des tailles d'arbres

FIG. 1 – Anatomie de la grammaire

La complexité des arbres factorisés est illustrée par la figure 2 représentant une vue simplifiée d'un des arbres verbaux canoniques pour la voix active. Cet arbre `#111` résulte du croisement de 25 classes terminales, comprend 43 nœuds plus 3 nœuds d'alternatives et 1 nœud d'entrelacement, et est contrôlé par 35 gardes[11]. Il est difficile d'obtenir le taux exact de factorisation atteint, mais voici néanmoins quelques paramètres indicatifs pour essayer de l'estimer :

d'arbres auxiliaires que nécessaire pour éviter d'avoir des pseudo-arbres enveloppants et que certains phénomènes syntaxiques pouvant produire des arbres enveloppants ont été bridés pour obtenir des arbres TIG.

[10]Cette non lexicalisation partielle est guidée par des raisons pragmatiques (limitation du nombre d'arbres) mais également linguistiques. Elle ne remet pas en cause la notion de domaine de localité sémantique des arbres TAG. Au contraire, l'accroche d'une participiale sur un nom, par exemple, est non lexicalisée car distincte (sémantiquement) de la construction d'une participiale.

[11]Un tel arbre avec toutes ses gardes et décorations serait extrêmement difficile à écrire à la main (2171 lignes de XML TAGML), ce qui justifie d'autant plus le recours à une méta-grammaire.

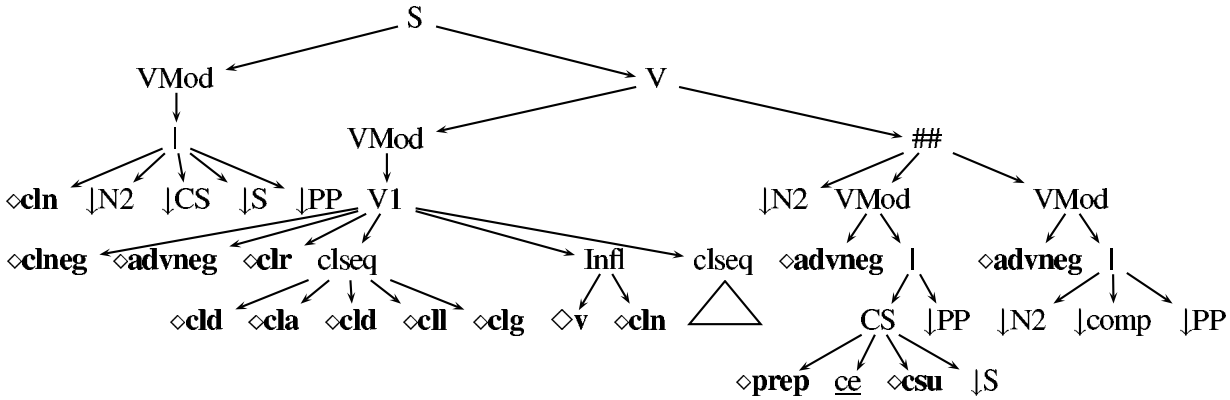


FIG. 2 – Arbre #111 (simplifié)

subjreal = 5 choix de réalisations pour un sujet pré-verbal : `cln` (clitique nominatif), `N2` (groupe nominal), `CS` (phrase), `S` (infinitive), `PP` (infinitive introduite par « d' ») ;
subjpos choix entre « pas de sujet », « sujet pré-verbal », « sujet pré-verbal et clitique post-verbal », « clitique post-verbal », « sujet post-verbal » ;
argspos = 3 entrelacements entre 2 arguments post-verbaux et un sujet post-verbal, tous optionnels ;
clseqpos = X choix entre 2 séquences de clitiques pré ou post verbaux optionnels : `cld`, `cla`, `cld`, `cll`, `clg` ;
clopt = 2 clitiques pré-verbaux optionnels : `clneg` (« ne »), `clr` (réflexif) ;
neg = 3 positions de négation (advneg « que » sur les arguments verbaux ou advneg pré-verbal pour les infinitives) ;
arg1real = 3 réalisations possibles pour arg1 : `N2` (objet), `comp` (attribut), `PP` (groupe prépositionnel nominal, adjectival, infinitif, complétif, ...) ;
arg2real = 2 réalisations possibles (au moins) pour arg2 (CS,PP) ;

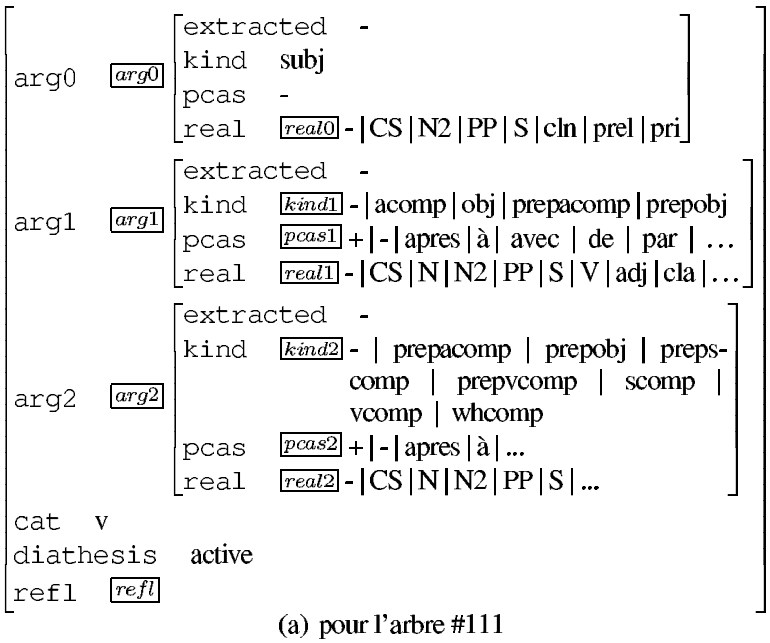
Divers calculs approchés donnent un taux de factorisation entre 10 000 et 100 000, sans même prendre complètement en compte les disjonctions apportées par certaines gardes. Ce taux d'expansion astronomique est bien entendu très largement au-dessus de la réalité si on prend mieux en compte la satisfiabilité des gardes. Néanmoins, même un taux d'expansion plus réaliste de l'ordre de 100 ou 1000 montre l'intérêt des ces structures factorisées, difficilement constructibles manuellement sans le support apporté par les méta-grammaires. De plus, les expériences menées ont montré que l'analyse reste efficace même avec des structures aussi complexes[12].

Un arbre comme l'arbre `#111` couvre de fait de très nombreuses constructions syntaxiques, bien plus en général que celles autorisées pour un mot donné. L'ancrage entre mots et arbres, ainsi que la sélection des constructions autorisées pour un mot se fait par unification entre des structures de traits appelées *hypertag* associés aux arbres et aux mots (Kinyon, 2000)[13]. Ainsi la figure 3(a) donne l'hypertag produit pour l'arbre `#111` tandis que la figure 3(b) montre un des hypertags mentionnés pour « promettre » dans notre lexique. Parmi toutes les constructions possibles pour l'arbre `#111`, l'unification sélectionne uniquement les constructions verbales canoniques autorisant un objet optionnel et un groupe prépositionnel optionnel introduit par

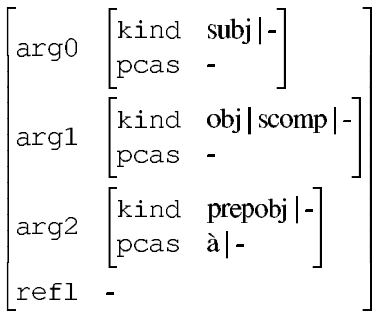
[12]En général, les mauvaises performances résultent plutôt d'arbres sous-contraints faisant exploser le nombre d'analyses.

[13]Ce mécanisme d'ancrage par hypertags rend caduque la notion trop rigide de famille d'arbres présente dans l'architecture XTAG (Doran *et al.*, 1994).

« à »[14]. Le lien entre hypertag $\mathcal{H}$ et des constructions syntaxiques se fait grâce aux variables présentes dans $\mathcal{H}$ et dans les décorations des nœuds ou dans les équations des gardes.



(a) pour l'arbre #111



(b) pour «promettre»

FIG. 3 – Hypertags

5 Expériences

L'analyseur hybride TAG/TIG compilé à partir de la grammaire a été testé sur divers corpus tout au long de la phase de développement et pendant la campagne EASY. Les analyses s'effectuent sur des treillis de mots (pour gérer les ambiguïtés morpho-syntaxiques et les mots inconnus) en s'appuyant sur un lexique de plus de 400 000 formes fléchies fournissant des informations de sous-catégorisation pour les verbes. Nous n'avons pas utilisé d'étiqueteur morpho-syntaxique. L'analyseur s'appuie sur une stratégie d'analyse tabulaire descendante gauche-droite et peut rendre soit une analyse complète de la phrase soit un ensemble d'analyses partielles couvrant au mieux l'entrée. Les analyses sont extraites sous forme de forêts partagées de dérivations, convertibles en forêts partagées de dépendances. Ces forêts nous servent de base pour calculer un *taux moyen d'ambiguïté par mot* α défini comme le nombre moyen d'arcs de dépendances atteignant un mot moins un[15].

Corpus	#phrases	% couv.	temps moyen (s)	temps médian (s)	ambiguïté
EUROTRA / OLD	334	95.80 / 89.22	1.81 / 0.70	1.27 / 0.54	0.7 / 0.3
TSNLP / OLD	1661	93.38 / 86.15	0.72 / 0.43	0.56 / 0.33	0.4 / 0.2
MD10x20 / OLD	5000	63.18 / 43.06	2.85 / 1.97	1.80 / 1.30	0.8 / 0.5
EASY	34438	42.45 / -	5.55 / -	1.61 / -	0.6 / -

TAB. 1 – Résultats (avec un *timeout* de 100s)

[14] La sélection des constructions avec une complétive (`scomp`) se fait avec un autre arbre.

[15] Pour une analyse non-ambiguë, tout mot sauf la « tête » de la phrase est atteignable par une seule dépendance. Le nombre maximal d'analyses pour un taux d'ambiguïté α et une phrase de longueur n est en $O((1+\alpha)^n)$.

La table 1 fournit des résultats d'analyses complètes de 2 versions successives de l'analyseur pour les jeux de tests EUROTRA et TSNLP ainsi que pour MD10x20, un corpus journalistique de phrases de longueur comprise entre 10 et 20 extraites (naïvement) du « *Monde Diplomatique* » et pour le corpus fourni pour la campagne EASY (couvrant divers styles : journalistique, littéraire, oral, mail, médical, questions/réponses). Les résultats de couverture sont excellents sur les jeux de tests, en particulier à cause d'un vocabulaire relativement restreint pour lequel notre lexique est complet. Sur le corpus MD10x20 qui est relativement homogène et pour lequel un minimum d'adaptation du lexique a été effectué, les résultats restent honorables. Les résultats sont moins bons pour EASY qui est très hétérogène[16]. Ce manque de couverture traduit bien évidemment des manques dans la méta-grammaire, en particulier sur les coordinations complexes, les superlatives et les comparatives, ainsi que sur les cadres de sous-catégorisation pour les catégories non-verbales et sur certaines articulations de phrase. Cependant, le manque de couverture provient également de notre lexique qui est très récent et ne fournit pas nécessairement des informations syntaxiques complètes voire correctes pour tous les mots[17].

La table 1 fournit aussi des résultats pour une version antérieure de l'analyseur (OLD) qui illustrent l'importance du suivi constant des grammaires. En effet, nous avons effectué, sans réel contrôle, des modifications de dernière minute avant EASY pour essayer d'améliorer la couverture (extension des clivées, généralisation abusive des incises, articulation des phrases par la ponctuation, gestion naïve des verbes support, ...). Ces modifications ont bien augmenté la couverture, mais, mal contrôlées, elles ont fait doubler les taux d'ambiguïté et les temps d'analyse (avec en première approximation, une relation linéaire entre temps et taux d'ambiguïté).

Enfin, sans corpus de référence, il nous est impossible pour l'instant de fournir des résultats concernant la précision des analyses (complètes ou partielles). Nous avons effectué de nombreuses vérifications manuelles sur les vues graphiques des forêts mais attendons maintenant les résultats de la campagne EASY pour avancer.

6 Conclusion

Notre méta-grammaire est encore loin d'être complète mais l'expérience montre néanmoins que les méta-grammaires rendent possible le développement rapide de grammaires à relativement large couverture. Il est à noter que ce développement a été en partie freiné par le manque d'information dans le lexique, en particulier pour avoir une discrimination plus fine des adverbes et pour traiter les sous-catégorisations des adjectifs et des noms.

Les extensions apportées aux méta-grammaires ainsi que les améliorations de notre environnement de travail se sont révélées très utiles. Néanmoins, concevoir une méta-grammaire reste un exercice délicat demandant une solide expertise linguistique et une utilisation systématique d'outils de tests et de visualisation. Il nous semble aussi souhaitable d'ajouter de nouveaux types de contraintes, même si elles peuvent s'exprimer à l'aide des contraintes actuelles, comme des contraintes d'exclusion entre nœuds, des contraintes de cardinalité pour exprimer des règles topologiques, ou des contraintes de rangs exprimables dans les gardes. Pour aller dans le sens de grammaires paramétrables (autorisant divers niveaux de langue) ou pour aller vers des formalismes cibles distincts, il serait utile de regrouper les contraintes par *contextes* à l'intérieur des

[16] On peut aussi préciser que les phrases pour EASY sont en moyenne plus longues.

[17] Mais nous exploitons progressivement les résultats d'analyse pour repérer et corriger les entrées incorrectes ou incomplètes.

classes de manière à pouvoir plus facilement n'en exploiter qu'une partie lors de la compilation (en sélectionnant un ensemble de contextes).

La factorisation des arbres, rendue possible par l'emploi de gardes et d'opérateurs réguliers, nous semble une approche générique extrêmement prometteuse pour contrôler l'explosion combinatoire du nombre de structures grammaticales produites, permettant ainsi de construire des analyseurs syntaxiques plus efficaces. Les arbres factorisés peuvent être complexes mais leur description au niveau de la méta-grammaire reste simple.

Le formalisme cible TAG que nous avons utilisé est judicieux mais néanmoins pas suffisamment puissant pour exprimer élégamment certains phénomènes syntaxiques comme les incises ou certaines extractions (comme l'extraction de génitifs dans « *de qui lis-tu un livre* »). Nous envisageons d'évoluer vers des formalismes cibles permettant d'exprimer plus de sous-spécification dans les arbres, comme par exemple les *Local Multi Component TAG*, avec l'ambition, à terme, de réduire la distance entre les méta-grammaires et le formalisme cible.

Les outils mentionnés dans cet article ainsi que la méta-grammaire sont librement disponibles[18].

Références

ABEILLÉ A. (2002). *Une grammaire électronique du français*. Paris : CNRS Editions.

ALONSO M. A. & DÍAZ V. J. (2003). Variants of mixed parsing of TAG and TIG. *Traitement Automatique des Langues (T.A.L.)*, **44**(3), 41–65.

CANDITO M.-H. (1999). *Organisation modulaire et paramétrable de grammaires électroniques lexicalisées*. PhD thesis, Université Paris 7.

CARROLL J., NICOLOV N., SMETS M., SHAUMYAN O. & WEIR D. (1998). Grammar compaction and computation sharing in automata-based parsing. In *Proceedings of Tabulation in Parsing and Deduction (TAPD'98)*, p. 16–25, Paris (FRANCE).

CLÉMENT L. & KINYON A. (2003). Generating parallel multilingual LFG-TAG grammars from a metaGrammar. In *Proc. of ACL'03*.

DORAN C., EGEDI D., HOCKEY B. A., SRINIVAS B. & ZAIDEL M. (1994). XTAG system — a wide coverage grammar for English. In *Proc. of the 15th International Conference on Computational Linguistics (COLING'94)*, p. 922–928, Kyoto, Japan.

GAIFFE B., CRABBÉ B. & ROUSSANALY A. (2003). Représentation et gestion du lexique d'une grammaire d'arbres adjoints. *Traitement Automatique des Langues (T.A.L.)*, **44**(3).

HARBUSCH K. & WOCH J. (2004). Integrated natural language generation with schema-tree adjoining grammars. In C. HABEL & E. THOMAS PECHMANN, Eds., *Language Production*. Mouton De Gruyter.

JOSHI A. K. (1987). An introduction to tree adjoining grammars. In A. MANASTER-RAMER, Ed., *Mathematics of Language*, p. 87–115. Amsterdam/Philadelphia : John Benjamins Publishing Co.

KINYON A. (2000). Hypertags. In *Proc. of COLING*, p. 446–452.

NEDERHOF M.-J., SATTA G. & SHIEBER S. (2003). Partially ordered multiset context-free grammars and free-word-order parsing. In *In 8th International Workshop on Parsing Technologies (IWPT'03)*, p. 171–182.

SCHABES Y. & WATERS R. C. (1995). Tree insertion grammar : a cubic-time, parsable formalism that lexicalizes context-free grammar without changing the trees produced. *Fuzzy Sets Syst.*, **76**(3), 309–317.

VILLEMONTE DE LA CLERGERIE E. (2002). Construire des analyseurs avec DyALog. In *Proc. of TALN'02*.

[18] Sur `http://atoll.inria.fr/packages/packages.html`.

XMG : Un Compilateur de Méta-Grammaires Extensible*

Denys Duchier (1), Joseph Le Roux (2), Yannick Parmentier (2)
(1) LIFL - UMR CNRS 8022- Université des Sciences et Technologies de Lille
Bâtiment M3 59655 Villeneuve d'Ascq Cédex
duchier@lifl.fr
(2) INRIA / LORIA - Institut National Polytechnique de Lorraine - Université Henri Poincaré Nancy 1
615, rue du Jardin Botanique - 54 600 Villers-Lès-Nancy
{leroux,parmenti}@loria.fr

Mots-clefs : Grammaires, compilation, ressources linguistiques, Grammaires d'Arbres Adjoints, Grammaires d'Interaction

Keywords: Grammars, compilation, linguistic resources, Tree Adjoining Grammars, Interaction Grammars

Résumé Dans cet article, nous présentons un outil permettant de produire automatiquement des ressources linguistiques, en l'occurence des grammaires. Cet outil se caractérise par son extensibilité, tant du point de vue des formalismes grammaticaux supportés (grammaires d'arbres adjoints et grammaires d'interaction à l'heure actuelle), que de son architecture modulaire, qui facilite l'intégration de nouveaux modules ayant pour but de vérifier la validité des structures produites. En outre, cet outil offre un support adapté au développement de grammaires à portée sémantique.

Abstract In this paper, we introduce a new tool for automatic generation of linguistic resources such as grammars. This tool's main feature consists of its extensibility from different points of view. On top of supporting several grammatical formalisms (Tree Adjoining Grammars and Interaction Grammars for now), it has a modular architecture which eases the integration of modules dedicated to the checking of the output structures. Furthermore, this tool offers adapted support to the development of grammars with semantic information.

* Nous tenons à remercier Benoît Crabbé, Claire Gardent, Guy Perrier et Eric De La Clergerie pour les nombreuses discussions qui nous ont aidées dans le cadre de ce travail.

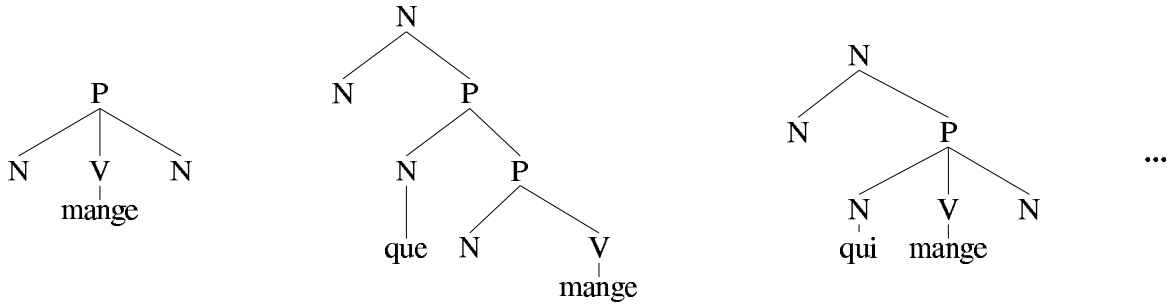
1 La production automatique de grammaires

Dans le cadre du développement d'applications de TALN (analyseurs syntaxiques par exemple), un certain nombre de ressources linguistiques est nécessaire, parmi lesquelles des grammaires. Nous nous intéressons ici aux grammaires fortement lexicalisées associant à chaque mot un ensemble de structures syntaxiques décrivant son comportement dans chaque emploi.

1.1 Le besoin de production automatique

Par opposition aux grammaires de règles, les grammaires fortement lexicalisées utilisées en linguistique informatique, comme les grammaires d'arbres adjoints lexicalisées (TAG), les grammaires d'interaction (IG) ou les grammaires catégorielles reportent la connaissance de la langue dans le lexique. Un tel lexique peut être vu comme une fonction associant à un mot de la langue l'ensemble des structures syntaxiques représentant ses usages. Pour avoir une bonne couverture, il est nécessaire que le lexique associe à chaque mot un maximum de structures. Par exemple pour analyser les expressions ci-dessous, le lexique doit associer au verbe *mange* (en TAG) les structures suivantes :

Jean mange la pomme. *La pomme que Jean mange ...* *Jean qui mange une pomme.* ...



Cette taille de lexique ne va pas sans problème[1]. Comme les règles syntaxiques sont *éclatées* à travers le lexique, l'analyse d'un nouveau phénomène linguistique ou, pire, sa révision peuvent avoir de graves conséquences sur la cohérence de la grammaire. Si les premiers lexiques à large couverture furent écrits à la main, leur génération automatique devient de plus en plus pressante, de manière à (1) pouvoir garantir la cohérence de l'ensemble et (2) pouvoir réviser une grammaire rapidement, même quand le lexique devient important. Nous avons développé un formalisme, dont l'implantation est XMG, qui, à partir d'une description métagrammaticale concise génère toutes les structures associées aux mots du lexique. Notre approche se veut la plus indépendante possible des formalismes grammaticaux et capable de gérer simultanément les aspects syntaxiques et sémantiques du lexique.

1.2 Principes de la production automatique

Le lexique est extrêmement redondant : d'une part, une même structure syntaxique peut être associée à plusieurs entrées lexicales (*e.g.* les verbes transitifs auront un grand nombre de structures syntaxiques en commun) et d'autre part les différentes structures lexicales partagent des

[1] Les problèmes purement informatiques ne sont pas évoqués ici.

fragments communs importants (*e.g.* le fragment sujet-verbe se retrouve de très nombreuses fois). L'idée qui s'est imposée dès les premiers travaux de génération automatique de lexique par (Candito, 1996) est de ne décrire que ces fragments élémentaires, pouvant hériter les uns des autres, puis de les combiner pour former toutes les structures syntaxiques du lexique. Ces combinaisons constituent la justification linguistique de cette approche[2]. En effet elles doivent expliquer la formation des différentes structures apparentées (par exemple la relation entre les formes actives et passives d'un même verbe). Les différents formalismes métagrammaticaux doivent donc rendre exprimables la notion de fragments réutilisables, la spécification de leur structure et la manière de les combiner pour produire des structures complètes. Par exemple, (Candito, 1996) explique les croisements par un ensemble de contraintes (de 3 types) que doivent vérifier les structures finales et (Gaiffe, Crabbé et Roussanaly, 2002) les justifie en assouplissant la notion de contrainte par une approche où besoins et ressources doivent être satisfaits. Notre approche les explique par deux actions primitives *l'accumulation* et *la composition disjonctive* auxquelles peuvent être ajoutées des contraintes (sensibilité aux ressources, comme par exemple un langage de couleur, voir section 2.2). Mais XMG se distingue des approches antérieures par trois aspects fondamentaux :

1. XMG est **multi-formalisme**, c'est à dire qu'il ne se limite pas à un formalisme syntaxique en particulier,
2. XMG est **extensible**, d'une part on peut lui ajouter des niveaux de description pour traiter de nouveaux formalismes (aussi bien syntaxiques que sémantiques), et d'autre part, on peut l'étendre en définissant des modules de contraintes additionnelles que les structures grammaticales produites doivent respecter,
3. la description métagrammaticale est vue comme un **programme logique** dont XMG est le compilateur, ce qui nous permet de réutiliser des techniques issues de la programmation logique.

1.3 L'extensibilité de XMG

XMG présente une grande originalité vis à vis des autres cadres méta-grammaticaux : il est à la fois multi-formalisme [3] (génère aussi bien des TAG que des IG), et extensible. Cette extensibilité est atteinte d'une part grâce à un langage de contrôle muni des opérations d'accumulation et de composition disjonctive qui sont générales, et ne décrivent que des relations de fragment à fragment. Ce langage est donc indépendant du formalisme grammatical cible et permet d'exprimer des combinaisons très fines. D'autre part, dans XMG, chaque fragment peut contenir un nombre arbitraire de niveaux de description (appelés aussi *dimensions*) tels que le niveau syntaxique, sémantique, etc. Ces dimensions ne sont pas complètement indépendantes puisqu'elles peuvent partager de l'information (en particulier pour l'interface syntaxe / sémantique). En outre, chaque dimension est munie de son propre langage de description qui varie selon le type d'information contenue. Par exemple, pour les TAG et les IG la dimension syntaxique est représentée par un langage de description d'arbres intégrant l'unification et où chaque nœud est équipé d'une structure de traits, tandis que pour les grammaires de dépendances extensibles (XDG) les entrées lexicales seraient des structures de traits atomiques. Enfin, XMG est extensible par l'ajout de modules spécifiques pour traiter les structures intermédiaires produites. Il existe deux types de

[2]Les approches antérieures ne s'intéressaient qu'au problème de redondance.

[3]Les approches de méta-grammaires précédentes ne permettaient que la production de grammaires TAG, bien que les travaux de (Clément et Kinyon, 2003) montrent qu'il est possible d'adapter le résultat obtenu pour générer des grammaires fonctionnelles lexicales (LFG).

modules : (1) ceux qui créent de nouvelles structures (par exemple, pour TAG, en produisant des arbres à partir de descriptions) et (2) ceux qui filtrent les structures correctes suivant des critères linguistiques (voir section 2.2).

A ce jour, la support du multi-formalisme a été validé dans XMG via TAG et IG, cependant il n'est pas encore possible de sélectionner la dimension *syntaxique* à utiliser (seule un langage de description d'arbres a été implanté) : l'extensibilité déclarative n'est pas encore atteinte.

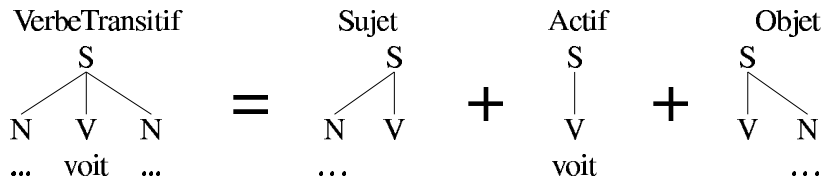
Par la suite, nous allons tout d'abord présenter le procédé de compilation de XMG, ensuite nous verrons concrètement comment produire automatiquement une grammaire TAG, enfin nous aborderons la question de l'intégration de XMG dans une chaîne de TALN.

2 Un procédé de compilation basé sur des techniques de programmation logique

Dans cette section, nous allons décrire le procédé de compilation de méta-grammaire utilisé par XMG. Plus précisément, nous allons présenter (a) le langage abstrait dont XMG est l'implantation. Notons que nous ne présentons pas ici la syntaxe *concrète* du langage (voir section 3 pour cela), mais le langage de plus bas niveau dans lequel cette syntaxe est traduite lors de la compilation. Nous présenterons également (b) l'architecture de XMG, qui par sa modularité, s'adapte aisément à différents formalismes grammaticaux.

2.1 Un langage de représentation étendu

Combinaison de fragments Nous avons vu à la section précédente que la compilation d'une méta-grammaire correspondait à la combinaison de fragments[4]. Nous pouvons définir de telles combinaisons sous forme de règles de réécriture. Par exemple, considérons l'arbre syntaxique associé à une entrée lexicale telle que *voit* (*i.e.* un verbe transitif) dans le formalisme TAG. Disposant des fragments d'arbres $Sujet$, $Actif$ et $Object$, nous pouvons réécrire l'arbre $VerbeTransitif$ comme la conjonction de ces 3 fragments :



Ce qui s'écrit également comme suit :

$$VerbeTransitif \rightarrow Sujet \wedge Actif \wedge Objet \quad (1)$$

Nous nous ramenons ainsi au formalisme des *grammaires de clauses définies* (DCG), dans lequel les terminaux ne seraient pas des mots mais des fragments d'arbres. La compilation d'une méta-grammaire s'identifie alors à la compilation d'un programme logique. Dans XMG, la

[4]Nous laissons provisoirement de côté la question de la contrainte des combinaisons, voir 2.2.

méta-grammaire est décrite au moyen d'un langage de représentation, qui comprend les notions d'*abstraction* (*cf* 2) et de *composition conjonctive et disjonctive* (*cf* 3).

$$\textit{Clause} \quad ::= \quad \textit{Nom} \rightarrow \textit{But} \tag{2}$$

$$\textit{But} \quad ::= \quad \textit{Description} \mid \textit{Nom} \mid \textit{But} \vee \textit{But} \mid \textit{But} \wedge \textit{But} \tag{3}$$

On remarque que XMG fournit un langage expressif incluant la disjonction et de ce fait introduit de l'indéterminisme dans la combinaison des fragments d'arbres. Ainsi, nous pouvons préciser l'exemple précédent de l'arbre associé aux verbes transitifs en spécifiant que le sujet peut être sous forme canonique *ou* sous forme relative, etc. Cela s'énonce via la règle suivante :

$$\textit{Sujet} \quad \rightarrow \quad \textit{SujetCan} \vee \textit{SujetRel} \vee \ldots \tag{4}$$

Description des fragments Les fragments dénotés par les abstractions (nommées également *classes*) de notre langage de représentation sont décrits au moyen de contraintes de dominance :

$$\textit{Description} ::= x \rightarrow y \mid x \rightarrow^{+} y \mid x \rightarrow^{*} y \mid x \prec y \mid x \prec^{+} y \mid x \prec^{*} y \mid x[f{:}E] \mid x(p{:}E) \tag{5}$$

où x, y représentent des variables de noeuds, $\rightarrow$ la dominance immédiate, $\rightarrow^{+}$ la dominance stricte, $\rightarrow^{*}$ la dominance large, $\prec$ la précédence immédiate, $\prec^{+}$ la précédence stricte, $\prec^{*}$ la précédence large, $x[f{:}E]$ l'association du trait f au noeud x et $x(p{:}E)$ l'association de la propriété p à ce même noeud x.

Ainsi, nous disposons d'un langage suffisamment expressif pour supporter les formalismes syntaxiques basés sur les descriptions d'arbres, telles que les grammaires TAG et IG.

Représentation sémantique L'extensibilité de XMG se retrouve également dans le fait qu'il offre un support adéquat à l'intégration d'une représentation sémantique dans la méta-grammaire. En effet, notre langage de représentation permet non seulement de placer dans les classes des informations syntaxiques (*e.g.* descriptions d'arbres), mais également des représentations sémantiques. A l'heure actuelle, un langage à base de structures prédicatives a été implanté :

$$\textit{Description} ::= \ell{:}p(E_1, ..., E_n) \mid \neg\ell{:}p(E_1, ..., E_n) \mid E_i \ll E_j \tag{6}$$

Ici, $\ell{:}p(E_1, ..., E_n)$ représente le prédicat p avec les arguments $E_1, .., E_n$, et étiqueté par ℓ, $\neg$ l'opérateur de négation, et $E_i \ll E_j$ la portée entre les variables sémantiques E_i et E_j.

Un tel langage peut servir, par exemple, à associer aux arbres des informations de type sémantique plate comme dans (Gardent et Kallmeyer, 2003).

En reprenant notre comparaison entre méta-grammaire et DCG se pose la question de la distinction entre information syntaxique et information sémantique. Cela est rendu possible par l'utilisation de *dimensions*, correspondant chacune à un *accumulateur* spécifique dans le formalisme des *grammaires de clauses définies étendues* (EDCG) (Van Roy, 1990).

Ainsi, la combinaison de classes (*i.e.* d'informations syntaxiques ou sémantiques) donne lieu à l'accumulation de leur contenu dans des structures dédiées. L'expression (3) vue précédemment est alors étendue en remplacant $Description$ par :

$$\textit{Dimension} \mathrel{+}= \textit{Description}$$

Ce type accumulation permet de traiter plusieurs dimensions. A l'heure actuelle, XMG dispose de 3 dimensions, notées **syn**, **sem** et **dyn**. Les deux premières représentent les dimensions syntaxique et sémantique et la dernière une dimension utilisée pour l'accumulation d'informations lexicales.

Notons que ces dimensions peuvent partager des variables logiques, ce qui permet un développement relativement naturel d'une interface syntaxe / sémantique au niveau méta-grammatical.

Portée des variables Dans les approches antérieures les noms de variables ont une portée globale à la méta-grammaire. Ce qui pose des problèmes de conflits de noms, passé un certain nombre de classes. Dans XMG, les identifiants ont par défaut une portée limitée à la clause, à laquelle est intégré un procédé d'export d'identifiants. Ainsi, il est possible de spécifier avec précision le domaine de visibilité d'une variable. Plus précisément, à chaque clause est associée une structure de traits contenant les identifiants exportés, (2) devient :

$$\mathit{Clause} \quad ::= \quad \langle f_1{:}E_1, \ldots, f_n{:}E_n \rangle \Leftarrow \mathit{Nom} \quad \rightarrow \quad \mathit{But} \tag{7}$$

Et l'appel de classe s'accompagne de l'accès à la structure d'export (notée *Var* ici), (3) est remplacée par :

$$\mathit{But} \quad ::= \quad \mathit{Dim} \mathrel{+}= \mathit{Description} \mid \mathit{Var} \Leftarrow \mathit{Nom} \mid \mathit{But} \vee \mathit{But} \mid \mathit{But} \wedge \mathit{But} \tag{8}$$

On peut accéder alors à un identifiant *X* via la notation pointée *Var.X*.

2.2 Une architecture modulaire

En section 1, nous avons présenté les caractéristiques de XMG, dont le fait qu'il supporte plusieurs formalismes syntaxiques. Cette caractéristique est fortement liée à l'architecture modulaire du compilateur.

Des modules dédiés La compilation d'une méta-grammaire se fait en plusieurs phases, dont certaines diffèrent suivant le formalisme. Chacune de ces phases est prise en charge par un module spécifique.

Actuellement XMG comporte 3 modules principaux :

a) La partie avant (*i.e.* le compilateur proprement dit) traduit la méta-grammaire en instructions exprimées dans un langage de plus bas niveau.

b) Ces instructions sont ensuite exécutées par une machine virtuelle (MV) de type *Warren's Abstract Machine* (WAM, voir (Ait-Kaci, 1991)).

Cette MV réalise l'unification des structures de données de la méta-grammaire (*i.e.* structures de traits associées aux noeuds, traits polarisés pour IG, etc), puis l'accumulation des dimensions pour une combinaison de classes donnée. En sortie de la MV, nous disposons de données accumulées dans chaque dimension, dans le cas des TAG, des descriptions d'arbres dont il faut calculer les solutions.

c) En plus de la partie avant et de la MV, qui sont communes aux formalismes des TAG et des IG, XMG intègre un module de résolution de descriptions d'arbres. Ce module est programmé sous forme d'un résolveur de contraintes (voir (Duchier et Niehren, 2000) pour une description complète du procédé).

Un résolveur extensible La modularité dans XMG est encore étendue par la programmation de modules additionnels optionnels et paramétrables. Ces modules permettent d'étendre les fonctionnalités du module de résolution pour, par exemple, contraindre les modèles produits par XMG selon des critères spécifiques, appelés aussi *principes*.

Les principes instanciables présents dans XMG actuellement sont de 3 types :

i) un principe de **couleurs.** Comme annoncé à la section 1.2, dans un contexte de développement de grammaires à large couverture par combinaison de fragments, une idée majeure consiste à contraindre les combinaisons acceptables en intégrant un système de gestion de ressources. Pour gérer les ressources en TAG, XMG intègre un langage de couleurs. Ce langage permet d'indiquer quels fragments d'arbres nécessitent quels autres fragments pour former un modèle valide (une présentation de l'emploi d'un langage de couleurs pour produire une grammaire TAG est donnée dans (Crabbé et Duchier, 2004)).

ii) un principe d'**unicité** paramétré par une propriété de noeuds. Ce principe permet de garantir la validité des solutions produites par XMG par rapport à une contrainte linguistique d'unicité telle que : *dans un arbre TAG, il ne peut y avoir deux extractions*[5].

iii) un principe de **rang** paramétré par un nombre entier, permettant de réaliser l'ordonnancement des clitiques dans les arbres TAG produits.

3 Cas d'étude : une méta-grammaire pour TAG

Voyons à présent comment écrire une méta-grammaire pour TAG avec XMG. Pour cela nous allons introduire la syntaxe *concrète* de XMG, qui est destinée à être traduite dans le langage *abstrait* présenté précédemment lors de la compilation. Notons que nous n'entrons pas ici dans les aspects du développement de grammaires à large couverture, nous nous focalisons sur un exemple simple, introductif vis à vis des possibilités offertes par XMG.

Nous allons considérer la méta-grammaire permettant de générer les arbres TAG pour les verbes transitifs à l'actif avec sujet canonique *ou* extrait *et* objet sous forme canonique (exemple 1 de la section 2.1)[6].

Les déclarations La première information que doit fournir notre méta-grammaire consiste en la spécification du résolveur utilisé (si nous en utilisons un). Ici nous allons utiliser le principe d'unicité de la fonction grammaticale sujet :

```
use unicity with (fg = suj) dims (syn)
```

Cette instruction indique à XMG que le résolveur ne doit accepter que les solutions pour lesquelles il n'y a qu'un seul noeud ayant la propriété *fg* = *suj* (respect du principe d'unicité).

Ensuite, nous déclarons les types de données utilisés. Ces types permettent de typer les propriétés et structures de traits qui seront associées aux noeuds dans les descriptions syntaxiques :

```
type CAT={n,v,p}
type FG={suj,obj}
property fg : FG
feature cat : CAT
```

[5]Nous adoptons cette convention dans un contexte méta-grammatical, bien que certains exemples de phrases à double extraction existent (cf (Abeillé, 2002)).

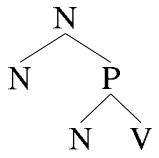
[6]Nous ne présentons pas le formalisme TAG ici, le lecteur est dirigé vers (Joshi et Schabes, 1997) pour une telle introduction.

Les classes A présent, nous pouvons définir le contenu de nos classes, *i.e.* les fragments d'arbres. Considérons le fragment correspondant au sujet canonique. Celui-ci est composé de 3 noeuds x, y et z, tels que $x \rightarrow y \wedge x \rightarrow z \wedge y \prec z$. Ce qui s'écrit comme suit dans la syntaxe de XMG ($x \rightarrow y$ correspond à `node ?X{node ?Y}`) :

```
class SujetCan
export ?X ?Z
declare ?X ?Y ?Z
{ <syn>{
        node ?X [cat = s]{
                node ?Y (fg=suj)[cat=n]
                node ?Z [cat = v]
                }
        }
}
```

La classe ObjetCan s'écrit de façon similaire. La classe SujetRel correspond à l'arbre du sujet en position relative. Le fragment d'arbre correspondant étant :

```
class SujetRel
export ?Y ?Z
declare ?X ?Y ?Z ?U ?V
{ <syn>{
        node ?U [cat = n]{
                node ?V [cat = n]
                node ?X [cat = p]{
                        node ?Y (fg=suj)[cat=n]
                        node ?Z [cat = v]
                        }
                }
        }
}
```



Il ne nous reste plus alors qu'à définir les classes Actif et VerbeTransitif. La classe Actif contient deux noeuds x, y tels que $x \rightarrow y$ et la classe VerbeTransitif se définit en suivant la règle donnée en (4) :

```
class VerbeTransitif
declare ?SU ?OB ?AC
{
       { ?SU = SujetCan[] | ?SU = SujetRel[] } ;
         ?OB = ObjetCan[] ; ?AC = Actif[]      ;
         ?SU.?X = ?OB.?X  ; ?SU.?Z = ?OB.?Y    ;
         ?SU.?X = ?AC.?X  ; ?SU.?Z = ?AC.?Y
}
```

Dans cette classe, on remarque la présence de conjonctions représentées par le symbole `;` et d'une disjonction représentée par `|`. On utilise la variable `SU` (respectivement `OB` et `AC`) pour désigner la structure de traits d'export de la classe SujetCan *ou* de la classe SujetRel (respectivement de la classe ObjetCan *et* de la classe Actif).

On remarque ici que nous procédons par égalité entre noeuds pour contraindre la combinaison des fragments d'arbres. Il s'agit du procédé de base. Pour passer au développement de méta-grammaires de taille importante, l'emploi d'un langage de couleur offre une meilleure flexibilité dans la définition des classe et permet d'alléger la gestion des noms de variables.

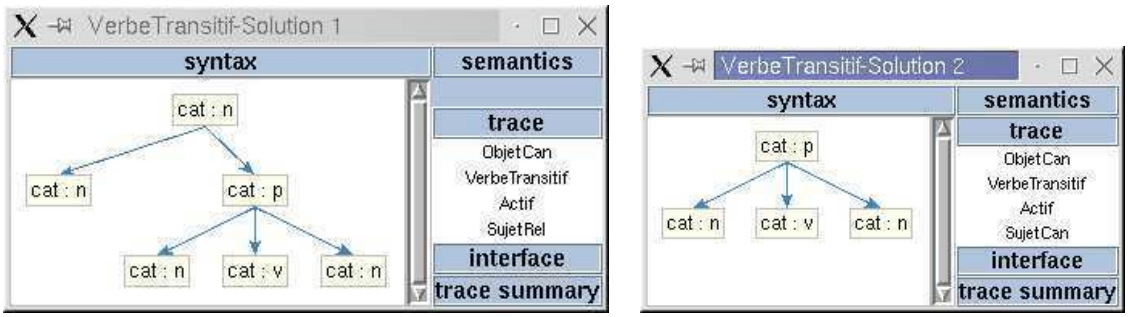


FIG. 1 – Arbres TAG produits par XMG

Les évaluations Une fois les classes de notre méta-grammaire écrites, nous pouvons demander à XMG de calculer tous les arbres TAG correspondants. Cela se fait en demandant l'*évaluation* de la classe : `value VerbeTransitif`

Résultat La classe VerbeTransitif génère deux solutions, qui correspondent aux arbres de la figure 1.

4 Intégration dans une chaîne de TALN

Le système XMG produit des grammaires au format XML[7] utilisables pour l'analyse syntaxique ou la génération :

Analyse syntaxique Nous avons vu en section 1 que l'un des buts de la production automatique de grammaires était d'éviter les problèmes liés à la redondance entre structures syntaxiques. Dans le cas de TAG, on évite aussi cette redondance en utilisant un lexique à 3 niveaux : lemmes, formes fléchies et structures syntaxiques. Dans ce contexte, la méta-grammaire produit non pas des arbres qui seraient associés chacun à une ou plusieurs unité(s) lexicale(s), mais un ensemble de structures syntaxiques *non-ancrées* L'association structure syntaxique – unité(s) lexicale(s) est alors réalisée par l'analyseur syntaxique. Pour cela, il est nécessaire d'ajouter à ces structures non-ancrées certaines informations sur la morphologie des unités lexicales qui peuvent être accueillies. Ces informations prennent la forme de structures de traits (voir (Crabbé, Gaiffe et Roussanaly, 2003)). XMG, par sa dimension *dyn* contenant des matrices attributs-valeurs, offre un support adéquat pour cette association.

Génération L'utilisation de grammaires produites automatiquement en génération pose des problèmes différents. Un générateur aura pour but de produire un ensemble de phrases à partir d'une représentation sémantique. Là aussi, pour éviter la redondance, il y a un découpage structures syntaxiques – unité(s) lexicale(s). Par contre, leur association s'accompagne de l'instanciation des arguments du prédicat sémantique dans l'arbre. Pour pouvoir accéder à ces arguments dans la structure syntaxique, nous pouvons utiliser la dimension *dyn* et la coindexation entre traits (voir (Gardent et Kow, 2004)).

[7] les DTD pour grammaires TAG et IG sont fournies dans le système XMG disponible librement à l'adresse `http://sourcesup.cru.fr/xmg`.

Conclusion

XMG offre un cadre de travail adapté au développement de grammaires de taille relativement importante (voir (Crabbé, 2005a)). Une grammaire TAG du Français à large couverture a ainsi pu être développée par B. Crabbé (Crabbé, 2005b), celle-ci est en cours d'évaluation en analyse syntaxique sur la suite de tests TSNLP. Les premiers résultats sont encourageants, la couverture de la grammaire produite étant supérieure à 75%. Pour donner un ordre d'idée, cette méta-grammaire contient 246 classes, représentant 55 familles (*ie* cadres de sous-catégorisation), et générant 5075 arbres non-ancrés en 20 minutes de compilation (sur Pentium 4 - 2,66 Ghz et 1 Go de mémoire vive).

Nous travaillons à l'heure actuelle à la production de grammaires à portée sémantique dans une optique d'analyse syntaxique combinée avec un calcul sémantique. Nous visons également l'intégration à XMG d'une bibliothèque de dimensions ayant chacune un langage de représentation propre. Cette bibliothèque a pour but d'offrir à l'utilisateur des outils adaptés lui permettant de créer des instances de méta-grammaire pour un formalisme cible arbitraire.

Références

Abeillé A. (2002), Une grammaire électronique du français , *CNRS Editions, Paris.*

Ait-Kaci H. (1991) , Warren's Abstract Machine : A Tutorial Reconstruction, *Logic Programming : Proc. of the Eighth International Conference*, K. Furukawa , MIT Press, Cambridge, MA.

Candito M.H. (1996), A principle-based hierarchical representation of LTAGs , *Proceedings of the 15th International Conference on Computational Linguistics (COLING'96), Kopenhagen.*

Clément L., Kinyon A. (2003), Generating LFGs with a MetaGrammar , *Proceedings of the 8th International Lexical Functional Grammar Conference, Saratoga Springs, NY.*

Crabbé B. (2005a), Grammatical development with XMG, *Fifth International Conference on Logical Aspects of Computational Linguistics (LACL05), Bordeaux.*

Crabbé B. (2005b), Représentation informatique de grammaires fortement lexicalisées - Application à la grammaire d'arbres adjoints, *Thèse de Doctorat (á paraître), Université Nancy 2.*

Crabbé B., Gaiffe B., Roussanaly A. (2003), Une plate-forme de conception et d'exploitation d'une grammaire d'arbres adjoints lexicalisés, *Actes de la conférence TALN'2003 Batz-sur-mer.*

Crabbé B., Duchier D. (2004), Metagrammar Redux , *International Workshop on Constraint Solving and Language Processing - CSLP 2004, Copenhagen.*

Duchier D., Niehren J. (2000), Dominance Constraints with Set Operators, *Proceedings of the First International Conference on Computational Logic (CL2000).*

Gaiffe B., Crabbé B., Roussanaly A. (2002), A New Metagrammar Compiler , *Proceedings of the 6th International Workshop on Tree Adjoining Grammars and Related Frameworks (TAG+6), Venice.*

Gardent C., Kallmeyer L. (2003), Semantic construction in FTAG , *Proceedings of the 10th meeting of the European Chapter of the Association for Computational Linguistics, Budapest.*

Gardent C., Kow E. (2004) , Génération et sélection de paraphrases grammaticales , *journée ATALA sur la génération de Langue Naturelle, Paris.*

Joshi A., Schabes Y. (1997), Tree-Adjoining Grammars , *Handbook of Formal Languages* , G. Rozenberg and A. Salomaa , Springer, Berlin, New York.

Van Roy P. (1990), Extended DCG Notation : A Tool for Applicative Programming in Prolog, *Technical Report UCB/CSD 90/583, Computer Science Division, UC Berkeley.*

Grammaire d'Unification Sens-Texte : modularité et polarisation

Sylvain Kahane (1), François Lareau (2)

(1) Modyco, U. Paris 10 / Lattice, U. Paris 7
sk@ccr.jussieu.fr
(2) OLST, U. de Montréal / Lattice, U. Paris 7
francois.lareau@umontreal.ca

Résumé – Abstract

L'objectif de cet article est de présenter l'état actuel du modèle de la Grammaire d'Unification Sens-Texte, notamment depuis que les bases formelles du modèle ont été éclaircies grâce au développement des Grammaires d'Unification Polarisées. L'accent est mis sur l'architecture du modèle et le rôle de la polarisation dans l'articulation des différents modules — l'interface sémantique-syntaxe, l'interface syntaxe-morphotopologie et les grammaires décrivant les différents niveaux de représentation. Nous étudions comment les procédures d'analyse et de génération sont contrôlables par différentes stratégies de neutralisation des différentes polarités.

This article presents the Meaning-Text Unification Grammar's current state, now that its formal foundations have been clarified with the development of Polarized Unification Grammars. Emphasis is put on the model's architecture and the role of polarization in linking the various modules — semantic-syntax interface, syntax-morphotopology interface and the well-formedness grammars of each representation level. We discuss how various polarity neutralization strategies can control different analysis and generation procedures.

Mots Clés – Keywords

Théorie Sens-Texte, interface syntaxe-sémantique, synchronisation, grammaire d'unification polarisée, grammaire de dépendance, grammaire topologique, génération de textes.

Meaning-Text Theory, syntax-semantics interface, synchronization, polarized unification grammar, dependency grammar, topology grammar, text generation.

1 Introduction

L'objectif de cet article est de présenter l'état actuel du modèle de la Grammaire d'Unification Sens-Texte [GUST], notamment depuis que ses bases formelles ont été éclaircies grâce au développement des Grammaires d'Unification Polarisées [GUP]. Le

formalisme des GUP permet de simuler élégamment la plupart des formalismes basés sur la combinaison de structures, notamment les grammaires de réécriture, TAG, LFG ou HPSG (Kahane 2004). Toutefois, GUP a été initialement développé pour donner une assise formelle solide à GUST, mais la formalisation de GUST en GUP n'a encore jamais été présentée.

L'architecture de GUST est basée sur la théorie Sens-Texte (Mel'cuk 1997 ; Kahane 2001). Plusieurs niveaux de représentation de la phrase sont considérés : sémantique, syntaxique, morphotopologique et phonologique. Ces niveaux sont bien séparés et chacun possède sa grammaire propre. La structure sémantique est un graphe, la structure syntaxique un arbre de dépendance, la structure morphotopologique un arbre ordonné, et la structure phonologique une chaîne linéaire. Sur ces structures peuvent s'ajouter d'autres structures comme une hiérarchie logique ou une structuration informationnelle au niveau sémantique ou encore une structure prosodique au niveau phonologique, mais nous n'en parlerons pas ici. De plus, ces niveaux sont ordonnés, du plus profond au plus proche de la surface, ce qui nous donne trois modules d'interface pour passer du sens au « texte » : sémantique-syntaxe, syntaxe-morphotopologie et morphotopologie-phonologie.

L'implémentation de GUST nécessite donc un formalisme capable de manipuler différents types de structures (graphe, arbre, arbre ordonné, chaîne) et de pouvoir les apparier. Comme on le verra, le formalisme des GUP est un formalisme mathématique spécialement conçu pour manipuler aussi simplement que possible de telles structures. Par ailleurs, GUP contrôle la saturation des structures qu'il combine par une polarisation de leurs objets : on peut comparer la construction d'une phrase en GUP à la formation d'une molécule en chimie par la neutralisation de la valence des atomes.

Notre implémentation de GUST fait un grand usage de la polarisation en contrôlant la construction des objets (et donc la saturation des structures) par une polarité propre à chaque module, que ce soit une grammaire de bonne formation des structures d'un niveau donné ou une grammaire d'interface entre deux niveaux. De plus, chaque module utilise la polarité des modules adjacents pour son articulation avec eux. L'ordre dans lequel nous neutraliserons ces différentes polarités va décider d'une procédure particulière en analyse ou en génération. Les procédures en largeur résultent de la neutralisation successive de toutes les polarités propres à un module (neutralisation de toute la structure sémantique, puis syntaxique, etc), tandis que les procédures en profondeur résultent d'une neutralisation en cascade de toutes les polarités introduites par un objet, c'est-à-dire que dès qu'un objet est construit à un niveau donné, on cherche à neutraliser les polarités des autres niveaux associées à cet objet plutôt que de construire d'autres objets du même niveau. Malgré une architecture stratifiée (séparation des niveaux et des interfaces), notre modèle peut donc tout à fait gérer une interaction complexe entre les différents modules et simuler une analyse ou une génération incrémentale (tentative de neutralisation de la structure du premier mot de la phrase du niveau phonologique jusqu'au niveau sémantique, puis du deuxième mot, etc).

La définition de GUP sera brièvement rappelée dans la section 2. Dans la section 3, nous proposerons des grammaires de bonne formation pour les différents types de structures que nous considérons dans GUST. Dans la section 4, nous introduirons la notion de grammaire de correspondance et nous montrerons comment GUP permet d'écrire des interfaces.

2 Grammaires d'unification polarisées [GUP]

Les grammaires d'unification polarisées sont des grammaires permettant de générer des ensembles de structures finies. Une structure repose sur des *objets*. Par exemple, pour un graphe (orienté), les objets sont des nœuds et des arcs. Chaque arc est lié à deux nœuds par les fonctions *source* et *cible*. Ce sont ces *fonctions* qui fournissent la structure proprement dite.

Une *structure polarisée* est une structure dont les objets sont polarisés, c'est-à-dire associés par une fonction à une valeur appartenant à un ensemble fini P de *polarités*. L'ensemble P est muni d'une opération commutative et associative notée « . », appelée *produit*. Un sous-ensemble N de P contient les polarités dites *neutres*. Une structure polarisée est dite *neutre* si tous les objets de cette structure sont neutres. Nous utiliserons ici un système de polarités P = {●,○,●} (que nous appellerons ainsi : ● = noire = saturée, ○ = blanche = contexte obligatoire et ● = grise = neutre absolu), avec N = {●,●}, et un produit défini par le tableau suivant (où ⊥ représente l'impossibilité de se combiner) :

·	●	○	●
●	●	○	●
○	○	○	●
●	●	●	⊥

Tableau 1 — Le produit des polarités

Les structures peuvent être combinées par *unification*. L'unification de deux structures A et B donne une nouvelle structure A⊕B obtenue en « collant » ensemble ces structures par l'identification d'une partie des objets de A avec une partie de ceux de B. Lorsque A et B sont unifiées, la polarité d'un objet de A⊕B obtenu par identification d'un objet de A et d'un objet de B est le produit de leurs polarités. Toutes les fonctions associées aux objets unifiés sont nécessairement identifiées (comme le sont les traits quand on unifie deux structures de traits).

Une *grammaire d'unification polarisée* (GUP) est définie par une famille finie T de types d'objets (avec des fonctions associées aux différents types d'objets), un système (P,.) de polarités, un sous-ensemble N de P de polarités neutres, et un ensemble fini de structures élémentaires polarisées, dont les objets sont décrits par T et dont une peut être marquée comme la structure initiale. Les structures *générées* par la grammaire sont les structures neutres obtenues par combinaison de la structure initiale et/ou d'un nombre fini de structures élémentaires. Rappelons que le formalisme est monotone (avec l'ordre ● < ○ < ● sur les polarités) et que les structures peuvent être combinées dans n'importe quel ordre.

Nous verrons dans la section suivante deux exemples de GUP qui génèrent respectivement les graphes sémantiques et les arbres syntaxiques. Tous nos exemples seront illustrés par (1) :

(1) *Pierre mange deux pommes.*

3 Grammaires de bonne formation

3.1 Grammaire sémantique

Nos représentations sémantiques sont essentiellement basées sur un graphe de relations prédicat-argument entre sémantèmes (les signifiés des unités lexicales et grammaticales de la phrase). On peut superposer sur ce graphe d'autres informations pour encoder la structure informationnelle ou les relations de portée (Mel'cuk 2001, Kahane 2005). Les nœuds d'une représentation sémantique représentent les sémantèmes et les arcs représentent les relations prédicat-argument. Notre grammaire $\mathcal{G}_{\text{sém}}$ est donc une grammaire de graphe dont les objets portent chacun une polarité, notée $p_{\text{sém}}$ dans la suite[1], qui indique quel est le nœud construit par la règle (polarisé en noir) et quels sont les nœuds constituant la valence sémantique du prédicat (polarisés en blanc) (Figure 1).

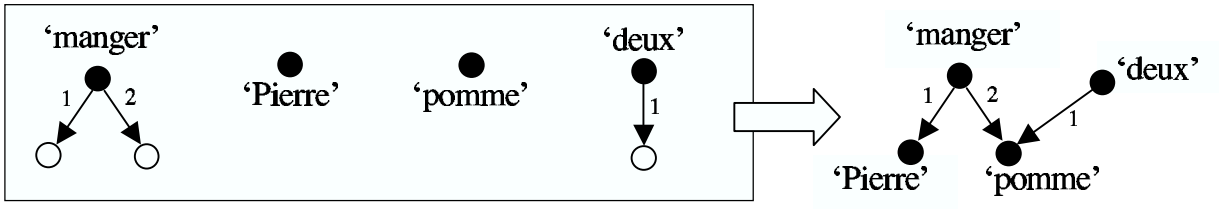


Figure 1 — Un extrait de $\mathcal{G}_{\text{sém}}$ générant le graphe sémantique de (1)

3.2 Grammaire syntaxique

Nos représentations syntaxiques sont des arbres de dépendance dont les nœuds sont étiquetés par des représentations lemmatisées de mots. Plus exactement le nœud syntaxique est étiqueté par le nom d'un lexème et des fonctions dites *grammaticales* lient ce nœud à d'autres objets représentant les grammèmes[2]. Les arcs de l'arbre sont étiquetés par des fonctions syntaxiques, (qui varient d'une langue à l'autre). Voir Figure 2 à droite la représentation syntaxique de (1).

Nous présentons notre grammaire syntaxique $\mathcal{G}_{\text{synt}}$ avec une seule polarité p_{synt}, qui comme précédemment indique quels sont les objets construits par la règle et permet de vérifier que les fonctions associées à chaque objet sont bien instanciées : la source et la cible pour les arcs, l'étiquette, la partie du discours (*pdd*) pour les nœuds et les fonctions grammaticales éventuelles (*nombre* pour les noms, *mode*, *temps, nombre* et *personne* pour les verbes, etc.)[3]. Une deuxième polarité, $p_{\text{synt-gouv}}$, masquée ici, sert à vérifier que la structure est bien un arbre, c'est-à-dire que chaque nœud est gouverné une et une seule fois, à l'exception du sommet.

---

[1] Par convention, nous faisons référence aux polarités par le nom de la fonction qui les associe aux objets qui les portent. Ainsi, à tous les objets des règles de $\mathcal{G}_{\text{sém}}$ est associée une fonction $p_{\text{sém}}$ qui retourne comme valeur une des polarités de l'ensemble P décrit plus haut.

[2] Nous indiquons ces fonctions en italiques dans nos figures, suivies du grammème en question. Les grammèmes sont eux-mêmes liés aux lexèmes par des fonctions inverses que nous masquons. Ce double lien rend les deux objets indissociables : l'unification de l'un force l'unification de l'autre.

[3] Nous présentons un système grammatical simplifié en masquant par exemple la détermination pour le nom ou la finitude et la voix pour le verbe.

La Figure 2 présente un extrait de $\mathcal{G}_{synt}$ permettant de générer l'arbre syntaxique de (1). La toute première règle est la structure initiale, qui doit être utilisée une et une seule fois et qui correspond au sommet de l'arbre (et qui devra recevoir une polarité $p_{synt\text{-}gouv}$ noire indiquant qu'il ne peut être gouverné). Les quatre règles suivantes sont des règles lexicales ; elles indiquent la partie du discours des lexèmes et les grammèmes qui leur sont nécessaires. On notera que ces règles ne contrôlent pas la valence syntaxique, qui sera contrôlée par l'interface sémantique-syntaxe. Les trois règles qui suivent sont des règles sagittales ; elles décrivent différentes relations syntaxiques possibles (et contrôle la structure d'arbre par une polarisation $p_{synt\text{-}gouv}$ noire du nœud dépendant, non indiquée ici). Les autres règles sont des règles grammaticales. Ces règles peuvent dépendre du contexte : ainsi l'indicatif nécessite une relation sujet et vice-versa, la présence d'un numéral impose le pluriel ou celle d'un sujet non pronominal la 3e personne du verbe. Notons également que l'introduction d'un grammème peut dépendre d'un autre grammème : l'indicatif par exemple exigera un temps, mais pas l'infinitif[4].

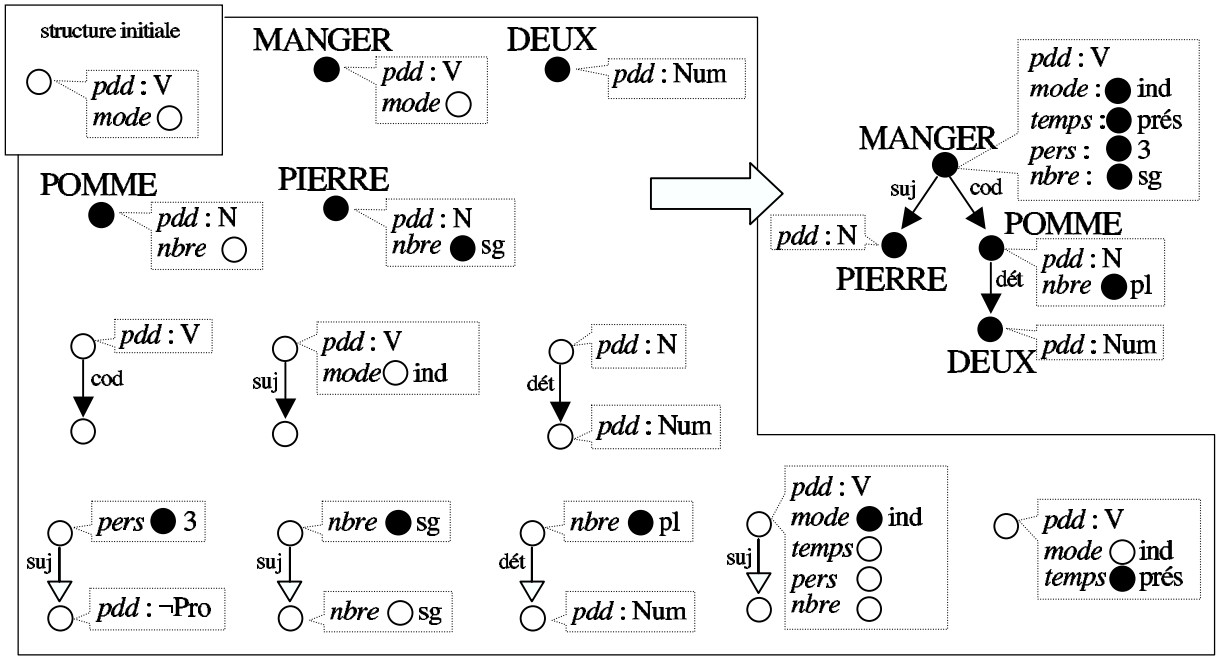


Figure 2 — Un extrait de $\mathcal{G}_{synt}$ générant la représentation syntaxique de (1)

Nous présenterons la grammaire morphotopologique en même temps que l'interface syntaxe-morphotopologie, dans la section 4.3. Nous laisserons de côté la phonologie.

4 Interfaces

Nous introduisons d'abord la notion de grammaire de correspondance avant d'introduire les deux interfaces considérées ici : sémantique-syntaxe et syntaxe-morphotopologie.

[4] Nous considérons l'opposition entre l'infinitif et l'infinitif dit « passé » d'ordre aspectuel.

4.1 Grammaires de correspondance

Une grammaire de correspondance $\mathcal{G}$ est une grammaire qui met en correspondance des structures appartenant à deux ensembles, que nous notons $\mathcal{A}$ et $\mathcal{B}$. Les règles de $\mathcal{G}$ mettent en correspondance des structures élémentaires composant les éléments de $\mathcal{A}$ et de $\mathcal{B}$. On peut considérer trois fonctionnements pour $\mathcal{G}$: équatif, transductif et génératif, selon qu'on fournit en entrée deux, une ou zéro structures (Kahane 2000). Une grammaire équative vérifie que deux structures fournies se correspondent, une grammaire transductive traduit une structure fournie en une autre, tandis qu'une grammaire générative génère un couple de structures en correspondance.

Considérons le fonctionnement équatif de $\mathcal{G}$, qui est le plus élémentaire : $\mathcal{G}$ partitionne les deux structures en un même nombre de fragments qui se correspondent deux à deux par les règles de la grammaire. En termes de polarités, cela signifie que les deux structures en entrée doivent être déclarées comme des ressources que $\mathcal{G}$ devra toutes « consommer » (assurant ainsi que les deux structures ont été totalement mises en correspondance). Les objets des deux structures recevront donc une polarité $p_{\mathcal{G}}$ blanche que les règles de $\mathcal{G}$ devront neutraliser. Les règles de $\mathcal{G}$ contiendront quant à elles des objets (des niveaux de représentation de $\mathcal{A}$ et de $\mathcal{B}$) de polarité $p_{\mathcal{G}}$ noire mis en correspondance. Ainsi, en neutralisant les deux structures fournies, $\mathcal{G}$ les met en correspondance.

Voyons maintenant comment la même grammaire $\mathcal{G}$ fonctionne de manière transductive. C'est ce mode de fonctionnement qui modélise les processus linguistiques de synthèse et d'analyse. On fournit à $\mathcal{G}$ une structure A de $\mathcal{A}$ dont tous les objets portent une polarité $p_{\mathcal{G}}$ blanche. On déclenche alors un jeu de règles afin de neutraliser A, ce qui nous construit une structure B synchronisée avec A. Une grammaire de bonne formation des structures de $\mathcal{B}$ doit maintenant vérifier que B appartient bien à $\mathcal{B}$. La structure B, tout en étant neutre pour $\mathcal{G}$ (tous ses objets portent une polarité $p_{\mathcal{G}}$ noire), doit donc déclencher la grammaire de $\mathcal{B}$. Elle doit pour cela être entièrement blanche en polarité $p_{\mathcal{B}}$. En somme, chaque module, qu'il soit une grammaire de bonne formation ou une grammaire d'interface, possède une polarité propre contrôlant la construction des objets par cette grammaire, mais pour assurer l'appel des modules adjacents, les objets des règles de $\mathcal{A}$ et de $\mathcal{B}$ reçoivent, en plus de leurs polarités respectives $p_{\mathcal{A}}$ et $p_{\mathcal{B}}$, une polarité $p_{\mathcal{G}}$ blanche, tandis qu'il faut ajouter aux règles de $\mathcal{G}$ une polarité blanche $p_{\mathcal{A}}$ pour les éléments du niveau de représentation de $\mathcal{A}$ et $p_{\mathcal{B}}$ pour les éléments du niveau de $\mathcal{B}$. Un objet construit par $\mathcal{A}$ aura donc une double polarisation $p_{\mathcal{A}}$-$p_{\mathcal{G}}$ (●,○), tandis que l'élément correspondant construit par $\mathcal{G}$ aura une double polarisation $p_{\mathcal{A}}$-$p_{\mathcal{G}}$ (○,●). Ceci nous donne un système à quatre polarités {(○,○),(○,●),(○,●),(●,●)} équivalent au système {○,–,+,●} de Bonfante *et al.* 2004 et Kahane 2004. Ces couples de polarités sont les *polarités d'articulation*.

4.2 Interface sémantique-syntaxe

L'interface sémantique-syntaxe $\mathcal{I}_{\text{sém-synt}}$ est une grammaire de correspondance entre l'ensemble des graphes sémantiques décrit par $\mathcal{G}_{\text{sém}}$ et l'ensemble des arbres syntaxiques décrit par $\mathcal{G}_{\text{synt}}$.

Nous allons illustrer la présentation de $\mathcal{I}_{\text{sém-synt}}$ par son fonctionnement transductif en synthèse. On fournit à la grammaire un graphe sémantique comme celui de la Figure 1 à droite, à la différence que chacun des objets reçoit maintenant une double polarisation $p_{\text{sém}}$-$p_{\text{sém-synt}}$(●,○). Les objets des règles de $\mathcal{I}_{\text{sém-synt}}$ portent tous une polarité $p_{\text{sém-synt}}$ (blanche ou noire, selon qu'ils sont ou non construits par la règle en question), mais ceux du niveau sémantique (resp. syntaxique) auront en plus une polarité $p_{\text{sém}}$ (resp. p_{synt}) blanche. La Figure 3 présente un extrait de la grammaire d'interface qui traite le graphe sémantique de (1). Seule la polarité $p_{\text{sém-synt}}$ est représentée.

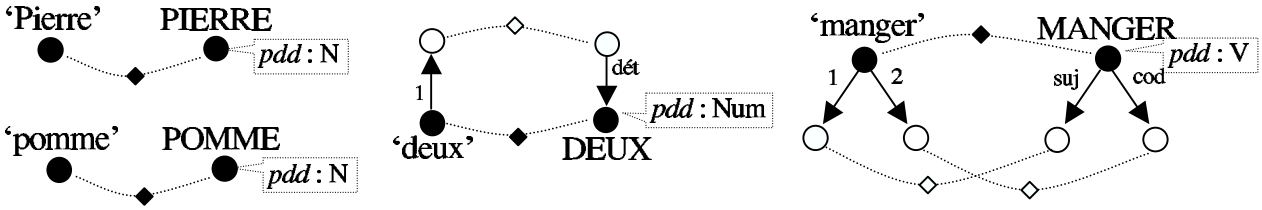


Figure 3 — Un extrait de la grammaire d'interface $\mathcal{I}_{\text{sém-synt}}$

Dans la Figure 4, nous donnons trois structures : le graphe sémantique de (1) avec la double polarisation $p_{\text{sém}}$-$p_{\text{sém-synt}}$, le résultat après neutralisation par $\mathcal{I}_{\text{sém-synt}}$ de toutes les polarités qui pouvaient l'être (c'est-à-dire $p_{\text{sém-synt}}$ et $p_{\text{sém}}$, mais pas p_{synt}), puis le résultat après application de $\mathcal{G}_{\text{synt}}$ sur l'arbre syntaxique construit par $\mathcal{I}_{\text{sém-synt}}$. Les polarités qui servent à l'articulation avec le module adjacent sont indiquées en décalé. A noter que les grammèmes d'accord (comme le *nbre* et *pers* pour le verbe) n'ont pas de contrepartie sémantique directe et ne doivent donc pas porter de polarité $p_{\text{sém-synt}}$.

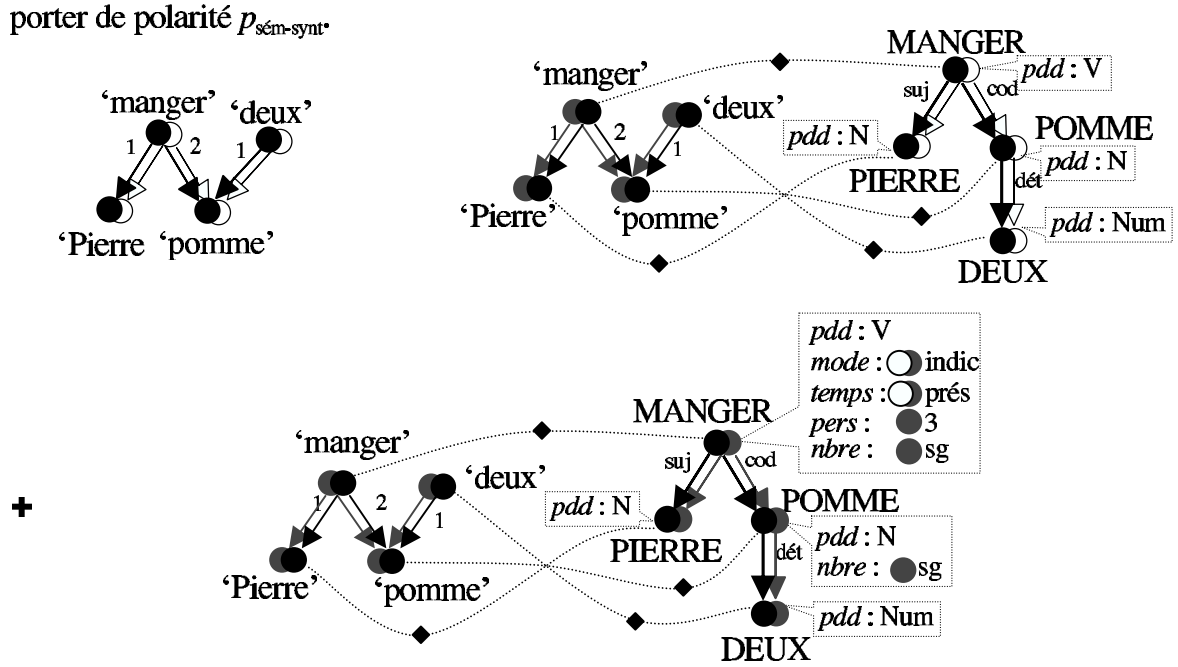


Figure 4 — Le graphe sémantique de (1) avant et après l'application de $\mathcal{I}_{\text{sém-synt}}$ et $\mathcal{G}_{\text{synt}}$

De même que $\mathcal{G}_{\text{sém}}$, la grammaire syntaxique $\mathcal{G}_{\text{synt}}$ doit être enrichie pour assurer l'articulation avec à la fois l'interface sémantique-syntaxe et l'interface syntaxe-morphotopologie. Comme on peut le voir dans de bas de la Figure 4, l'application de $\mathcal{G}_{\text{synt}}$ sur la sortie de $\mathcal{I}_{\text{sém-synt}}$ permet de neutraliser toutes les polarités p_{synt} des objet syntaxiques construits par $\mathcal{I}_{\text{sém-synt}}$. Cependant, il n'est pas possible d'introduire le lexème MANGER sans lui attribuer un mode (*cf.* Figure 2)

et en lui attribuant le mode indicatif, par exemple, on force l'utilisation d'un grammème de temps et de grammèmes d'accord en personne et en nombre. La grammaire $\mathcal{G}_{\text{synt}}$, en neutralisant les polarités p_{synt} de la sortie de $\mathcal{I}_{\text{sém-synt}}$, introduit donc des grammèmes qui ne figuraient pas dans cette sortie et qui pour certains (pas les grammèmes d'accord) auront une polarité $p_{\text{sém-synt}}$ blanche (Figure 4 en bas). Cette polarité blanche sur les grammèmes de mode et de temps ainsi introduits devra donc être neutralisée par des règles de $\mathcal{I}_{\text{sém-synt}}$. Or, rien dans la structure sémantique de départ ne peut correspondre à ces grammèmes. Qu'à cela ne tienne ! Le jeu des polarités de GUST permet de modifier en cours de route la représentation donnée en entrée afin de satisfaire des contraintes provenant d'autres niveaux de représentation. Supposons que notre modèle comprenne les deux règles données en Figure 5. La première est une règle de $\mathcal{G}_{\text{sém}}$ qui donne une représentation d'un des sens du temps présent. La seconde est une règle de $\mathcal{I}_{\text{sém-synt}}$ qui fait correspondre à ce sens le grammème de temps présent. Ces deux règles peuvent s'appliquer dans le sens inverse de la synthèse qui est en cours pour venir modifier la structure de départ en lui ajoutant le sens du présent (nous laissons de côté la question du mode pour des raisons de clarté). On obtient alors la structure sémantique à droite.

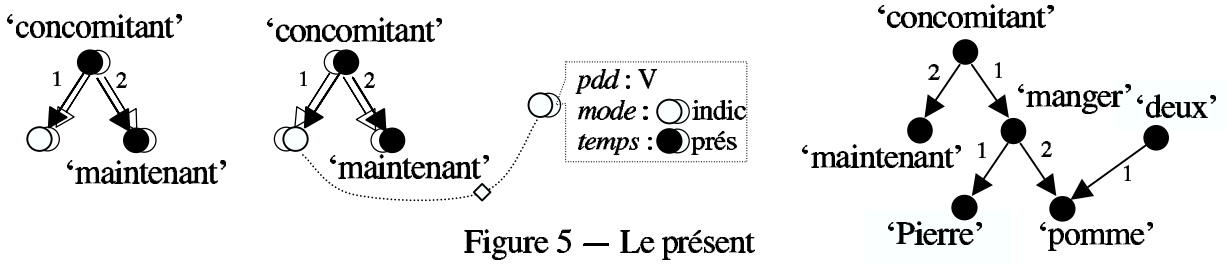


Figure 5 — Le présent

Bien entendu, il existe d'autres règles qui peuvent saturer les besoins en mode et en temps, et on obtiendra autant de structures sémantiques modifiées qu'il y a de façons de satisfaire les contraintes syntaxiques. Dans le cadre d'un système de génération de texte, il faut alors que le programme choisisse, parmi les graphes sémantiques proposés, celui qui convient le mieux selon les connaissances du monde auxquelles il a accès. Dans le cas qui nous préoccupe, le système aura ainsi à répondre à la question « Quand Pierre mange-t-il deux pommes : est-ce dans le passé, dans le présent ou dans le futur ? ». Ce mécanisme modélise bien le fait que les sens grammaticaux ne sont pas nécessairement exprimés parce qu'ils répondent à un besoin communicatif de la part du locuteur, mais plutôt parce qu'ils sont imposés par la langue (voir à ce sujet Polguère 1998). Ce type de mécanisme, en plus d'être plausible d'un point de vue cognitif, pourrait s'avérer particulièrement utile dans le cadre de la traduction automatique, où la représentation sémantique obtenue par analyse dans la langue source n'est pas forcément exprimable dans la langue cible. Il peut également opérer en situation d'analyse. C'est alors la chaîne sonore qui est donnée en entrée et qui peut être modifiée si elle n'est pas tout à fait bien formée, ce qui permettrait de contourner certains problèmes liés au bruit ou aux erreurs de performance. Ces hypothèses n'ont toutefois pas encore été validées par expérimentation.

4.3 Interface syntaxe-morphotopologie

Nous allons simplifier la présentation de l'interface syntaxe-morphotopologie au maximum en ne traitant pas réellement la morphologie (nous associons directement un lexème et les grammèmes qui lui correspondent à une forme fléchie) et en adoptant une structure topologique conventionnelle du type de celle des grammaires syntagmatiques (*cf.* Gerdes &

Kahane 2004 pour une étude de la topologie du français et Gerdes & Kahane 2001 pour une modélisation formelle). La grammaire de bonne formation topologique est donc équivalente à une grammaire de réécriture hors-contexte (*cf.* Kahane 2004 pour la simulation en GUP des grammaires de réécriture) : elle construit un arbre de constituants topologiques et son principal objectif est d'indiquer dans quel ordre se trouvent les constituants frères (*cf.* règle R1 équivalente à S → GN GV)[5]. L'interface syntaxe-morphotopologie contient trois types de règles : des règles pour le sous-constituant tête (*cf.* règle R2 disant que GV est la tête de S et que donc GV et S sont synchronisés avec le même nœud syntaxique), des règles pour le placement d'un dépendant (*cf.* règle R3 disant que le sujet correspond au GN sous S) et des règles de « morphologie » (*cf.* règle R4 disant que *mange* ⇔ $\text{MANGER}_{\text{ind, prés, 3, sg}}$). Cette interface est assez proche d'une grammaire LFG (Bresnan 1999), assurant comme elle la synchronisation d'une structure de constituants avec une structure de dépendance, mais la polarisation permet d'expliciter ce que chaque règle construit réellement et de découper davantage l'information (par exemple, nos trois premières règles correspondent à une unique règle LFG : S → GN[↓=↑Subj] GV[↓=↑]). Dans la Figure 6, nous n'indiquons que la polarité propre à chaque module (p_{topo} dans R1 et $p_{\text{synt-topo}}$ dans les autres).

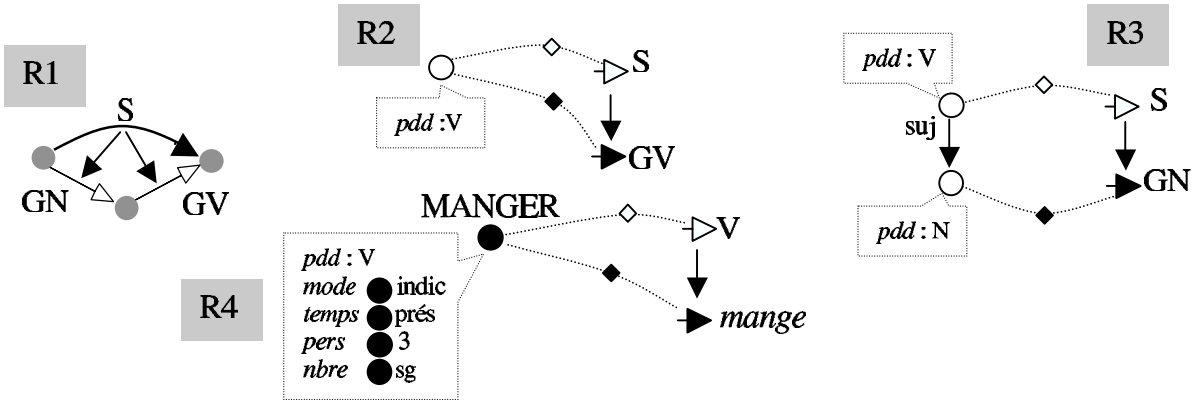


Figure 6 — Morphotopologie

5 Conclusion

Notre article porte essentiellement sur l'architecture d'un modèle linguistique. Nous avons proposé un modèle modulaire (3 grammaires de bonne formation et 2 grammaires d'interface dans la version simplifiée présentée ici) permettant de gérer différents niveaux d'organisation et différents types de structures (graphe, arbre, arbre ordonné) et pourtant nous utilisons un unique formalisme. Cela nous permet de combiner les règles dans n'importe quel ordre et permet aux différents modules d'interagir constamment. De plus, la saturation des structures ainsi que l'interaction des modules sont soigneusement contrôlées grâce à la polarisation (généralement multiple) de tous les objets. Chaque polarité ayant un rôle bien déterminé, nous pouvons diriger l'application du modèle. On peut ainsi choisir entre une procédure en largeur et une procédure en profondeur : en privilégiant la neutralisation de tous les objets portant une

[5] Les nœuds gris dans la règle R1 servent à indiquer le début et la fin des constituants. Il reçoivent une polarité p_{topo} grise, c'est-à-dire qu'ils sont absolument neutres et peuvent se combiner librement à chaque fois qu'une unification d'arc l'oblige. Ils n'auront pas de polarité $p_{\text{synt-topo}}$, puisqu'ils n'ont pas de correspondant au niveau syntaxique. A noter que, comme pour le niveau syntaxique, une deuxième polarité, non considérée ici, est nécessaire pour assurer la structure d'arbre et vérifier que chaque « arc » est gouverné.

polarité propre à un module, les modules s'appliqueront les uns à la suite des autres, tandis qu'en privilégiant la neutralisation de toutes les polarités associées à un objet, chaque élément de la structure d'entrée sera traité du sens au texte (ou vice-versa). Il est important de noter que la procédure en profondeur n'est possible que parce qu'un même formalisme a été utilisé pour l'ensemble des modules. Nous avons également vu que, grâce à l'interaction étroite des différents modules, une GUST peut accepter en entrée une représentation sémantique sous-spécifiée où tous les sens flexionnels sont absents, et générer quand même des phrases bien formées en ajoutant à la représentation initiale les sens flexionnels qu'impose la langue. Enfin, comme l'ont montré Bonfante *et al.* (2004), la polarisation permet un filtrage efficace des règles réduisant les calculs sans issue.

Le développement d'un modèle du français est en cours. Le noyau central de la topologie du français est décrit par Gerdes & Kahane (2004), la syntaxe des temps du français est décrite par Lareau (2004), d'autres points de l'interface sémantique-syntaxe (les distorsions entre syntaxe et sémantique et les questions de portée des quantificateurs) sont étudiés par Kahane (2003, 2005). L'implémentation de GUST dans le formalisme GUP est également en cours. Mais davantage que la couverture, l'accent est mis sur la propreté du traitement théorique et l'adéquation de la formalisation.

Références

BRESNAN J. (1999), *Lexical-Functional Syntax*, Blackwell.

BONFANTE G., GUILLAUME B. & PERRIER G. (2004), Polarization and abstraction of grammatical formalisms as methods for lexical disambiguation, *Actes CoLing,* Genève, 303-309.

GERDES K. & KAHANE S. (2001), Word order in German: A formal dependency grammar using a topological hierarchy, *Actes ACL*, Toulouse, 220-227.

GERDES K. & KAHANE S. (2004), L'amas verbal au cœur d'une modélisation topologique du français, *Actes Journées de la syntaxe – Ordre des mots dans la phrase française, positions et topologie*, Bordeaux, 8 p.

KAHANE S. (2000), Des grammaires formelles pour définir une correspondance, *Actes TALN*, Lausanne, 197-206.

KAHANE S. (2001), Grammaires de dépendance formelles et théorie Sens-Texte, Tutoriel, *Actes TALN, vol. 2*, 17-76.

KAHANE S. (2004), Grammaires d'unification polarisées, *Actes TALN*, Fès, 233-242.

KAHANE S. (2005), Structure des représentations logiques, polarisation et sous-spécification, *Actes TALN*, Dourdan, 10 p.

LAREAU F. (2004), *Vers un modèle formel de la conjugaison française dans le cadre des grammaires d'unification Sens-Texte polarisées*, Document pour l'examen général de synthèse, Montréal, Université de Montréal.

MEL'CUK I. (1997), *Vers une linguistique sens-texte : leçon inaugurale*, Paris, Collège de France.

MEL'CUK I. (2001), *Communicative Organisation of Natural Language*, Benjamins.

NASR A. (1995), A formalism and a parser for lexicalised dependency grammars, *4th Int. Workshop on Parsing Technologies*, State University of New York Press.

POLGUERE A. (1998), Pour un modèle stratifié de la lexicalisation en génération de texte, *TAL, 39:2*, 57-76.

SHIEBER S. M. & SCHABES Y. (1990), Synchronous tree-adjoining grammars, *Proceedings of the 13th Int. Conference on Computational Linguistics*, vol. 3, 253-258, Helsinki, Finland.

Indexation sémantique au moyen de coupes de redondance minimale dans une ontologie

Florian Seydoux & Jean-Cédric Chappelier
Faculté Informatique et Communications
Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH–1015 Lausanne, Suisse
{florian.seydoux,jean-cedric.chappelier}@epfl.ch

Mots-clefs : Indexation sémantique, Recherche documentaire, Redondance minimale, Ontologie.

Keywords: Semantic Indexing, Information Retrieval, Minimal Redundancy, Ontology.

Résumé Plusieurs travaux antérieurs ont fait état de l'amélioration possible des performances des systèmes de recherche documentaire grace à l'utilisation d'indexation sémantique utilisant une ontologie (p.ex. WordNet). La présente contribution décrit une nouvelle méthode visant à réduire le nombre de termes d'indexation utilisés dans une indexation sémantique, en cherchant la coupe de redondance minimale dans la hiérarchie fournie par l'ontologie. Les résultats, obtenus sur diverses collections de documents en utilisant le dictionnaire EDR, sont présentés.

Abstract Several former works have shown that it is possible to improve information retrieval performances using semantic indexing, adding additional information coming from a thesaurus (e.g. WordNet). This paper presents a new method to reduce the number of "concepts" used to index the documents, by determining a minimum redundancy cut in the hierarchy provided by the thesaurus. The results of experiments carried out on several standard document collections using the EDR thesaurus are presented.

1 Introduction

L'utilisation de connaissances sémantiques dans le cadre de la Recherche Documentaire (RD) n'est pas nouvelle. On voit se dégager dans la littérature scientifique principalement trois champs d'application : *l'expansion de requêtes* (Voorhees, 1994; Moldovan & Mihalcea, 2000), *la désambiguïsation sémantique (WSD)* (Ide & Véronis, 1998; Wilks & Stevenson, 1998) et *l'indexation sémantique*. C'est dans ce dernier cadre que se situe le travail présenté ici.

L'indexation sémantique consiste à utiliser, pour indexer des documents, le(s) sens des mots qu'ils contiennent, au lieu ou en plus des mots[1] eux-mêmes comme c'est le cas en RD classique,

Ce travail a été financé par le projet n°200020–103529 du Fond National Suisse pour la Recherche Scientifique.

[1] Habituellement, leurs lemmes ou leurs racines (*stems*).

de manière à améliorer tant le rappel (par le biais des relations de synonymie) que la précision (en traitant correctement les cas d'homographie/polysémie).

Les différentes expériences rapportées à ce sujet dans la littérature font cependant état de résultats peu concluants, parfois même contradictoires : si certains observent que l'ajout de ce type d'information, réalisée de manière automatique, dégrade les performances de leur système (Salton, 1968; Harman, 1988; Voorhees, 1993; Voorhees, 1998), pour d'autres au contraire une amélioration significative est obtenue (Richardson & Smeaton, 1995; Smeaton & Quigley, 1996; Gonzalo *et al.*, 1998a; Gonzalo *et al.*, 1998b; Mihalcea & Moldovan, 2000).

Bien qu'il semble souhaitable pour un système de RD de prendre en compte un maximum d'informations, en particulier des informations de nature sémantique, un tel accroissement des termes d'indexation peut se révéler contre-productif, ou tout du moins ne pas développer son plein potentiel. En effet, une forte augmentation du nombre de termes d'indexation a non seulement comme conséquences de prolonger notablement les temps de traitement, mais surtout affecte les performances sur le plan de la précision : tenter de discriminer quelques documents parmi un ensemble sur la base d'un très grand nombre de critères est difficile à réaliser, la « distance » – généralement une similarité ou une dissemblance – entre chaque paire de documents tendant à devenir à peu près la même (effet « *curse of dimensionality* »).

Ce problème n'est pas nouveau et il existe déjà un certain nombre de techniques visant à limiter la taille du jeu d'indexation : en plus de celles procédant par filtrage (en utilisant par exemple un anti-dictionnaire (*stoplist*), la catégorie morpho-syntaxique, ou encore les fréquences d'occurrence), la limitation du nombre de termes d'indexation a aussi été envisagée au moyen de techniques statistiques issues de l'analyse des données (analyse en composantes principales, analyse factorielle discriminante) (Deerwester *et al.*, 1990; Hofmann, 1999). Cependant, la plupart de ces techniques ne sont pas nécessairement adaptées lorsque l'on est en présence d'informations supplémentaires sur les termes d'indexation ayant une structure formelle (au lieu de statistique). L'objectif des travaux présentés dans cette contribution est précisément d'utiliser une ressource sémantique externe (i.e. additionnelle aux données de recherche documentaire proprement dites) structurée, de type ontologie, en vue d'augmenter la richesse de l'indexation. La spécificité de ce travail par rapport à des travaux antérieurs similaires, qui utilisent des « *synsets* » ou des hyperonymes de *WordNet* comme termes d'indexation (Gonzalo *et al.*, 1998a; Gonzalo *et al.*, 1998b; Whaley, 1999; Mihalcea & Moldovan, 2000), est d'essayer de faire un pas supplémentaire en sélectionnant les « concepts » à utiliser comme termes d'indexation au moyen d'un critère issu de la théorie de l'information, la *Coupe de Redondance Minimale* (CRM, voir figure 1), que l'on applique à la relation inclusive « est-un » (hyperonymie) obtenue ici par le biais de la taxonomie (anglaise) *EDR* (Miyoshi *et al.*, 1996).

2 Coupe de redondance minimale

2.1 Objectifs

Le choix du « concept hyperonyme »[2] à utiliser pour représenter un mot est un choix délicat : un concept trop général dégradera les performances du système en diminuant la précision, tandis

[2] Nous désignons par « concept hyperonyme » un nœud non feuille dans l'ontologie. Les feuilles de l'ontologie représentent les mots.

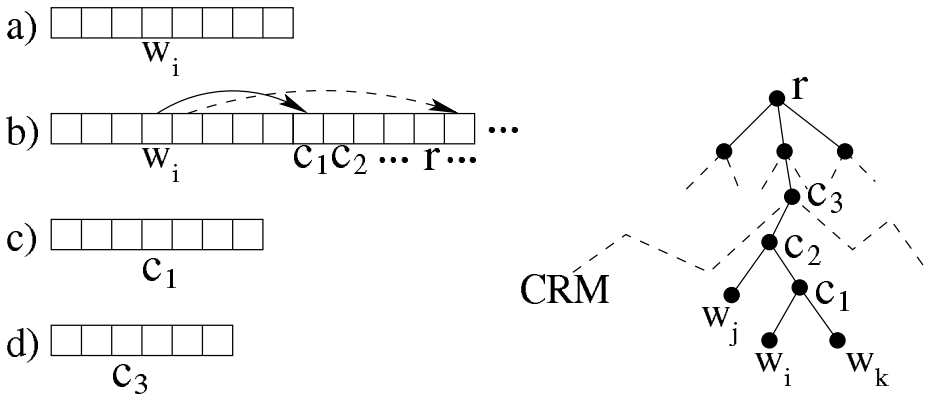


FIG. 1 – Différentes méthodes d'indexation : (a) traditionnelle, au moyen des mots, racines (*stems*) ou lemmes ; (b) utilisant une ontologie sémantique (illustration de droite), chaque terme d'indexation de (a) est augmenté par tout ou partie des « concepts » le recouvrant ; cela conduit à une explosion du nombre de termes d'indexation ; (c) indexation par les concepts de plus bas niveau ($\approx$ indexation par « synsets ») : chaque terme d'indexation est *remplacé* par son concept hyperonyme direct, factorisant ainsi tous les mots dominés par ce concept ; on réduit donc le nombre de termes d'indexation, tout en permettant de détecter la similarité entre documents contenant ces mots ; (d) indexation par une Coupe de Redondance Minimale (CRM) : chaque terme d'indexation est remplacé par l'un de ses concepts hyperonymes, déterminé par la CRM. Cela restreint d'avantage le nombre de termes d'indexation, le nombre de mots couverts (factorisés) par chacun d'eux étant plus grand qu'avec le concept hyperonyme direct.

qu'un concept trop spécifique ne permettra pas de réduire significativement le nombre de termes d'indexation et conservera la distinction entre mots de sens proches.

Pour déterminer le niveau adéquat des concepts d'indexation, nous faisons ici le choix de ne prendre en considération des coupes dans l'ontologie (une coupe étant un ensemble minimal[3] de nœuds définissant une partition sur les feuilles), en considérant que chaque nœud représente alors l'ensemble des feuilles qu'il recouvre.

Le problème est de trouver une stratégie permettant d'identifier une coupe « optimale » en un temps acceptable. Pour une tâche relativement similaire, Li (1998) propose d'utiliser le critère MDL (*Minimum Description Length*). Si ce critère est facilement calculable, il a comme inconvénient, du moins lorsque appliqué à l'ontologie EDR, de très souvent sélectionner la racine de l'ontologie comme coupe « optimale » ; ce qui n'est pas vraiment adéquat pour la tâche considérée ! Nous nous proposons donc ici d'employer un autre critère, fondé sur la théorie de l'information, permettant d'identifier une coupe pour laquelle la *redondance d'information* est minimale, c'est-à-dire une coupe qui équilibre le plus possible les degrés de description des mots factorisés en tenant compte de la probabilité d'occurrence de ces mots.

2.2 Critère de redondance minimale

Soient $\mathcal{N} = \{n_i\}$ l'ensemble des nœuds (concepts ou mots) et $\mathcal{W}$ l'ensemble des feuilles (mots uniquement) contenus dans l'ontologie considérée. On définit alors une coupe Γ comme un sous-ensemble minimal[3] de $\mathcal{N}$ recouvrant $\mathcal{W}$. Une coupe probabilisée $M = (\Gamma, P)$ est une

[3] Par « minimal », on entend qu'aucun nœud de la coupe ne peut en être retiré sans en diminuer la couverture.

paire composée d'une coupe Γ et d'une distribution de probabilités P sur Γ. On notera $|\Gamma|$ le nombre de nœuds de la coupe (et par extension: $|M| = |\Gamma|$).

Dans la suite, nous considérons la coupe $M = (\Gamma, P_f)$ probabilisée par les fréquences d'occurrences des mots correspondant aux feuilles de l'ontologie : $P_f(n_i) = f(n_i)/|D|$, où $f(n_i)$ représente le nombre d'occurrences du concept (ou mot) n_i dans les données D. Pour calculer $f(n_i)$, on admet qu'il y a occurrence de n_i lorsqu'il y a occurrence de l'un des $w_i \in {n_i}^{++}$ mots hyponymes de n_i, où n^{++} représente la fermeture transitive de n^+, ensemble des successeurs de n.

La redondance $R(M)$ d'une coupe probabilisée $M = (\Gamma, P)$ est définie par (Shannon, 1948):

$$R(M) = 1 - \frac{H(M)}{\log|M|}, \quad \text{avec} \qquad H(M) = -\sum_{n\in\Gamma} P(n) \cdot \log P(n).$$

Minimiser la redondance revient à maximiser le rapport entre l'entropie des éléments de la coupe et sa valeur maximale possible ($\log|M|$) ; le but est donc de trouver une coupe probabilisée M qui maximise le critère $\mathcal{C}_H$:

$$\mathcal{C}_H = \begin{cases} 0 & \text{si } |M| \leq 1, \\ \frac{H(M)}{\log|M|} & \text{sinon.} \end{cases}$$

Un tel critère pose cependant quelques difficultés en pratique: d'une part, il ne permet pas d'identifier une coupe optimale unique, mais un *ensemble* de coupes possibles ; d'autre part, l'optimum local sur une partie de l'ontologie est conditionné par l'optimum sur le reste (et inversement). Pour identifier les modèles satisfaisant le critère global, il faudrait donc le calculer pour l'ensemble des coupes possibles.

La première difficulté peut être surmontée de manière relativement aisée, par exemple en ne retenant qu'une coupe choisie au hasard, ou en favorisant celles admettant le plus de nœuds, ou encore en guidant le choix selon la profondeur moyenne des nœuds.

Pour être calculable, la seconde difficulté implique par contre de renoncer à l'optimalité globale. Néanmoins, il est possible d'utiliser un algorithme de programmation dynamique permettant d'obtenir une coupe acceptable (heuristique). Cet algorithme consiste à choisir, pour un sous-arbre[4] dans l'ontologie, une coupe optimale parmi celles constituées des successeurs directs de la racine de ce sous-arbre et les sous-coupes « optimales » de chacun de ces successeurs, obtenues de manière similaire. Plus formellement, l'algorithme récursif donné en table 1 est appliqué à partir de la racine de l'ontologie[5].

2.3 Exemple

Pour illustrer le fonctionnement de la technique de sélection des coupes décrite précédemment, admettons que l'on dispose de l'ontologie présentée en figure 2 ; les valeurs indiquées en regard

[4] Bien que les ontologies utilisées présentent usuellement une structure de graphe orienté sans cycle (DAG), nous simplifierons ici le propos en considérant qu'il s'agit d'arbres. Cette approximation, qui n'invalide en rien les raisonnements exposés ici, n'est évidemment pas faite en pratique.

[5] En pratique, plusieurs optimisations sont introduites (notamment, les successeurs feuilles d'un nœud sont nécessairement compris dans la sous-coupe optimale pour ce nœud) ; mais elles ne changent rien à l'aspect fondamental présenté ici.

ALGORITHME CRM

Entrée : un nœud t (dans une hiérarchie).

Sortie : CRM : *une* coupe de redondance minimale sous ce nœud.

Si $t \in \mathcal{W}$
  $CRM \leftarrow \{t\}$
Sinon
  Pour $n_i \in t^+$
    $\gamma_i \leftarrow \text{CRM}(n_i)$
    $\vartheta_i \leftarrow \{n_i\}$
  Pour $1 \leq k \leq n := |t^+|$
    $\Gamma_k \leftarrow \bigcup_{j \in [1:n \setminus k]} \gamma_j \cup \vartheta_k$
  $\Gamma_{n+1} \leftarrow \bigcup_{j \in [1:n]} \gamma_j$
  $\Gamma_{n+2} \leftarrow \bigcup_{j \in [1:n]} \vartheta_j$
  $CRM \leftarrow \text{Argmax}_{\Gamma_j : 1 \leqslant j \leqslant n+2} \left(\mathcal{C}_H(\Gamma_j) \right)$

où Argmax retourne une coupe possible réalisant ce maximum.

TAB. 1 – Algorithme de recherche heuristique d'une CRM.

des feuilles correspondent aux fréquences d'occurrences des mots y-relatifs obtenues sur un corpus fictif.

Pour la coupe $\Gamma = [\text{ANIMAL}, \text{PLANTE}, \text{TRANSPORT}]$, on obtient la valeur du critère $\mathcal{C}_H$:

n_i	ANIMAL	PLANTE	TRANSPORT
$f(n_i)$	18	30	1
$P_f(n_i)$	0.3673	0.6122	0.0204
$-P_f(n_i)\log_2 P_f(n_i)$	0.5307	0.4334	0.1146
$\mathcal{C}_H(\Gamma) = \frac{1.0787}{\log_2(3)} = 0.6806$			
$R(\Gamma) = 1 - \mathcal{C}_H(\Gamma) = 0.3194$			

Dans un tel cas de figure, en examinant l'ensemble des 2036 différentes coupes possibles, on trouverait que le critère sur la coupe optimale (indiquée sur la figure 2) vaut 0.874. L'algorithme de recherche par optimum local trouve une coupe pour laquelle le critère est légèrement inférieur: 0.810 ; mais son obtention ne nécessite l'évaluation que de 36 coupes différentes.

3 Expériences

Nous avons effectué un jeu d'expériences en utilisant les collections standards ADI, TIME, MED, CACM et CISI[6] du projet SMART (Salton, 1971), ainsi qu'une ontologie produite à partir du dictionnaire électronique *EDR* (Miyoshi *et al.*, 1996).

EDR est organisée en cinq dictionnaires de différents types, plus ou moins indépendants les uns

[6] Disponibles à l'adresse *ftp://ftp.cs.cornell.edu/pub/smart/*.

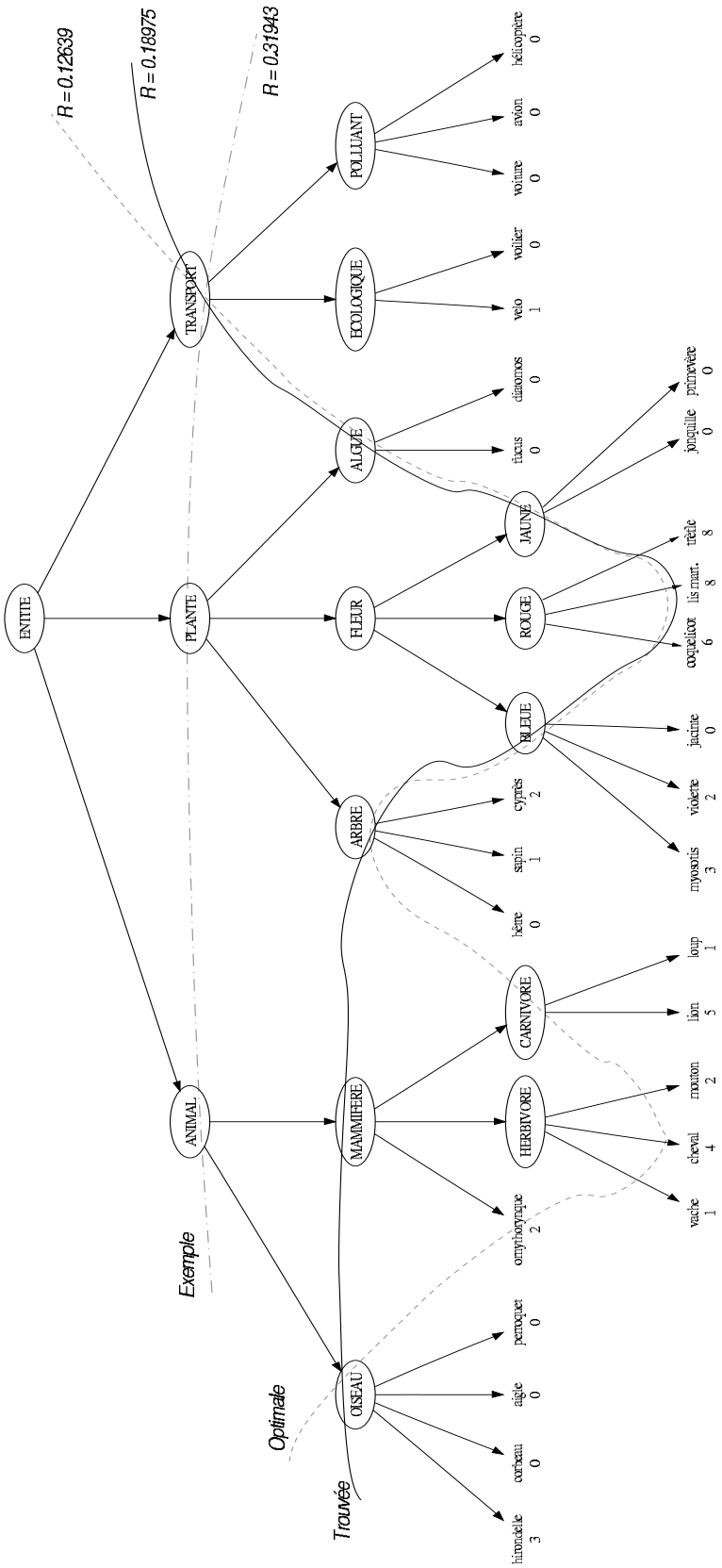


FIG. 2 – Exemple de coupes dans une ontologie.

des autres. Parmi l'ensemble de ces dictionnaires, les deux suivants sont utilisés pour constituer l'ontologie:

le dictionnaire des mots anglais, qui rassemble les informations morphologiques (prononciation, découpage syllabique, inflexion, ...) et syntaxiques (catégorie morpho-syntaxique, dénombrabilité, flexions, ...) pour un peu plus de 240′000 graphies différentes (correspondant à ≈ 420′000 mots), et permet de relier ces graphies avec les informations du dictionnaire des concepts. Les graphies de ce dictionnaire sont principalement (mais pas exclusivement) des lemmes ; il comporte également un nombre important de multi-termes (> 113′000), figurant des mots composés et expressions idiomatiques.

le dictionnaire des concepts, qui décrit à peu près 490′000 concepts, organisés hiérarchiquement entre eux selon des relations d'hyponymie/hyperonymie (chaque concept pouvant avoir plusieurs hyponymes et hyperonymes). Un certain nombre de relations sémantiques binaires supplémentaires (telles que objet-action, agent-action, agent-but) sont par ailleurs décrites, mais nous ne les utilisons pas ici. Remarquons qu'un nombre important de concepts (environ la moitié) ne sont pas directement associés à des mots ; ces concepts ne peuvent être définis et appréhendés qu'au travers de leurs relations avec les autres concepts.

Le système de RD utilisé est le modèle vectoriel SMART, combiné à un lemmatiseur externe[7], qui fait également office de segmenteur (*tokenizer*) et d'étiqueteur morpho-syntaxique. Un filtrage par catégorie grammaticale est réalisés (ne sont conservés que les noms, adjectifs et verbes), mais nous n'utilisons pas d'anti-dictionnaire et ne faisons pas de filtrage fréquentiel.

Les transformations du jeu d'indexation sont obtenues en prétraitant les données soumises au système de RD:

1. en premier lieu, les diverses informations textuelles (principalement titre et contenu) des documents sont agrégées, et les autres informations (auteurs, sources, etc.) supprimées ; documents et requêtes sont ensuite segmentés et lemmatisés ;
2. on cherche ensuite les correspondances entre les mots contenus dans les documents et ceux décrits dans l'ontologie ; on tente d'établir en priorité une correspondance avec la graphie, et s'il n'y en a pas, avec sa forme lemmatisée ; les mots sans correspondance sont indexés de manière traditionnelle ; les taux de couverture[8] sur les différentes collections sont de l'ordre de 90%.
3. on procède ensuite à l'expansion de la hiérarchie des concepts relatifs aux mots conservés pour l'ensemble des documents ; selon les différents cas expérimentés, on prendra soit la *totalité des concepts* possibles (en tablant sur un renforcement mutuel des concepts « corrects » induit par les multiples co-occurrences), soit uniquement le *concept le plus probable* (dans l'absolu pour le mot donné – cette information est présente dans l'ontologie utilisée) ;
4. on détermine ensuite une coupe optimale selon le critère C_H, au moyen de l'algorithme CRM présenté en section 2.2 ;
5. finalement, on substitue les mots des documents et des requêtes par les identificateurs des concepts de la coupe qui les subordonnent.

[7] Le système Sylex 1.7 (© 1993-98 DECAN INGENIA).

[8] Par « *couverture* », on désigne la fraction des occurrences des mots couverts par l'ontologie.

	mesure	(a)	(b)	(c)	(d)
corpus ADI (82 documents)					
tous les concepts, tf.idf	taille index	1800	14748	10099	1292
	précision	**0.3578**	0.3134	0.3356	0.2458
	rappel	0.6984	0.7126	**0.7406**	0.6017
tous les concepts, sans pondération	précision	0.2497	0.1219	**0.2550**	0.1607
	rappel	0.5996	0.3452	**0.6708**	0.5130
concept le plus probable, tf.idf	taille index	1800	5255	2888	658
	précision	0.3578	0.4060	**0.4274**	0.2052
	rappel	0.6984	0.7306	**0.7217**	0.5200
concept + probable, sans pondération	précision	0.2497	0.1376	**0.2939**	0.1466
	rappel	0.5996	0.3727	**0.7141**	0.4911
corpus TIME (423 documents)					
tous les concepts, tf.idf	taille index	21815	93707	70091	6760
	précision	**0.5496**	0.4231	0.4536	0.2683
	rappel	**0.8901**	0.7642	0.8036	0.6026
tous les concepts, sans pondération	précision	**0.3288**	0.0337	0.2353	0.0370
	rappel	**0.7755**	0.1021	0.5709	0.1387
concept le plus probable, tf.idf	taille index	21815	53140	31612	4814
	précision	0.5496	0.5143	**0.5565**	0.2729
	rappel	0.8901	0.8760	**0.9053**	0.5162
concept + probable, sans pondération	précision	0.3288	0.0346	**0.3692**	0.0372
	rappel	**0.7755**	0.1201	0.7590	0.1322
corpus MED (1033 documents)					
tous les concepts, tf.idf	taille index	11893	51712	38524	4078
	précision	**0.4607**	0.3029	0.2996	0.2336
	rappel	**0.5547**	0.3903	0.3794	0.3142
tous les concepts, sans pondération	précision	**0.3623**	0.0105	0.1905	0.0229
	rappel	**0.4574**	0.0246	0.2749	0.0513
concept le plus probable, tf.idf	taille index	11893	30284	18109	2888
	précision	**0.4607**	0.4266	0.4518	0.0743
	rappel	**0.5547**	0.5169	0.5404	0.1042
concept + probable, sans pondération	précision	**0.3623**	0.0105	0.3229	0.0132
	rappel	**0.4574**	0.0313	0.4230	0.0368
corpus CISI (1460 documents)					
tous les concepts, tf.idf	taille index	10019	53453	39544	3516
	précision	**0.1733**	0.1043	0.1139	0.0740
	rappel	**0.2318**	0.1627	0.1675	0.1294
tous les concepts, sans pondération	précision	**0.0687**	0.0232	0.0569	0.0282
	rappel	**0.1239**	0.0376	0.0963	0.0492
concept le plus probable, tf.idf	taille index	10019	26246	14993	1894
	précision	0.1733	0.1590	**0.1825**	0.0602
	rappel	**0.2318**	0.2131	**0.2313**	0.0895
concept + probable, sans pondération	précision	0.0687	0.0201	**0.0805**	0.0221
	rappel	0.1239	0.0403	**0.1300**	0.0435
corpus CACM (3204 documents)					
tous les concepts, tf.idf	taille index	10053	51712	38524	4078
	précision	**0.2865**	0.1293	0.1935	0.1089
	rappel	**0.4534**	0.2579	0.3617	0.1999
tous les concepts, sans pondération	précision	**0.1555**	0.0133	0.1447	0.0320
	rappel	**0.3082**	0.0306	0.2549	0.0699
concept le plus probable, tf.idf	taille index	10053	25207	14681	2670
	précision	**0.2865**	0.2358	0.2804	0.0645
	rappel	0.4534	0.3834	**0.4567**	0.1090
concept + probable, sans pondération	précision	**0.1555**	0.0230	0.1472	0.0245
	rappel	**0.3082**	0.0302	0.2926	0.0385

TAB. 2 – Résultats des différentes expériences sur différents corpus. (a) : mots uniquement ; (b) : mots + concepts ; (c) : hyperonymes directs et (d) : hyperonymes dans CRM (cf aussi fig. 1).

On trouvera dans la table 2 les valeurs de précision (« *11-pt prec* ») et de rappel (« *30 doc* »)[9] fournies par le système SMART. Toutes les expériences sont par ailleurs conduites en utilisant soit le schéma de pondération classique (« tf.idf »), soit sans pondération.

On constate que l'indexation par hyperonymes directs obtient des résultats sensiblement égaux au système de base, mais pour un rappel plus élevé. L'indexation par CRM dégrade par contre les performances.

4 Conclusion

Les résultats obtenus sur ces expériences ne sont malheureusement pas concluants quant à l'utilisation du critère CRM pour l'indexation sémantique. Cependant, plusieurs remarques sont à apporter :

- Le critère utilisé ici ne permet pas de sélectionner, ni même d'influencer, le niveau de profondeur dans l'ontologie de la coupe obtenue. Au vu de la réduction drastique du jeu d'indexation et des mauvaises performances obtenues, il semble que ce critère, ou du moins l'heuristique implémentée, sélectionne une coupe située trop haut dans la hiérarchie, ce qui a comme conséquence évidente de faire baisser la précision. La bonne performance de la coupe au niveau des concept hyperonymes directs nous permet de croire qu'il doit y avoir un niveau plus adapté, plus proche des feuilles, pour la CRM.
  On pourrait par exemple limiter considérablement l'espace de recherche de la coupe idéale en empêchant de considérer des nœuds situés « trop hauts » dans la hiérarchie. Une piste à explorer pour améliorer tant l'adéquation de la coupe sélectionnée avec un processus d'indexation que la recherche de cette coupe elle-même consisterait à explorer les gains possibles en terme de redondance à partir de la coupe uniquement constituée de feuilles, et en dirigeant la recherche vers le haut de la hiérarchie, plutôt que de haut en bas à partir de la racine, comme dans l'heuristique présentée ici.
- Par ailleurs, en conservant l'idée d'une action sur le jeu d'indexation lui-même, il serait intéressant d'examiner de quelle manière les pondérations (e.g. « tf.idf »), utilisées uniquement lors de la recherche des documents proprement dite, devraient être prises en compte lors de la détermination de la coupe.
- Finalement, les résultats présentés ici restent à corroborer avec ceux à obtenir avec d'autres ontologies, en particulier WordNet, qui a une structure assez différente d'EDR.

Pour terminer, soulignons que l'intérêt de la technique présentée dépasse le cadre de la stricte recherche documentaire. Celle-ci pourrait en effet s'avérer utile, et peut être même plus prometteuse, pour d'autres domaines d'application tels que la classification de documents ou le résumé automatique.

Références

DEERWESTER S. C., DUMAIS S. T., LANDAUER T. K., FURNAS G. W. & HARSHMAN R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, **41**(6),

[9] Il s'agit là de mesures standard: la « *11-pt precision* » est la moyenne des précisions pour les taux de rappels 0.0, 0.1, ..., 1.0, où la précision au taux de rappel 0.0 est la précision maximale obtenue sur l'ensemble des documents pertinents retrouvés ; le « *rappel 30 doc* » est le taux de rappel après 30 documents retournés.

391–407.

GONZALO J., VERDEJO F., CHUGUR I. & CIGARRAN J. (1998a). Indexing with WordNet synsets can improve text retrieval. In *Proc. of the COLING/ACL 1998 Workshop on Usage of WordNet for Natural Language Processing*, p. 38–44.

GONZALO J., VERDEJO F., PETERS C. & CALZOLARI N. (1998b). Applying EuroWordNet to multilingual text retrieval. *Journal of Computers and the Humanities*, **32**(2-3), 185–207.

HARMAN D. (1988). Towards interactive query expansion. In *Proc. of the 11th Annual Int. ACM-SIGIR Conference on Research and development in information retrieval*, p. 321–331.

HOFMANN T. (1999). Probabilistic latent semantic indexing. In *proc. of the 22th International Conference on Research and Development in Information Retrieval (SIGIR)*, p. 50–57.

IDE N. & VÉRONIS J. (1998). Word sense disambiguation: The state of the art. *Computational Linguistics*, **24**(1), 1–40.

LI H. (1998). A probabilistic approach to lexical semantic knowledge acquisition and structural disambiguation. Master's thesis, Graduate School of Science, University of Tokyo.

MIHALCEA R. & MOLDOVAN D. (2000). Semantic indexing using WordNet senses. In *Proc. of ACL Workshop on IR & NLP*.

MIYOSHI H., AMD M. KOBAYASHI K. S. & OGINO T. (1996). An overview of the EDR electronic dictionary and the current status of its utilization. In *Proc. of COLING*, p. 1090–1093.

MOLDOVAN D. I. & MIHALCEA R. (2000). Using wordnet and lexical operators to improve internet searches. *IEEE Internet Computing*, **4**(1), 34–43.

RICHARDSON R. & SMEATON A. F. (1995). *Using WordNet in a Knowledge-Based Approach to Information Retrieval*. Rapport interne CA-0395, Dublin City University, Glasnevin, Dublin 9, Ireland.

SALTON G. (1968). *Automatic Information Organization and Retrieval*. McGraw-Hill.

SALTON G. (1971). *The SMART Retrieval System – Experiments in Automatic Document Processing*. Prentice Hall.

SHANNON C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, **27**, 379–423.

SMEATON A. F. & QUIGLEY I. (1996). Experiments on using semantic distances between words in image caption retrieval. In *Proc. of 19th Int. Conf. on Research and Development in Information Retrieval*, p. 174–180.

VOORHEES E. M. (1993). Using WordNet to disambiguate word senses for text retrieval. In *Proc. of 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, p. 171–80.

VOORHEES E. M. (1994). Query expansion using lexical-semantic relations. In *Proc. 17th Annual Int. ACM-SIGIR Conf. on Research and Development in Information Retrieval*, p. 61–69.

VOORHEES E. M. (1998). Using WordNet for text retrieval. In C. FELLBAUM, Ed., *WordNet: An Electronic Lexical Database*, chapter 12, p. 285–303. MIT Press.

WHALEY J. M. (1999). *An Application of Word Sense Disambiguation to Information Retrieval*. Rapport interne PCS-TR99-352, Dartmouth College, Computer Science, Hanover, NH.

WILKS Y. & STEVENSON M. (1998). Word sense disambiguation using optimised combinations of knowledge sources. In *Proc. of the 17th Int. Conf. on Computational Linguistics*, p. 1398–1402.

Recherche en corpus de réponses à des questions définitoires

Véronique Malaisé[1, 2] Thierry Delbecque[2, 3] Pierre Zweigenbaum[2, 3, 4]
(1) DRE de l'Institut National de l'Audiovisuel
4, avenue de l'Europe, 94366 Bry-sur-Marne Cedex
(2) INSERM, U729, 75006 Paris
(3) INALCO, CRIM, 75343 Paris Cedex 07
(4) Assistance Publique - Hôpitaux de Paris, STIM/DSI, 75674 Paris Cedex 14
vmalaise@ina.fr, thd@biomath.jussieu.fr, pz@biomath.jussieu.fr

Mots-clefs : Systèmes de questions-réponses, repérage d'énoncés définitoires, patrons lexico-syntaxiques, médecine

Keywords: Question-answering systems, mining definitions, lexico-syntactic patterns, medicine

Résumé Les systèmes de questions-réponses, essentiellement focalisés sur des questions factuelles en domaine ouvert, testent également d'autres tâches, comme le travail en domaine contraint ou la recherche de définitions. Nous nous intéressons ici à la recherche de réponses à des questions « définitoires » portant sur le domaine médical. La recherche de réponses de type définitoire se fait généralement en utilisant deux types de méthodes : celles s'appuyant essentiellement sur le contenu du corpus cible, et celles faisant appel à des connaissances externes. Nous avons choisi de nous limiter au premier de ces deux types de méthodes. Nous présentons une expérience dans laquelle nous réutilisons des patrons de repérage d'énoncés définitoires, conçus pour une autre tâche, pour localiser les réponses potentielles aux questions posées. Nous avons intégré ces patrons dans une chaîne de traitement que nous évaluons sur les questions définitoires et le corpus médical du projet EQueR sur l'évaluation de systèmes de questions-réponses. Cette évaluation montre que, si le rappel reste à améliorer, la « précision » des réponses obtenue (mesurée par la moyenne des inverses de rangs) est honorable. Nous discutons ces résultats et proposons des pistes d'amélioration.

Abstract Question-answering systems mostly focus on open-domain, factoid questions, but also test other tasks such as restricted-domain and « definitional » questions. We address here the search for definitional questions in the medical domain. Searching for answers to definitional questions generally resorts to two kinds of methods : those which mostly rely on the contents of the target corpus, and those which call on external resources. We have chosen to limit ourselves to the first kind. We present an experiment in which we reuse lexico-syntactic patterns, formerly designed for another task, to locate answers to definitional questions. We have integrated these patterns in a processing chain which we evaluate on the medical definitional questions and corpus of project EQueR (evaluation of French QA systems). This evaluation shows that, while recall still needs to be increased, the « precision » of the obtained answers (as measured through the mean reciprocal rank) is honorable. We discuss these results and propose directions for improvement.

1 Introduction

Les systèmes de questions-réponses ont été jusqu'ici principalement évalués sur des questions de type « factuel », dont la réponse attendue est un fait (Voorhees & Tice, 2000). Des questions recherchant des définitions ont cependant été introduites lors de la campagne d'évaluation TREC-12 QA de 2003 (Voorhees, 2003). Sur les 500 questions de TREC-12, 50 portaient sur des définitions : 30 concernaient un personnage (« *Who is Andrea Bocceli ?* »), 10 une organisation (« *What is ETA in Spain ?* »), et 10 d'autres « choses » (« *What is feng shui ?* ») [1].

La campagne EQueR[2] pour l'évaluation de systèmes de questions-réponses en français a également inclus des questions « définitoires ». Une particularité de cette campagne est d'avoir mis en place, à côté d'une traditionnelle tâche de questions-réponses en domaine ouvert, une tâche de questions-réponses à domaine restreint (corpus et questions de domaine médical). Dans la tâche médicale, sur 200 questions, 70 (soit plus d'un tiers) portaient sur des définitions, dont aucune ne concernait un personnage ou une organisation.

Cet article porte sur cette recherche de réponses à des questions « définitoires » médicales[3]. Dans les travaux antérieurs, deux types principaux de méthodes sont employées pour rechercher ce type de réponse : des méthodes endogènes, par application sur le corpus de patrons d'énoncés définitoires, et des méthodes exogènes, qui projettent sur ce corpus des définitions obtenues dans des ressources dictionnairiques externes ((Hildebrandt *et al.*, 2004), par exemple, combinent les deux). Nous avions à notre disposition les définitions d'une partie des termes du thesaurus MeSH, rédigées par l'équipe CISMeF du CHU de Rouen. Pour les questions portant sur des termes de ce thesaurus, la projection en corpus de ces définitions aurait pu être pertinente. Cependant, les questions ayant été préparées par l'équipe CISMeF, l'usage de ces définitions constituerait un biais[4]. Nous nous sommes donc focalisés sur une méthode endogène, et avons cherché à réutiliser des travaux réalisés sur le repérage d'énoncés définitoires à des fins de construction d'ontologie (Malaisé *et al.*, 2004). C'est cette méthode qui fait l'objet de cet article.

Après une revue de travaux existants (section 2), nous présentons le corpus médical EQueR (section 3), la méthode que nous avons mise en place (section 4) et ses résultats (section 5). Nous discutons ces résultats (section 6) et concluons (section 7).

2 Travaux antérieurs

Les travaux réalisés autour de la tâche TREC-QA de 2003 constituent une source naturelle de bibliographie. Les réponses aux questions définitoires ont été considérées comme un ensemble de « pépites » d'information (Voorhees, 2003) correspondant à des éléments de définition à retrouver : éléments « vitaux », « non-vitaux » ou non pertinents. Les systèmes doivent ramener un maximum d'éléments vitaux et un minimum d'éléments non pertinents. Nous considérons

[1] La notion de « définition » d'un personnage peut sembler étrange, mais c'est le typage qui a été choisi dans cette campagne d'évaluation.

[2] EQueR fait partie du projet Technolangue EVALDA coordonné par ELDA ; contacts : Christelle Ayache (ELDA), Brigitte Grau (LIMSI) (`http://www.technolangue.net/article61.html`). La tâche médicale a été gérée par Magaly Douyère (CISMeF). L'évaluation a été menée en juillet 2004.

[3] La méthode présentée ici n'a pas été mise en place, faute de temps, dans le prototype que nous avons présenté à EQueR. Les connaissances utilisées ont néanmoins été préparées indépendamment des questions EQueR.

[4] Un examen *a posteriori* montre effectivement que 15 des réponses aux questions définitoires sont constituées par ces définitions, qui figurent dans certains documents du corpus.

également, suivant (Meyer, 2001), qu'un énoncé est « définitoire » s'il contient au moins un élément susceptible de servir de base à la construction d'une définition lexicographique. Cet élément peut être, par exemple, l'hyperonyme du terme ou une caractéristique qui permet de le distinguer d'autres termes proches dans le domaine. Idéalement, les énoncés que nous recherchons combinent ces deux types d'éléments. Parmi les participants à TREC 2003, (Hildebrandt *et al.*, 2004) s'appuient sur plusieurs sources de connaissances pour composer une réponse à une question définitoire. La première est obtenue en appliquant sur le corpus cible (AQUAINT) onze patrons recherchant des « pépites » de définition. La deuxième utilise le dictionnaire Merriam-Webster en ligne. La troisième consiste, en dernier recours, à collecter toutes les phrases du corpus contenant le terme cible. L'évaluation des réponses aux questions définitoires privilégiait largement le rappel par rapport à la précision. De ce fait, un système reposant essentiellement sur de simples techniques de recherche d'information, soumis par BBN après l'évaluation (Xu *et al.*, 2003), a obtenu de meilleurs résultats que les systèmes des participants. L'évaluation des définitions dans EQueR n'a pas mis en place une telle prime au rappel. Il était donc raisonnable de privilégier une approche plus précise.

Indépendamment des systèmes de questions-réponses, (Klavans & Muresan, 2001) se sont intéressées au repérage d'énoncés définitoires dans le domaine médical. Elles indiquent que 75 % des énoncés définitoires peuvent être retrouvés à l'aide des patrons qu'elles ont mis au point, et que ce rappel peut être augmenté par *bootstrapping*. Les différents travaux de recherche de définitions, que ce soit au moyen de patrons lexico-syntaxiques (Rebeyrolle, 2000), d'exploration contextuelle (Cartier, 1997) ou de règles (Klavans & Muresan, 2001), se basent schématiquement sur des indices ou *marqueurs* de la définition, associés à des contraintes concernant le voisinage lexical et/ou syntaxique du marqueur. Les contextes de ces marqueurs doivent vérifier en corpus l'ensemble des contraintes définies pour que l'énoncé correspondant soit considéré comme de type définitoire. La méthode retenue par (Malaisé *et al.*, 2004) reprend cette approche fondée sur des patrons lexico-syntaxiques ancrés sur des marqueurs.

3 Le corpus médical EQueR

3.1 Les documents

Le corpus médical EQueR est un sous-ensemble des documents indexés par le Catalogue et Index des Sites Médicaux Francophones (CISMeF, `http://www.chu-rouen.fr/cismef/`) : ceux de neuf « sites éditeurs »[5], auxquels viennent se joindre les documents référencés un lien plus loin sur le même site. L'ensemble comporte 5621 documents originellement au format HTML ou PDF (convertis en texte brut), pour un total d'environ 19 millions de mots.

3.2 Les questions définitoires

Les questions d'EQueR, comme celles de TREC, étaient typées au préalable : l'identifiant de la question indiquait s'il s'agissait d'une question factuelle, booléenne, définitoire ou à réponse

[5]Fédération Nationale des Centres de Lutte Contre le Cancer ; La Documentation Française ; Agence Française de Sécurité Sanitaire des Produits de Santé ; Agence Nationale d'Accréditation et d'Évaluation en Santé ; Orphanet, serveur d'informations sur les maladies rares ; site officiel du Sénat ; le CHU de Rouen ; Université de Rouen, restreinte à sa branche médicale ; site bilingue Santé Canada (ministère fédéral de la santé).

sous forme de liste. Parmi les 70 questions définitoires médicales d'EQueR, cinq portaient sur des acronymes : « *Comment l'IPS peut-il être défini ?* », etc. Dans le système que nous avons présenté à EQueR (système STIM-LIPN, voir (Delbecque *et al.*, 2005)), les acronymes étaient repérés par le classique patron « *expression (ACRONYME)* » ou sa variante « *ACRONYME (expression)* ». Ces patrons se sont déclenchés par exemple sur le passage « *L'index de pression systolique (IPS) [...]* ». Ils ont été appliqués sur l'ensemble du corpus et les résultats stockés dans une base de données, avant de recevoir les questions. Les questions repérées comme portant sur un acronyme ont été envoyées sur un traitement spécifique qui accède à cette base.

Les 65 questions restantes portent sur des termes à définir : « *Quelle est la définition de la désinfection ?* », « *Qu'est-ce que le syndrome du décalage horaire ?* ». Une chaîne spécifique a été conçue pour traiter ces questions. C'est sur ces traitements que nous nous concentrons ici.

4 Recherche d'énoncés définitoires pour un terme spécifique

La méthode que nous avons mise en œuvre peut se décrire en trois grandes parties :
- comme pour les « entités nommées » des questions factuelles (Delbecque *et al.*, 2005), nous cherchons à repérer et indexer au préalable tous les énoncés définitoires ;
- une question définitoire étant donnée, il faut en extraire le terme dont on cherche la définition ;
- il faut enfin sélectionner et classer les énoncés définitoires préindexés concernant ce terme.

4.1 Repérage d'énoncés définitoires en corpus

Nous nous sommes appuyés sur des travaux antérieurs, tant théoriques ((Fuchs, 1994), par exemple) qu'appliqués à la recherche en corpus (Rebeyrolle, 2000), et de nos propres corpus de test pour compiler une liste de marqueurs d'énoncés à intérêt définitoire. Nous avons mis au point des patrons lexico-syntaxiques à partir de ces marqueurs et les avons testés lors d'expérimentations antérieures visant à repérer des relations sémantiques entre termes (Malaisé *et al.*, 2004). Ces patrons sont appliqués sur un corpus préalablement analysé par Cordial Analyseur [6] : Cordial segmente en phrases, lemmatise, étiquette les mots et indique leurs relations syntaxiques. Nos patrons portent sur les lemmes et les catégories des mots.

Pour le présent travail, nous n'avons conservé que les patrons qui avaient donné les meilleurs résultats dans ces travaux antérieurs (voir le tableau 1), à savoir ceux modélisés autour de :
- verbes métalinguistiques : « *appeler* », « *nommer* », « *référer* », « *dénommer* », « *désigner* », « *dénoter* », « *signifier* », « *définir* » ;
- noms métalinguistiques associés à un ensemble de verbes supports : « *nom* », « *terme* », « *mot* », « *expression* », « *vocable* », « *appellation* », « *désignation* », « *dénomination* », « *concept* », « *notion* », « *acception* », associés à « *porter* », « *appliquer* », « *employer* », « *réserver* », « *utiliser* », « *donner* », « *renvoyer* », « *référer* », « *être* » ;
- indices de reformulation et d'hyperonymie : « *vouloir dire* », « *entendre par* », « *à savoir* », « *sorte de* », « *est un* », « *par exemple* » ;
- la parenthèse.

Nos patrons lexico-syntaxiques permettent d'extraire des « définitions candidates » et deux groupes syntaxiques dans ces énoncés, qui sont susceptibles de contenir l'élément défini : le *definiendum*. Ces groupes syntaxiques sont extraits selon deux modalités :

[6] http://www.synapse-fr.com/

Type de marqueur	*Patrons lexico-syntaxiques*
Verbes métalinguistiques	*VerbeMeta* NonPrécédéDe *« se »* ET NonSuiviDe [pas/.]
Noms métalinguistiques	*NomMeta* {0-6}MOTS *VerbeSupport* ; *VerbeSupport* {0-6}MOTS *NomMeta* ; *« être »* {0-1}MOT [le/ce] *NomMeta* ; sous [le/ce] NomMeta
Indices de reformulation ou d'hyperonymie	*« vouloir »* {0-n}MOTS *« dire »* NonSuiviDe *« que »* ; *« entendre »* {0-6}MOTS *« par »* ; *« par »* {0-6}MOTS *« entendre »* ; *« à savoir »* PrécédéDe *Ponctuation* ; *« à savoir »* NonPrécédéDe *Ponctuation* ; *Determinant « sorte de »* ; *« être »* {0-3}MOTS [le/la/les/un/une/des] *Nom* ; *« par exemple »* PrécédéDe 1MOT ET SuiviDe 1MOT
Parenthèse	*NomCommun* (1MOT) ; *NomCommun* (*« ou »* *Nom* ; *NomCommun* (*« qui »* [est/se] ; *Verbe* (*VerbeInfinitif*

TAB. 1 – Patrons lexico-syntaxiques pour le repérage d'énoncés définitoires.

– si le marqueur est un verbe, nous extrayons son sujet et son objet direct dans l'énoncé, s'il en contient, et sinon, nous extrayons respectivement :
  – le groupe syntaxique ayant la même fonction que le nom précédant le marqueur ;
  – le groupe syntaxique ayant la fonction du premier mot plein suivant le marqueur ;
– si le marqueur n'est pas un verbe, nous extrayons les groupes syntaxiques précédant et suivant le marqueur de la manière décrite ci-dessus.

Dans les cas où deux marqueurs doivent être présents dans la phrase (*définir* associé à *comme*,...), nous ne spécifions pas la position relative des deux marqueurs dans la phrase, et extrayons les sujets et objets ou les contextes droits et gauches du verbe. Ce procédé rudimentaire donne toutefois des résultats de l'ordre de 55 % de précision (Malaisé *et al.*, 2004) et permet de factoriser les patrons. La qualité de cette extraction dépend également de la qualité de la segmentation initiale des phrases et de leur analyse. Par exemple, dans l'énoncé (5598-1) (qui comporte un titre mal segmenté) *« RECOMMANDATIONS ET RÉFÉRENCES Les patients dyslipidémiques sont définis par une augmentation des taux1 sériques du cholestérol et ou des triglycérides. [...] »*, le marqueur *« défini »* a permis de repérer les deux groupes *« RÉFÉRENCES Les patients dyslipidémiques sont »* et *« par une augmentation des taux1 sériques du cholestérol et ou des triglycérides. Ils ont de ce fait »*.

Selon la phrase, l'un des deux groupes peut également être vide : pour l'énoncé (5601-6) *« [...] La SFHH recommande le terme de prédésinfection pour l' étape préalable à la désinfection ou à la stérilisation Opération utilisant des détergents contenant au moins un principe actif reconnu pour ses propriétés bactéricides , fongicides , sporicides ou virucides (SFHH) [...] »*, le système n'a extrait qu'un groupe « droit » : *« de prédésinfection [...] Opération utilisant des détergents contenant au moins un principe actif reconnu pour ses propriétés bactéricides »*.

Les énoncés définitoires candidats ainsi trouvés, avec le ou les deux groupes extraits correspondant aux positions hypothétiques du definiendum, sont notés dans une table, qui est indexée par les mots (leurs formes graphiques et lemmes) présents dans chacun des groupes. C'est cet index des definienda hypothétiques qui servira lors de la recherche de réponses.

4.2 Analyse de la question

Les questions posées sont traitées en fonction de leur type. Le traitement général appliqué par défaut a été réalisé par l'équipe du LIPN (Thierry Poibeau), et s'appuie sur une série de transducteurs mis en œuvre avec Unitex (`http://www-igm.univ-mlv.fr/~unitex/`). Pour les questions de type « définition » (par exemple, *« Quelle est la définition de "chimiothéra-*

pie" ? »), un traitement spécifique vise à extraire de la question le terme dont on cherche la définition[7]. Il procède par élimination, en supprimant de la question tous les mots considérés comme « vides » : principalement la copule, les déterminants, les particules interrogatives et les verbes de parole. Les mots des deux premières catégories (*« est »*, *« la »*, *« de »*) sont puisés dans plusieurs listes collectées dans des travaux antérieurs, qui ont été augmentées par une liste de particules interrogatives (*« quel »*, *« quelle »*, *« quelles »*, *« quels »*, *« quoi »*, *« comment »*,...) et d'autres mots (*« façon »*, etc.). Les mots qui désignent une définition (*« définition »*, *« définir »*, *« appeler »*...) ont été pris dans les listes des principaux marqueurs employés dans les patrons : ceux qui se sont appliqués sur le corpus. Les listes employées contiennent directement les formes fléchies des mots. Ainsi, pour la question ci-dessus, le terme restant est *« chimiothérapie »*.

4.3 Recherche de définitions en réponse à une question

Il s'agit ici de proposer des définitions pour le terme extrait d'une question. On va pour cela le chercher parmi les definienda hypothétiques relevés précédemment (section 4.1). Ils sont classés en fonction du nombre de mots du terme de la question qu'ils contiennnent. Dans les expériences présentées ici, nous avons imposé que tous les mots du terme recherché soient présents. Les réponses ont alors été classées dans l'ordre des documents du corpus. Dans l'évaluation EQueR, un système peut renvoyer jusqu'à cinq réponses ordonnées. Nous conservons donc les cinq premiers candidats. Une réponse se compose d'un passage (l'énoncé, tronqué à 250 caractères si nécessaire) et d'une réponse courte (la définition). Nous proposons comme définition celui des deux groupes qui ne contient pas le terme de la question (ou celui qui le contient si l'autre est vide). Si l'énoncé est tronqué, il est centré sur ce groupe.

Ainsi, à la question *« Qu'est-ce qu'une aniridie ? »* , l'énoncé (5590-2) donnera comme réponse « courte » *« comme l' absence totale d' iris »*, et comme passage *« Aniridie sporadique TITLE L' aniridie est une absence clinique d' iris [...] L' aniridie se définit comme l' absence totale d' iris . »*. Pour la question *« Quelle est la définition de l'asthme ? »*, l'énoncé (5586-2) donne la réponse courte erronée *« Définition Le clinicien »*, mais le passage correct *« National des Prescriptions et Consommations des Médicaments [...] Définition Le clinicien définit l' asthme comme un accès de dyspnée , de toux et de sifflement paroxystique , dont l' expression peut »*[8].

4.4 Évaluation

L'objectif de ce travail était de répondre aux questions de type définitoire posées dans la campagne EQueR d'évaluation de systèmes de questions-réponses, tâche médicale. Pour ne pas biaiser le système, la mise au point du module de recherche de réponse à des questions définitoires s'est faite indépendamment des questions EQueR. Nous avons utilisé pour cela un ensemble de 735 termes médicaux tirés du thesaurus MeSH pour lesquels nous disposions de définitions en français[9]. Ces termes nous ont servi à tester notre chaîne de traitement.

L'évaluation proprement dite s'est faite sur les questions de la campagne EQueR. Ce module ayant été terminé après la fin de l'évaluation officielle, nous avons nous-mêmes calculé les

[7]Ce terme sera alors stocké sous la forme graphique selon laquelle il apparaît dans la question.

[8]Ce passage n'a hélas pas été classé parmi les cinq premiers par notre système.

[9]Ces définitions sont celles mises au point par l'équipe CISMeF, dans une version du printemps 2004.

scores, en reprenant les principes utilisés dans EQueR. Cette évaluation repose sur un jugement humain de pertinence des réponses courtes et passages produits par le système, avec une possibilité de variation non négligeable[10]. Nous l'avons effectuée nous-mêmes, les jugements individuels et les chiffres synthétiques qui en découlent ne sont donc pas directement comparables à ceux de l'évaluation EQueR officielle. Nous estimons cependant qu'ils devraient rester dans le même esprit.

Pour EQuER, une réponse courte est jugée correcte si elle est juste et précise ; inexacte si elle n'est pas assez précise, incorrecte si elle n'est pas juste, et non justifiée si elle est correcte mais que le document indiqué ne la corrobore pas. Un passage est correct s'il contient au moins une partie d'une réponse juste (le reste étant dans le document) ou incorrect s'il ne contient pas assez ou pas du tout d'éléments corrects (non complétés par le document).

Nous avons noté nos réponses (courtes ou passages) sur trois niveaux : sûrement correcte, possiblement correcte, incorrecte. Le niveau intermédiaire vise à prendre en compte l'écart qui peut exister entre notre jugement et celui qu'auraient rendu les évaluateurs d'EQueR. Nous donnons ainsi deux séries de résultats, un « score strict » qui considère les « possiblement correctes » comme incorrectes, et un « score laxiste » qui les compte comme correctes.

La note assignée à une question est l'inverse du rang de la première bonne réponse renvoyée par le système. La note globale est la moyenne de ces inverses du rang.

5 Résultats

Le repérage des énoncés définitoires a été appliqué à l'ensemble du corpus médical EQueR[11]. 17 792 énoncés définitoires potentiels ont été repérés. Les patrons les plus productifs sont ceux centrés sur les marqueurs : *« définir »* (4950), *« exemple »* (4347), *parenthèse* (1851), *« appeler »* (1336), *« être »* (1272). Les 65 questions définitoires d'EQueR hors acronymes ont été analysées, et les termes extraits ont été recherchés dans les groupes représentant les definienda hypothétiques[12]. Notre système a proposé des réponses à 22 des 65 questions (tableau 2). Selon le jugement porté, entre 5 et 10 des réponses courtes et entre 9 et 16 des passages étaient corrects. Le rang des bonnes réponses trouvées varie du premier au cinquième (et dernier) rang. Dans la version stricte, le bon passage est en moyenne trouvé au troisième rang ($MRR = 0,33$), et la moitié des 22 questions n'obtient aucune bonne réponse. Dans la version laxiste, le bon passage est en moyenne au second rang ($MRR = 0,53$), et seules 6 de ces questions n'obtiennent pas de bonne réponse.

6 Discussion

Pour les raisons expliquées plus haut, il est difficile de comparer nos résultats à ceux obtenus pour les définitions dans les compétitions TREC-QA (pour 2003, f-mesure médiane à 0,192,

[10]Dans le cadre de TREC-QA, la marge d'erreur liée à des différences d'appréciation sur la pertinence de certaines réponses semble globalement ne pas modifier le classement des systèmes.

[11]Il a été programmé en XSLT, et son application sur l'ensemble du corpus EQueR médical prend 25 minutes sur un biprocesseur Xeon 2,4 GHz avec 1,2 Go de mémoire.

[12]Les programmes pour l'analyse des questions et la recherche des réponses aux 65 questions définitoires EQueR, implémentés en Perl, mettent 10 secondes pour traiter l'ensemble des questions sur la même machine.

		Score strict		Score laxiste	
n° EQueR	*Question*	C	P	C	P
MD28	*Quelle est la définition de chimiothérapie ?*	0	0	0	0,25
MD52	*Qu'est-ce qu'une aniridie ?*	0,5	0,5	0,5	0,5
MD56	*Qu'est ce qu'un mésothéliome ?*	1	1	1	1
MD57	*Qu'est-ce qu'une anomalie congénitale ?*	0	0	0	0
MD61	*Qu'est-ce qu'une anorexie ?*	0,33	0,33	0,33	0,33
MD62	*Quelle est la définition de la désinfection ?*	0	0,33	0	0,33
MD64	*Qu'est-ce que la radiothérapie ?*	0	0	0	0,5
MD66	*Quelle est la définition de l'asthme ?*	0	0	0	0
MD69	*Qu'est-ce que le séquençage ?*	0	0	0	0
MD70	*Qu'est-ce que le syndrome du décalage horaire ?*	1	1	1	1
MD71	*Qu'est-ce qu'un trouble dépressif ?*	0	0	1	1
MD75	*Qu'est-ce que l'Index de Pression Systolique ?*	0,2	1	0,2	1
MD77	*Qu'est-ce que la schizophrénie ?*	0	0	1	1
MD79	*Qu'est-ce qu'une amblyopie ?*	0	0	0,33	0,33
MD82	*Qu'est-ce qu'un scanner ?*	0	0	0	0
MD87	*Quelle est la définition de la génomique ?*	0	0	0	0
MD89	*Quelle est la définition du neuroblastome ?*	0	1	0	1
MD91	*Qu'est-ce que la thérapie génique ?*	0	1	1	1
MD94	*Qu'est-ce que la virémie ?*	0	1	0	1
MD98	*Qu'est-ce qu'une sialographie ?*	0	0	0	0
MRD153	*Que signifie le terme chimiothérapie ?*	0	0	0	0,5
MRD189	*Comment peut-on définir un trouble dépressif ?*	0	0	1	1
	Moyenne des inverses de rang (MRR)	0,138	0,326	0,335	0,534
	Nombre de réponses trouvées	5	9	10	16

TAB. 2 – Les 22 questions auxquelles le système a répondu, et le score des réponses fournies. C = réponse courte, P = passage. Les scores sont des inverses de rangs (un score de 0,33 correspond à une réponse trouvée au rang 3). Le score « laxiste » accepte des réponses moins complètes.

meilleure à 0,555) : ils comportaient de nombreuses questions sur des personnes ou des organisations et n'étaient pas calculés de la même façon (« pépites » de connaissance). La comparaison aux résultats d'EQueR est elle aussi malaisée, du fait de la part de jugement humain impliquée dans l'évaluation des réponses individuelles[13].

On peut néanmoins noter qu'un nombre relativement faible (un tiers) de questions obtiennent une réponse, et qu'une partie seulement de ces questions reçoivent une réponse correcte parmi les cinq premières proposées par le système (entre 40 et 73 % pour les passages). D'après les données globales dont nous disposons, nous avons calculé que dans EQueR, 39 questions définitoires ont obtenu au moins un passage correct de la part d'un participant, dont 35 avec la réponse courte correcte. Il est cependant important d'analyser l'origine de ces pertes de façon à augmenter le rappel du système actuel. Ce système étant composé de modules enchaînés, chaque module est susceptible de participer à cette perte d'information.

Le repérage des énoncés définitoires et des definienda hypothétiques est une première origine : dans les travaux précédents, leur rappel a été évalué à environ 50 %, et leur précision de 10 à 69 % suivant la complexité syntaxique des énoncés. De plus les patrons n'ont pas été adaptés au domaine médical, ce qui entraîne à la fois du bruit (marqueurs polysémiques en médecine) et du silence (patrons de définitions de type « médical » non modélisées). L'analyse des questions,

[13]Cette part pourra être réduite si tous les passages corrects sont relevés pour chaque question, ce qui a été fait au moins partiellement par les organisateurs.

réalisée ici de façon simpliste, a laissé passer plusieurs mots « vides »[14] non prévus ou dont la forme employée n'était pas prévue (« *dire* », « *sigle* », « *définie* », « *possible* », ...). Les termes ainsi obtenus (MD15 « *possible ostéosynthèse* », MD16 « *dire noyade sublétale* », ...) ne peuvent être trouvés en position de definiendum. Huit termes ont ainsi été mal identifiés. On peut espérer qu'une analyse employant par exemple des transducteurs et une lemmatisation comme pour les autres questions, donnerait des résultats plus précis. Enfin, dans les cas où plus de cinq réponses ont été trouvées, un meilleur classement des réponses pourrait augmenter le rappel si la bonne réponse ne fait partie des premiers passages renvoyés (une dizaine de cas). Par exemple, la méthode consistant à collecter toutes les phrases où apparaît le terme à définir, puis les mots les plus fréquents dans ces phrases, et à privilégier les phrases qui contiennent le plus grand nombre de ces mots (Xu *et al.*, 2003) semble une piste intéressante.

Il semblerait pertinent de comparer ce que donnerait l'application brute de cette méthode à ce que nous obtenons en filtrant à l'aide du repérage d'énoncés définitoires. On peut espérer que cette focalisation sur les énoncés définitoires réduit le bruit (même si elle ne le supprime pas, en particulier du fait que les patrons employés sont eux-mêmes sources de bruit) : on a vu qu'il n'était pas excessif dans les réponses données par notre système (MRR entre 0,33 et 0,53).

Ensuite, on peut supposer que l'usage de connaissances extérieures (dictionnaires, terminologies avec définitions, locaux ou interrogeables en ligne), que nous nous sommes interdit ici, devrait aider à localiser des définitions candidates non contraintes par les patrons dont nous disposons. Par exemple, certaines définitions sont données en plusieurs phrases, voire en faisant implicitement référence au titre du document, avec une présentation du type *« <TITLE>Noyade sublétale</TITLE> Définition [MeSH Scope Note ; traduction CISMeF] : immersion non fatale dans l'eau. Le sujet peut être réanimé. »* ou *« <TITLE>Adénite</TITLE> Définition [MeSH Scope Note ; traduction CISMeF] : inflammation des ganglions lymphatiques. »*. Ce type de présentation ne semble pas possible à détecter par des patrons génériques. En revanche, le fait de savoir que l'« *Adénite* » est une « *Inflammation des ganglions lymphatiques* » permet d'associer cet énoncé au terme correspondant. Cette information peut se trouver dans un dictionnaire médical, comme celui disponible en ligne à l'URL `http://www.AtMedica.com`.

Pour terminer, soulignons qu'une source importante de bruit est constituée par le corpus lui-même, obtenu par conversion en texte de documents HTML et PDF, conversion dont on sait qu'elle reste difficile à réaliser proprement de façon automatique. Sur un corpus de taille importante, une révision manuelle complète reste hors de portée, et les passages bruités sont nombreux. Ces passages sont à l'origine d'énoncés mal segmentés ou incohérents, que nos programmes n'ont pas su prendre en compte correctement. Ce point constitue aussi probablement une différence importante par rapport au corpus AQUAINT utilisé dans TREC.

7 Conclusion

Le système assemblé ici permet de proposer des réponses à des questions de type définitoire à partir du corpus médical EQueR. Si son rappel doit être amélioré, la précision des réponses proposées est honorable. Notons qu'il a trouvé quelques réponses correctes qu'aucun participant à EQueR n'a trouvées, par exemple celle de « *Quelle est la définition de la désinfection ?* ».

Nous avons souligné qu'au-delà des améliorations à apporter à la préparation du corpus et aux

[14] Particules interrogatives, verbes de parole, etc. (cf. section 4.2).

méthodes présentées elles-mêmes, essentiellement endogènes, un recours à des connaissances extérieures (définitions existantes) devrait aider à renforcer la détection des définitions correctes en corpus. Ce sera le thème de nos prochains travaux.

Remerciements

Nous remercions Magaly Douyère pour ses conseils sur l'évaluation des définitions trouvées.

Références

CARTIER E. (1997). La définition dans les textes scientifiques et techniques : présentation d'un outil d'extraction automatique de relations définitoires. In *2e Rencontres Terminologie et Intelligence Artificielle*, p. 127–140, Toulouse : ERSS.

DELBECQUE T., JACQUEMART P. & ZWEIGENBAUM P. (2005). Utilisation du réseau sémantique de l'UMLS pour la définition de types d'entités nommées médicales dans un système de questions-réponses : impact de la source des documents explorés. In *CORIA (COnférence en Recherche d'Informations et Applications)*, p. 101–115, Grenoble : CLIPS.

FUCHS C. (1994). *Paraphrase et énonciation*. Paris, Ophrys.

HILDEBRANDT W., KATZ B. & LIN J. (2004). Answering definition questions using multiple knowledge sources. In S. DUMAIS, D. MARCU & S. ROUKOS, Eds., *Actes HLT-NAACL*, p. 49–56, Boston, Massachusetts, USA : ACL.

KLAVANS J. & MURESAN S. (2001). Evaluation of the DEFINDER system for fully automatic glossary construction. *Journal of the American Medical Informatics Association*, **8**(suppl), 324–328.

MALAISÉ V., ZWEIGENBAUM P. & BACHIMONT B. (2004). Repérage et exploitation d'énoncés définitoires en corpus pour l'aide à la construction d'ontologie. In P. BLACHE, Ed., *Actes de TALN 2004 (Traitement automatique des langues naturelles)*, p. 269–278, Fès, Maroc : ATALA LPL.

MEYER I. (2001). Extracting knowledge-rich contexts for terminography. In D. BOURIGAULT, C. JACQUEMIN & M.-C. L'HOMME, Eds., *Recent Advances in Computational Terminology*, p. 279–302. Amsterdam : John Benjamins.

REBEYROLLE J. (2000). *Forme et fonction de la définition en discours*. Thèse de doctorat, Université de Toulouse II - Le Mirail.

VOORHEES E. M. (2003). Overview of the TREC 2003 question answering track. In E. M. VOORHEES & L. P. BUCKLAND, Eds., *Actes Twelfth Text Retrieval Conference (TREC 2003)*, p. 54–68, Washington DC : NIST.

VOORHEES E. M. & TICE D. M. (2000). The TREC-8 question answering track evaluation. In E. M. VOORHEES & D. K. HARMAN, Eds., *Eighth Text REtrieval Conference (TREC-8)*, NIST Special Publication 500-246, p. 83–105.

XU J., LICUANAN A. & WEISCHEDEL R. (2003). TREC2003 QA at BBN : Answering definitional questions. In E. M. VOORHEES & L. P. BUCKLAND, Eds., *Actes Twelfth Text Retrieval Conference (TREC 2003)*, p. 98–106, Washington DC : NIST.

QRISTAL, système de Questions-Réponses

Dominique Laurent, Patrick Séguéla

Synapse Développement
33 rue Maynard,
31000 Toulouse, France
{dlaurent, p.seguela}@synapse-fr.com

Mots-clés : Système de questions-réponses, recherche d'information, évaluation des systèmes de questions-réponses, extraction de réponse, recherche sur le Web, QRISTAL.

Keywords: Question Answering system, information retrieval, Question Answering evaluation, answer extraction, Web search strategy, QRISTAL.

Résumé

QRISTAL[1] (Questions-Réponses Intégrant un Système de Traitement Automatique des Langues) est un système de questions-réponses utilisant massivement le TAL, tant pour l'indexation des documents que pour l'extraction des réponses. Ce système s'est récemment classé premier lors de l'évaluation EQueR (Evalda, Technolangue[2]). Après une description fonctionnelle du système, ses performances sont détaillées. Ces résultats et des tests complémentaires permettent de mieux situer l'apport des différents modules de TAL. Les réactions des premiers utilisateurs incitent enfin à une réflexion sur l'ergonomie et les contraintes des systèmes de questions-réponses, face aux outils de recherche sur le Web.

Abstract

QRISTAL[1] is a question answering system which makes intensive use of natural language processing techniques, for indexing documents as well as for extracting answers. This system recently ranked first in the EQueR evaluation exercise (Evalda, Technolangue[2]). After a functional description of the system, its results in the EQueR exercise are detailed. These results and some additional tests allow to evaluate the contribution of each NLP component. The feedback of the first QRISTAL users encourage further thoughts about the ergonomics and the constraints of question answering systems, faced with the Web search engines.

---

[1] développé avec le soutien de l'ANVAR et de la Commission Européenne (TRUST, IST-1999-56416), cf. Amaral (2004), Laurent (2004).

[2] http://www.technolangue.net

1 Introduction

QRISTAL est un système de questions-réponses multilingue (français, anglais, italien, portugais, polonais) conçu pour extraire des réponses de documents placés sur un disque dur, ou pour extraire des réponses à partir du Web sur la base des pages ou passages retournés par des moteurs Web classiques (Google, MSN, AOL, etc.)

Le système reconnaît un grand nombre de formats (.html, .xml, .txt, .doc, .dbx, .hlp, .pdf, .ps, etc.), autorisant ainsi l'indexation de l'immense majorité des textes mais également des e-mails ou encore des fichiers d'aide.

Commercialisé depuis novembre 2004 pour la plate-forme Windows, ce système est destiné au grand public. Cependant, il est en cours d'intégration dans des applications professionnelles de recherche d'information. Chacun peut le tester sur le site www.qristal.fr, le corpus de test étant constitué du manuel de grammaire en ligne disponible sur http://www.synapse-fr.com/grammaire/GTM_0.htm

Notre système est fondé sur la technologie Cordial d'analyse syntaxique et sémantique du texte. Il se caractérise par une utilisation intensive des outils de TAL, entre autres l'analyse syntaxique, la désambiguïsation sémantique, la recherche des référents des anaphores, la détection des métaphores, la prise en compte des converses, le repérage des entités nommées ou encore l'analyse conceptuelle et thématique. L'utilisation professionnelle ou grand public a nécessité une optimisation constante des différents modules afin que le logiciel reste extrêmement rapide, l'utilisateur étant maintenant habitué à obtenir ce qui ressemble à des réponses dans un délai très court.

2 Architecture

L'architecture du système est modulaire. Le schéma général est décrit par la figure 1.

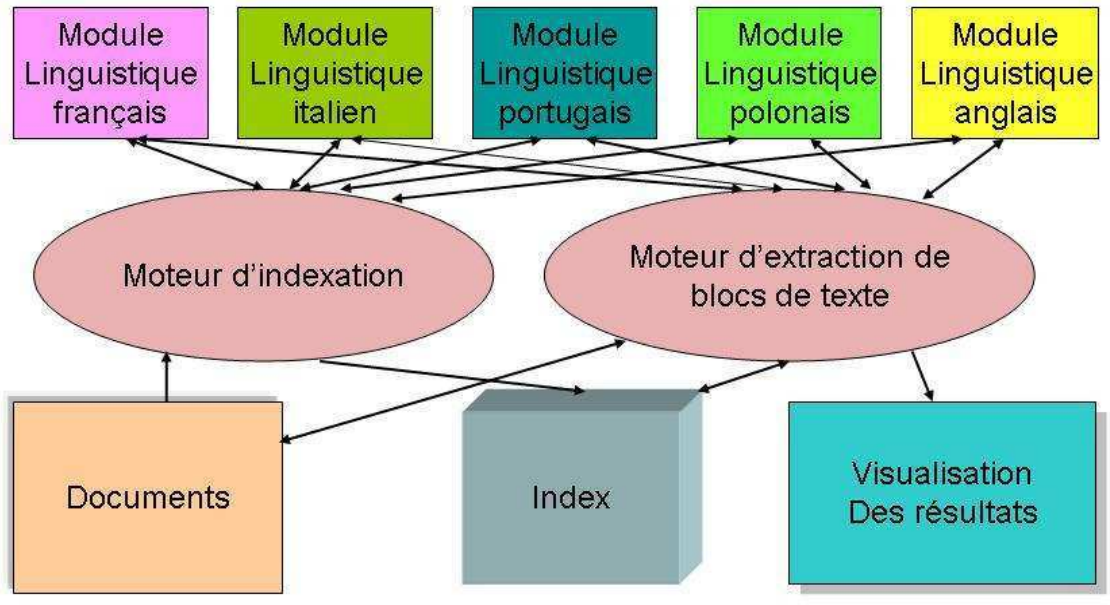


Figure 1. Architecture générale du système

Ce système est donc un moteur complet d'indexation et d'extraction de réponses. Toutefois l'indexation n'est effectuée que pour les documents "statiques", la recherche sur le Web se faisant à l'aide d'un métamoteur, par conséquent sans indexation préalable des pages.

2.1 Indexation multicritères

Au-delà du schéma général de fonctionnement du système, la figure 2 décrit le processus d'analyse linguistique effectué lors de l'indexation, lors de l'analyse de la question et lors de l'extraction de la réponse.

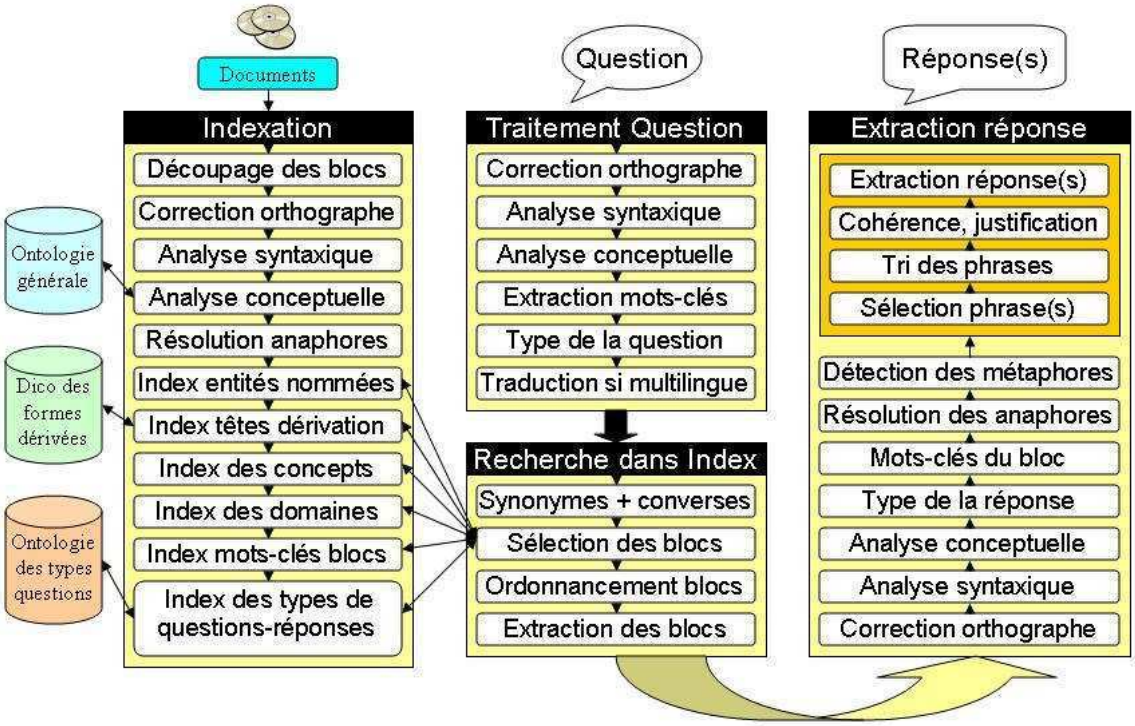


Figure 2. Processus d'analyse linguistique lors de l'indexation

Les textes sont convertis en Unicode puis découpés en blocs de longueur fixe, actuellement un kilo-octet. Cette découpe permet de réduire la taille de l'index car seul le nombre d'occurrences d'un lemme donné par bloc est sauvé dans l'index, ce nombre constituant par ailleurs un indice précieux de la pertinence de chaque bloc lors de la recherche d'un lemme donné dans l'index. En fait le mot "lemme" est ici fautif car sont indexées des têtes de dérivation. Par exemple "symétriques", "asymétrie", "dissymétriques", "symétriseraient" ou "symétrisable" seront indexés dans la même entrée "symétrie", réduisant de fait encore la taille de l'index sur disque.

Chacun des blocs de texte est analysé syntaxiquement et sémantiquement. À partir des informations issues de cette analyse, plusieurs index sont constitués :

- index des têtes de dérivation (les têtes pouvant être des sens de mots comme "symétrie"),
- index des noms propres (si ces noms propres figurent dans nos dictionnaires),

- index des expressions (connues de nos dictionnaires d'expressions d'environ 50 000 entrées, comme "frein moteur", "prendre l'air", "par inadvertance"...),
- index des entités nommées (repérées dans le texte, comme "George W. Bush" ou "Société Nationale des Chemins de fer Français"),
- index des concepts (sur deux niveaux de l'ontologie générale : 256 catégories, comme la visibilité et 3 387 sous-catégories, comme l'éclairage ou la transparence),
- index des domaines (186 domaines, comme l'aéronautique, l'agriculture, etc.),
- index des types de questions-réponses (distance, vitesse, définition, causalité...),
- index des mots-clés du texte.

Le processus d'indexation est similaire pour chacune des langues, ce qui permet de disposer de données indépendantes de la langue (numéros de concepts dans l'ontologie, numéro de domaine, type de question-réponse) qui fournissent des bases intéressantes pour retrouver dans une langue des réponses à des questions posées dans une autre langue.

En français, le taux de désambiguïsation grammaticale correcte (distinction nom-verbe-adjectif-adverbe) est supérieur à 99%, le taux de désambiguïsation sémantique est d'environ 90% pour 9 000 mots polysémiques et environ 30 000 sens pour ces mots (nombre de sens nettement inférieur à celui du Larousse par exemple, les dictionnaires d'expressions couvrant déjà un grand nombre de sens idiomatiques). La vitesse d'indexation varie entre 200 et 400 Mo par heure avec un Pentium 3 GHz, selon la taille et le nombre des fichiers à indexer.

L'indexation des types de questions est sans doute l'un des aspects les plus originaux de notre système. Lors de l'analyse des blocs à indexer, les réponses éventuelles sont repérées, par exemple un nom de fonction pour une personne ("boulanger", "ministre", "directeur de cabinet"...) une date de naissance ("né le 28 avril 1958"), un lien de causalité ("dû à l'accumulation de neige", "en raison du gel"...) ou de conséquence ("entraînant de graves perturbations", "facilitant la gestion du trafic"...), et le bloc est alors indexé comme pouvant fournir une réponse du type repéré.

La typologie comprend actuellement 86 types, dont des types factuels (dimension, surface, poids, vitesse, pourcentage, température, prix, nombre d'habitants, nom d'œuvre, etc.) mais aussi de nombreux types non factuels (forme, possession, jugement, but, causalité, opinion, comparaison, classification, etc.). Lors de l'évaluation EQueR (voir §3), 492 questions sur 500 ont été classées selon cette typologie avec seulement 6 erreurs (exemple d'erreur, sur la question 391, "Quels sont les cinq nouveaux membres non-permanents du Conseil de Sécurité de l'ONU ?", le type repéré est "noms de personnes" au lieu de "noms de pays").

L'index des mots-clés du texte est également une originalité de notre système. Il est rendu nécessaire par la découpe des textes en blocs. Du fait de cette découpe, des blocs spécifiques peuvent ne pas contenir les sujets du texte bien que les phrases de ces blocs portent sur ces sujets. L'index des mots-clés permet d'ajouter une plus-value aux textes portant a priori sur le sujet recherché, lequel sujet peut être une notion, une personne, un événement, etc.

2.2 Extraction de réponse

Lorsque l'utilisateur pose sa question, celle-ci est analysée, syntaxiquement et sémantiquement. Le type de question-réponse est déterminé. Relevons cependant ici que l'analyse

sémantique de la question est plus complète que l'analyse effectuée sur les textes car l'énoncé est généralement court. De ce fait, la désambiguïsation sémantique est plus incertaine, par manque de contexte. Si l'utilisateur a la possibilité de "forcer" tel ou tel sens de mot, cette option reste peu utilisée. Aussi calculons-nous un poids pour chaque sens possible et ce poids entre en compte dans la recherche dans l'index (exemple : sens 1 à 20%, sens 2 à 65%, sens 3 à 5% de probabilité). Ainsi les erreurs éventuelles de désambiguïsation sémantique sont tempérées par ces coefficients qui permettent de "remonter" des blocs correspondant à d'autres sens mais ayant d'autres caractéristiques recherchées, par exemple des réponses potentielles au type de la question, un nom propre identique, un même thème, etc.

Après analyse de la question, si le corpus visé est sur disque, les différents index sont consultés et les blocs les mieux placés pour ces index sont réanalysés. Sur le Web, des requêtes sont générées vers les moteurs Web classiques. Dans les pages de résultats retournées par les moteurs, les bribes ("snippets") ou les pages indiquées par les liens (pour les questions non factuelles) sont analysées.

Comme indiqué figure 2, l'analyse des blocs sélectionnés est similaire à l'analyse effectuée lors de l'indexation ou lors de l'analyse de la question avec, par exemple, une désambiguïsation sémantique des mots polysémiques. Toutefois cette analyse se double d'un calcul de poids pour chacune des phrases, le poids étant fonction du nombre de mots et entités nommées trouvés dans cette phrase, de la présence ou non du type de réponse correspondant à la question, de l'accord entre les thèmes et domaines. C'est ce poids qui permettra ensuite le classement des réponses.

Après analyse, les phrases ou paragraphes semblant répondre à la question sont triés et une analyse complémentaire extrait les entités nommées ou les groupes de mots (parfois des propositions ou des listes) qui correspondent le mieux aux réponses. Cette extraction se fait en fonction des caractéristiques des entités nommées ou sur la base de nature syntaxique des groupes ou propositions.

Pour une question sur un corpus fermé, le temps de réponse est d'environ 3 secondes avec un Pentium 3 GHz. Sur le Web, les premières réponses sont généralement fournies au bout de 2 secondes, un affinage progressif ayant lieu et pouvant durer jusqu'à une quinzaine de secondes, selon le paramétrage utilisateur (nombre de moteurs, nombre de pages analysées, etc.)

3 Évaluation EQueR

QRISTAL a été évalué dans le cadre d'EQueR, campagne d'évaluation des systèmes de Questions-Réponses du projet EVALDA (voir GRAU 2004). Le projet EVALDA et, plus généralement, les projets Technolangue, ont été initiés par les Ministères français de l'Industrie, de la Recherche et de la Culture.

La campagne EQueR a été organisée par l'ELDA (Evaluations and Language resources Distribution Agency, www.elda.org) entre janvier 2003 et décembre 2004. Très similaire dans ses principes aux campagnes TREC-QA (USA) ou NTCIR (Japon), elle a pris la forme de deux tests distincts :

- 500 questions générales, principalement factuelles, sur un corpus journalistique et administratif de 1,5 Go.
- 200 questions, souvent non factuelles, sur un corpus médical d'articles scientifiques et de pages Web d'environ 50 Mo.

Les **500** questions générales se décomposaient en :

- **407** questions factuelles simples (ex: *Comment s'appelle le fils de Juliette Binoche ?)*
- **31** questions dont la réponse est une liste (ex.: *Quels sont les trois pays qui bordent la Bosnie-Herzégovine ?)*
- **32** questions dont la réponse est une définition (ex.: *Qu'est-ce que la NSA ?*)
- **30** questions binaires, à réponse oui ou non (ex.: *La carte d'identité existe-t-elle au Royaume-Uni ?*)

La métrique utilisée pour noter les résultats était le MRR (*Mean Reciprocal Rank*, voir http://trec.nist.gov/data/qa.html), c'est-à-dire 1 pour une réponse exacte en première position, 1/2 pour une réponse exacte en seconde position, 1/3 pour une réponse exacte en troisième position, etc. Seules 5 réponses étaient prises en compte, sauf pour les questions binaires ou une seule réponse justifiée était acceptée. Pour les questions dont la réponse était une liste, la métrique utilisée était le NIAP (*Non Interpolated Average Precision,* voir MONZ, 2003).

Chaque participant pouvait fournir deux fichiers de résultats avec deux grilles d'évaluation : le mode "passages" dans lequel, sur un passage d'au maximum 200 caractères, la réponse était jugée juste si elle était contenue dans ce passage; le mode "réponses exactes" où il fallait donner la réponse exacte et uniquement celle-ci.

Notre système de Questions-Réponses évalué pour EQUER était une version bêta de QRISTAL, et Synapse Développement participait là à sa première campagne d'évaluation de moteurs de questions-réponses.

Sur la tâche générale (500 questions), les résultats des sept participants ont été les suivants :

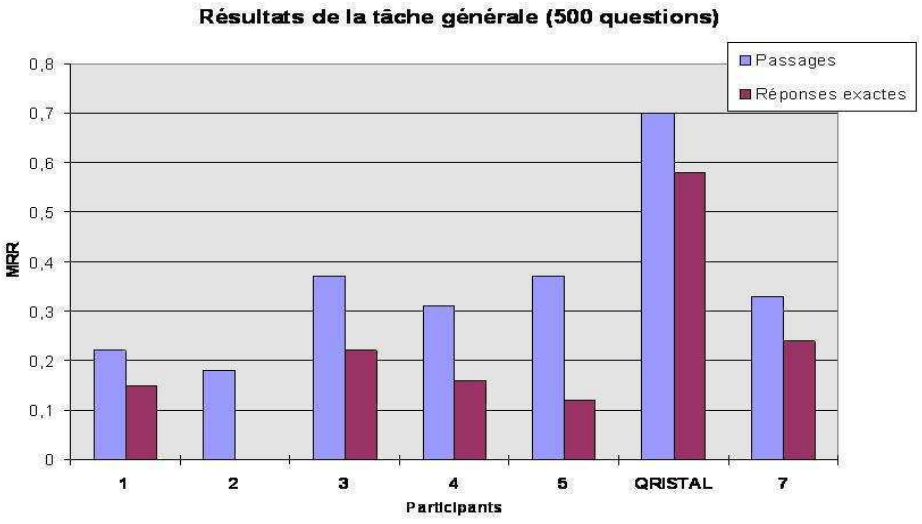


Figure 3. Résultats de la tâche générale

Sur la tâche spécialisée (200 questions sur un corpus médical), les résultats des cinq participants ont été les suivants :

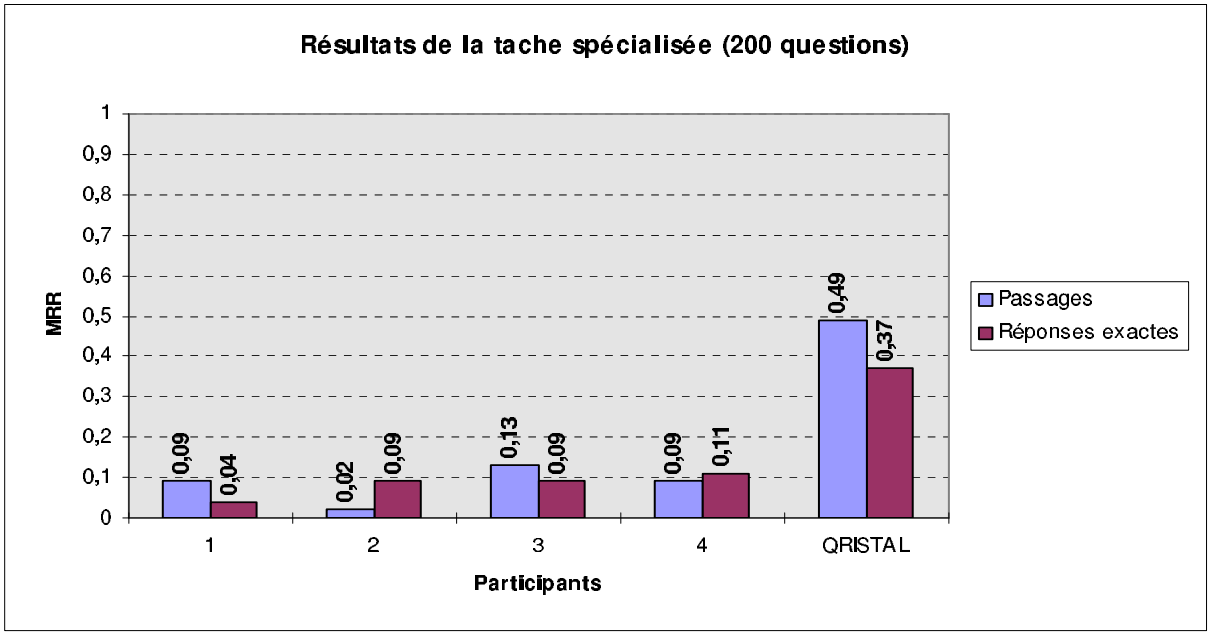


Figure 4. Résultats de la tâche spécialisée

Ces deux graphiques mettent en évidence le bon niveau de performance de QRISTAL, qui obtient les meilleurs résultats parmi les sept systèmes en compétition, sur les deux tâches. Son niveau de performance assez voisin de celui du meilleur moteur américain de TREC (MRR de 0,58 contre 0,68, voir Harabagiu, 2002 et Voorhees, 2003) ou du meilleur moteur japonais de NTCIR (MRR de 0,58 contre 0,61) sur la tâche générale.

Si l'on considère les différentes catégories de questions, QRISTAL présente des résultats homogènes selon les catégories, contrairement aux autres participants (figure 5).

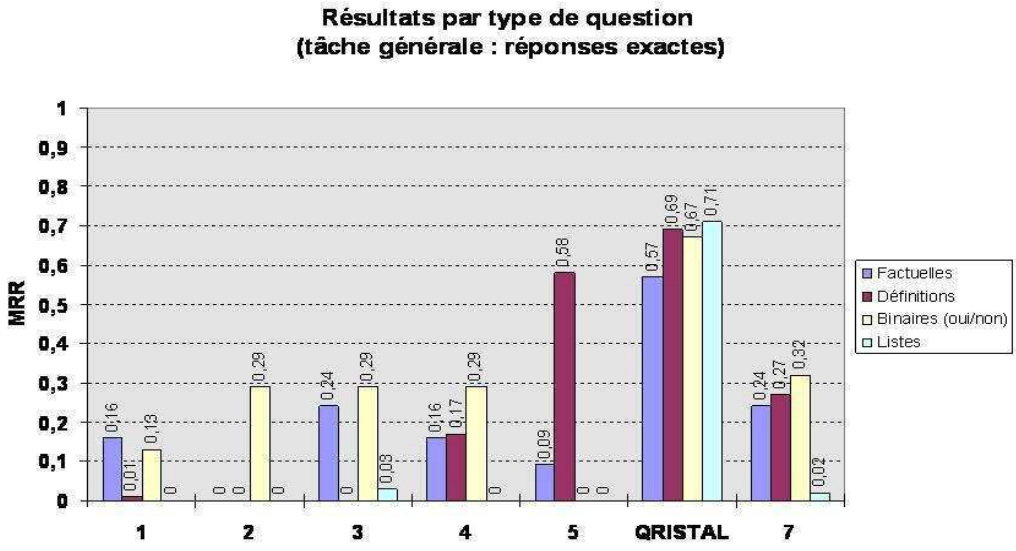


Figure 5. Résultats par type de question

Ces résultats correspondent au meilleur run fourni, celui correspondant à une passe complète d'indexation et d'extraction de réponse dans QRISTAL. Nous avions fourni un autre fichier de résultats dans lequel les textes analysés pour extraction de réponse étaient les textes retournés par un indexeur classique de recherche d'informations. Dans ce second fichier de résultats, les MRR du moteur de Synapse étaient de 0,64 (contre 0,70) pour les passages et de 0,48 (contre 0,58) pour les réponses exactes. Cette différence, significative statistiquement, indique que notre processus d'indexation multicritères est payant, car il permet de remonter de meilleurs textes à de meilleures positions qu'un moteur de recherche d'information classique.

Nous avons par ailleurs effectué un autre test afin d'évaluer l'impact de notre typologie de questions-réponses. En forçant à "inconnu" le type de toutes les questions analysées et en ne prenant pas en compte l'index des types de questions-réponses, nous avons alors obtenu un MRR de 0,46 contre 0,70 pour les passages, démontrant ainsi la pertinence de ce type de variable d'indexation et d'analyse, d'ailleurs déjà largement utilisé en questions-réponses, en particulier par les participants de TREC-QA.

4 Utilisation de QRISTAL

Les réactions des premiers utilisateurs de QRISTAL (quelques centaines après deux mois de commercialisation) permettent de tirer nombre d'enseignements sur la perception des systèmes de questions-réponses, les attentes et les illusions tant sur ces outils que sur les moteurs classiques de recherche sur le Web.

En matière de recherche sur disque dur, les réactions sont assez rares et très souvent positives. La vitesse de recherche est bonne et le fait que le moteur trouve des dérivés ou des synonymes des mots de la question améliore nettement le taux de documents retrouvés. Ainsi, sur un corpus de dépêches d'un mois de l'AFP, les moteurs classiques se révèlent incapables de retrouver le nom du président du Pérou sur l'interrogation "président Pérou" alors que QRISTAL retrouve "Alejandro Toledo" dans "le chef de l'état péruvien, Alejandro Toledo" par un double processus de synonymie et de dérivation.

Compte tenu du très grand nombre de pages indexées sur le Web, ce type d'absence de réponse ne se pose qu'avec de petits corpus, pas sur le Web, sauf pour des questions moins générales, où peu de pages contiennent la réponse. Dans ces cas-là, l'utilisateur conclura en général qu'"il n'y a pas de réponses sur le Web", ne percevant pas que ce silence résulte souvent de l'absence de traitement linguistique des moteurs classiques !

En matière de recherche sur le Web, les moteurs existants donnent le sentiment à l'utilisateur qu'ils disposent de la réponse quasi instantanément. En fait, ces moteurs fournissent des bribes de texte et des liens vers des pages, en aucun cas la réponse exacte à la question posée. Il faut au mieux lire quelques fragments, au pire ouvrir quelques pages, pour espérer obtenir une réponse. Ce processus demande toujours plusieurs secondes, d'autant que la découpe de bribes par les moteurs associe parfois des données issues de phrases différentes (ainsi une demande associant "surface" et un nom de pays donnera rarement la superficie du pays mais plutôt des tailles d'appartements situés dans ce pays). Et ceci n'est valable que pour des questions factuelles ou portant sur des personnes, les questions du type "pourquoi" ou "comment" étant habituellement hors de portée des moteurs de recherche du Web.

Utiliser les moteurs de recherche sur le Web suppose par ailleurs la maîtrise de la syntaxe de ces moteurs (rarement identique). Or cette maîtrise est peu commune. Certes, la plupart des utilisateurs de Google savent qu'il vaut mieux ne saisir que quelques mots, si possible des noms, mais peu connaissent l'usage des guillemets. Pour ces très nombreux utilisateurs, la saisie de questions en langage naturel est un atout de poids.

Reste que de nombreux traitements pouvant permettre d'améliorer la qualité des résultats finaux sont inenvisageables, tout simplement parce que l'utilisateur ne supporte pas d'attendre une réponse plus de trois à quatre secondes. Ainsi nous avions implémenté un dispositif de validation de la réponse en allant interroger à nouveau les moteurs avec les mots de la question et les mots de chacune des différentes réponses possibles (Magnini, 2002). Ce dispositif permettait d'obtenir une amélioration nette : le nombre de bonnes réponses fournies en première position passait de 47% à 58%. Mais le processus de validation demandait six à dix secondes supplémentaires et a donc dû être désactivé.

Améliorer un système de questions-réponses suppose donc une gestion extrêmement serrée du temps machine requis pour chacun des traitements. De sorte qu'il faut choisir avant tout les traitements offrant le meilleur rapport amélioration des résultats/temps machine utilisé.

5 Conclusion

QRISTAL est le premier moteur de questions-réponses commercialisé, auprès du grand public comme des professionnels. Ses résultats lors de l'évaluation EQUER montrent que l'usage intensif de technologies TAL pour l'analyse de la question et des textes indexés, ainsi que pour l'extraction de la réponse, donne de bons résultats, puisque le système se classe premier parmi les sept systèmes évalués.

Ces résultats, même s'ils sont du niveau des meilleurs prototypes internationaux, peuvent toutefois être encore considérés comme insuffisants, particulièrement lors de recherches sur le Web où l'absence d'indexation, l'utilisation d'un métamoteur (donc des résultats renvoyés par les moteurs) et des impératifs de rapidité, rendent plus incertaine l'extraction de réponses correctes dans de nombreux cas.

Même s'il paraît très vraisemblable que, dans quelques années, les moteurs booléens actuels seront remplacés par des moteurs en langage naturel, démontrer les avantages de ce type d'outil et renverser quelques illusions sur les moteurs de recherche actuels demandera du temps, ne serait-ce que parce que les prescripteurs sont souvent des experts en recherche booléenne !

Remerciements

Les auteurs remercient vivement Bruno Wieckowski et l'ensemble des ingénieurs et linguistes ayant participé au développement de QRISTAL

Références

AMARAL C., LAURENT D., MARTINS A., MENDES A., PINTO C. (2004), Design & Implementation of a Semantic Search Engine for Portuguese, *Proceedings of the Fourth Conference on Language Resources and Evaluation.*

CLARKE C. L. A., CORMACK G. V., LYNAM T. R. (2001), Exploiting Redundancy in Question Answering, *Proceedings of 24th Annual International ACM SIGIR Conference (SIGIR 2001)*, p. 358-365.

GRAU B.. (2004), L'évaluation des systèmes de question-réponse, *Évaluation des systèmes de traitement de l'information,* TSTI, p. 77-98, éd. Lavoisier.

HARABAGIU S., MOLDOVAN D., CLARK C., BOWDEN M., WILLIAMS J., BENSLEY J. (2002), Answer Mining by Combining Extraction Techniques with Abductive Reasoning, *Proceedings of The Twelfth Text Retrieval Conference (TREC 2003).*

LAURENT D., VARONE M., AMARAL C., FUGLEWICZ P. (2004), Multilingual Semantic and Cognitive Search Engine for Text Retrieval Using Semantic Technologies, *First International Workshop on Proofing Tools and Language Technologies*, Patras, Grèce.

MAGNINI B., NEGRI M., PREVETE R., TANEV H. (2002), Is It the Right Answer? Exploiting Web Redundancy for Answer Validation, *Proceedings of the 40th Annual Meeting of the ACL*, p. 425-432

MONZ C. (2003), From Document Retrieval to Question Answering, *ILLC Dissertation Series 2003-4*, ILLC, Amsterdam.

VOORHEES E. M.. (2003), Overview of the TREC 2003 Question Answering Track, NIST, 54-68 (http://trec.nist.gov/pubs/trec12/t12_proceedings.html).

Morphosémantique pour l'appariement de termes dans le vocabulaire médical : approche multilingue

Fiammetta Namer

UMR "ATILF" CNRS & Université Nancy2

Mots Clefs : morphologie, sémantique, multilinguisme, composition savante, relation lexicale, terminologie médicale

Keywords: morphology, semantics, multilingualism, neoclassical compounding, lexical relation, medical terminology

Résumé : Cet article s'intéresse à la manière dont la morphosémantique peut contribuer à l'appariement multilingue de variantes terminologiques entre termes. L'approche décrite permet de relier automatiquement entre eux les noms et adjectifs composés savants d'un corpus spécialisé en médecine (synonymie, hyponymie, approximation). L'acquisition de relations lexicales est une question particulièrement cruciale lors de l'élaboration de bases de données et de systèmes de recherche d'information multilingues. La méthode est applicable à au moins cinq langues européennes dont elle exploite les caractéristiques morphologiques similaires des mots composés dans les langues de spécialité. Elle consiste en l'intéraction de trois dispositifs : (1) un analyseur morphosémantique monolingue, (2) une table multilingue qui définit des relations de base entre les racines gréco-latines des lexèmes savants, (3) quatre règles indépendantes de la langue qui infèrent, à partir de ces relations de base, les relations lexicales entre les lexèmes contenant ces racines. L'approche décrite est implémentée en français, où l'on dispose d'un analyseur morphologique capable de calculer la définition de mots construits inconnus à partir du sens de ses composants. Le corpus de travail est un lexique spécialisé médical d'environ 29000 lexèmes, que le calcul des relations de synonymie, hyponymie et approximation a permis de regrouper en plus de 3000 familles lexicales.

Abstract: This paper addresses the issue of the interaction between morphosemantics and term variants extraction. The described method enables neoclassical compound nouns and adjectives of a biomedical specialized corpus to be automatically related by synonymy, hyponymy and approximation links. Acquiring lexical relations is a particularly crucial issue when elaborating multilingual databases and when developing cross-language information retrieval systems. This method can be applied at least to five European languages and exploits the similarity between the morphological characteristics of compound words in specialized domains. It requires the interaction of three techniques: (1) a language-specific morphosemantic parser, (2) a multilingual table defining basic relations between word roots, and (3) a set of language-independant rules to draw up the list of related terms. This approach has been fully implemented for French, on an about 29,000 terms biomedical lexicon, resulting to more than 3,000 lexical families.

1 Variation terminologique et morphologie

Dans le domaine bio-médical, comme dans toute langue de spécialité, l'extraction de variantes terminologiques constitue un enjeu important (Bourigault *et al.* 2001). L'objectif que la démarche présentée ici vise à atteindre, est de mettre à contribution la morpho-sémantique pour maîtriser l'appariement terminologique bilingue, voire translinguistique. Le but est l'enrichissement des relations entre termes dans les bases multilingues de connaissances. L'interrogation de ressources hétérogènes (bases de données, notices bibliographiques etc.)

dans plusieurs langues est une préoccupation constante dans les domaines de spécialité, qui a conduit au développement de plusieurs techniques pour l'établissement de terminologies multilingues. L'extraction terminologique et l'alignement de corpus parallèles (Gaussier 2001), sont deux étapes classiques dans la conception de tels systèmes (voir l'expérience de (Tran *et al.* 2003)). En matière de synergie entre terminologie et morphologie, différentes études et applications existent. Certaines se basent sur la reconnaissance de séquences au moyen de patrons (Daille 2001; Jacquemin, Tzoukermann 1999) d'autres utilisent plutôt des systèmes statistiques fondés sur l'apprentissage de règles (Hathout 2003; Grabar, Zweigenbaum 2000). Enfin, des travaux ont été menés dans le but de faire coopérer morphologie et terminologie bilingue, entre autre par (Chiao, Zweigenbaum 2003). Notre approche se situe plutôt dans la lignée des systèmes basés sur l'application de contraintes. L'analyseur morphologique DériF[1] (Namer 2003), qui constitue l'une des étapes de notre système, a été récemment adapté pour l'analyse du vocabulaire bio-médical, dans le cadre des projets UMLF et Vumef. Comme nous allons le voir, DériF se fonde sur la transposition de connaissances linguistiques pour apparier les lexèmes spécialisés au moyen de relations de trois types : (1) il relie le lexème analysé à sa base (*bactérien, bactérie*) même si celle-ci est d'origine gréco-latine (*hépatique, foie*), (2) ce lien est annoté au moyen d'une pseudo-définition (*amyotrophie* : "absence de développement des muscles") ; (3) grâce à l'adaptation des ressources spécifiques au domaine bio-médical, DériF calcule les relations de synonymie, hyponymie et approximation[2] entre les mots composés dits-savants (Fradin 2000; Warren 1990) : *hystérorragie* y est vu comme synonyme de *métrorragie*, hyponyme de *hystérorrhée*, voisin de *colporragie* ; l'intérêt de ces composés savants est qu'ils constituent à eux seuls près de la moitié des néologismes recensés dans les textes médicaux (Lovis et al. 1998). La possibilité de bâtir une méthode translinguistique vient du fait que, contrairement à la langue générale, la morphologie des lexèmes spécialisés obéit à des règles constructionnelles extrêmement proches dans toutes les langues européennes (Iacobini 2003). Pour les mêmes raisons, la démarche multilingue s'applique au calcul des liens lexicaux entre mots composés savants du vocabulaire médical[3] ; trois types de ressources intéragissent : un analyseur morphologique, une table établissant des relations de base entre les racines gréco-latines pouvant entrer dans la formation de mots, et un système de règles calculant les relations lexicales entre les termes. Alors que la conception d'un analyseur est une tâche qui doit être réitérée pour chaque nouvelle langue, nous allons voir que la table est une donnée unique multilingue et que le système de règles est indépendant de la langue choisie. L'approche a été implémentée en français, sur un lexique d'environ 29000 termes, et donne lieu à l'émergence d'environ 3000 familles lexicales. L'article s'organise comme suit. Nous présentons tout d'abord (§2) les connaissances et données sur lesquelles repose l'approche morphologique pour la définition multilingue de relations lexicales entre termes. Ensuite, (§3) nous développons la méthode utilisée pour réaliser cet objectif, et nous présentons (§4) les résultats obtenus en français. Ces résultats conduisent naturellement à une discussion et à des perspectives (§5) qui clôtureront cette présentation.

2 Genèse

Comme annoncé en §1, la méthode proposée s'appuie sur la synergie entre un analyseur morphologique basé sur règles, une table qui classe et annote les racines gréco-latines

---

[1] DériF a été développé lors des projets ACI MorTAL (G. Dal, CNRS) et UMLF (P. Zweigenbaum, INSERM), et RNTS VumeF (S. Darmoni, L@stics et JF Forget, Vidal) (Zweigenbaum et al. 2003; Darmoni et al. 2003).

[2] L'approximation (voisinage) subsume les notions de co-méronymie, de co-hyponymie et de compatibilité.

[3] Les exemples sont donnés en français, italien, espagnol, allemand anglais, notés respectivement : FR, IT, ES, DE,EN

utilisées dans les termes médicaux, et des règles de calcul de relations lexicales entre mots composés savants. Un certain nombre de constatations sont à l'origine de cette démarche qui est à la fois indépendante de la langue de travail, et spécifique aux domaines de spécialité proches du biomédical. (1) Les théories en morphologie lexicale[4] permettent de déduire la définition d'un mot morphologiquement complexe en fonction de celui de ses constituants. Donc, un système implémentant une telle approche théorique (comme DériF, cf. §4) est à même de calculer la pseudo-définition de mots inconnus à partir des procédés morphologiques mis en œuvre. (2) Quelle que soit la langue européenne considérée, les mots complexes en biomédecine contiennent dans leur grande majorité des racines gréco-latines (*gastr-, -phage, -hydr-*), qu'à la suite de (Haspelmath 2002) entre autres, nous nommerons éléments de formation, notés EFs. Un EF partage sa catégorie et son sens avec l'entrée lexicale contemporaine auquel il supplée (ainsi, *gastr-* signifie *estomac*$_{FR}$, et son type catégoriel est NOM). D'une langue à l'autre, la réalisation des EFs ne présente que de légères variations graphiques, et leur emploi dans la formation de termes de spécialité met en jeu des règles quasiment identiques (Iacobini 2003). Il en résulte que les EFs et les structures de mots complexes peuvent avantageusement être représentés par des symboles abstraits, qui gomment les différences entre les langues. Ainsi, le terme abstrait VASCUL--ITE[5] correspond à *vascul--ite*$_{FR}$, *Vascul--itis*$_{DE}$, *vascol--ite*$_{IT}$ et *vascul--itis*$_{ES/EN}$. (3) La dernière observation qui sous-tend cette approche, peut-être la plus importante, est l'exploitabilité des systèmes internationaux de classification (SNOMED, CIM-10, MesH), qui organisent la terminologie médicale au moyen notamment de relations lexicales (synonymie, méronymie, (co)hyponymie…). L'identité entre un EF et sa traduction rend transposables ces systèmes classificatoires pour l'organisation hiérarchique des EFs : de la même façon que *estomac* est une partie du *ventre*, tous deux étant décrits dans le chapitre *anatomie*, GASTR est une partie de ABDOMIN, les deux EFs se trouvant également sous le descripteur *anatomie*. On établit alors quatre types de relations lexicales entre les EFs : synonymie, notée = (OPT=OPHTALM, *vision*), hyponymie, notée < (BLAST *cellule embryonnaire* < CYT *cellule*), méronymie, notée ← (CORO *pupille* ← OCUL *œil*) et approximation, notée ~ (RHIN *nez* ~ OTO *oreille*).

3 Démarche

Rappelons que notre objectif est l'appariement multilingue de termes médicaux composés savants au moyen de relations lexicales calculées au cours de l'analyse morphologique de ces termes. Notre approche s'articule autour de trois types de données et techniques, qui répondent aux observations faites en §2 : un ensemble réduit de règles générales (§3.3) infèrent des relations lexicales entre les mots composés d'un corpus à partir de relations de base établies entre les EFs qui constituent ces termes, et réunies dans une table (§3.2) ; enfin, l'identification de ces EFs requiert l'intervention d'un analyseur morphologique (§3.1).

3.1 Analyseur Morphologique monolingue

Le processus de décomposition d'un lexème complexe en constituants est une tâche monolingue, dévolue à un analyseur morphologique qui peut fonctionner selon des approches diverses, allant de la simple segmentation (Lovis et al. 1995) à l'application de contraintes permettant d'annoter les résultats d'informations sémantiques (Namer 2003). Les résultats des analyseurs basés sur contraintes ont l'avantage d'associer à une décomposition hiérarchique la

[4] Nos travaux suivent des hypothèses liées à une morphologie de type léxématique, où sens et structure se calculent conjointement, et constituent une adaptation de la théorie élaborée à l'origine dans (Corbin 1987).

[5] Les EFs abstraits sont écrits en petites majuscules, les frontières entre EFs sont représentés par '--'

définition du mot analysé en fonction du procédé morphologique identifié. Ainsi, le sens d'un lexème obtenu par affixation est calculé à partir de celui de sa base, via la traduction de celle-ci, lorsqu'elle est réalisée sous forme d'EF : *hépatique*$_{ADJ}$ = "en relation avec le foie". Comme l'illustre la Fig.1, les EFs apparaissent très fréquemment dans la formation de termes suffixés, préfixés ou composés, toutes langues confondues. Contrairement à l'affixation, la composition construit un nom ou un adjectif en associant deux constituants (chacun peut être un lexème autonome ou un EF, et le cas échéant, d'origine grecque ou latine). Dans les langues romanes, la composition savante (*saxifrage*$_A$) se distingue de la composition dite populaire (*casse-pierre*$_A$) par la place occupée par le constituant tête (noté X), placé à droite du constituant modifieur, noté Y. Le sens du composé est fonction, entre autres, de sa catégorie et du rapport sémantique entre X et Y : le composé peut être de type **(a)** additif (*buccodentaire*$_A$ caractérise ce "qui concerne la bouche : *bucc* et les dents), **(b)** endocentrique (*gastralgie*$_N$ est hyponyme de douleur : *algie*, et affecte l'estomac : *gastr*) ou **(c)** exocentrique (*brachycéphale*$_A$ n'est pas hyponyme de tête : *céphal*(e), mais désigne ce(lui) "qui a une tête : *céphal* courte : *brachy*").

Lang.	Affixation[6]	*traduction*	EF abs.	Composition (type)	*traduction*	EFs abs
IT	epat#ico	*hépatique*	HEPAT	gastro--ectomia **(b)**	*gastrectomie*	GASTR, ECTOMI
FR	bucc#al	*buccal*	BUCC	bucco--dent(aire) **(a)**	*buccodentaire*	BUCC
EN	an#alges(ic)	*analgésique*	ALGES	thermo--algesia **(b)**	*thermoagésie*	THERM, ALGES
DE	Hypo#thermie	*hypothermie*	THERM	Thermo--taxis **(b)**	*thermotaxie*	THERM, TAXI
ES	intra#cefal(ico)	*intracéphalique*	CEPHAL	braqui--cefalo **(c)**	*brachycéphale*	BRACHY,CEPHAL

Figure 1 : Eléments de Formation et procédés morphologiques

3.2 Table multilingue des Eléments de Formation

EF (1)		Instanciation (2) Anglais	Allemand	Français[7]	Italien	Espagnol	CAT (3)	Chapitre SNOMED (4)	Relation lexicale (5)
GASTR	réal	gastr	Gastr	gastr	gastr	gastr	N	*ANATOMIE*	=STOMAC, ←ABDOMIN, ~HEPAT, ~ENTER, ~PANCREAT
	trad	stomach	Magen	estomac	stomaco	estomago			
ALGI	réal	algia/algy	algie	algie	algia	algia	N	*SYMPTOME*	=ODYN, ~ITE
	trad	pain	Schmerz	douleur	dolore	dolor			
ITE	réal	itis	ite	ite	ite	itis	N	*SYMPTOME*	~ALGI, ~ODYN
	trad	inflammation	Inflammation	inflammation	infiammazione	inflamación			
PHLEB	réal	phleb	Phleb	phléb	fleb	fleb	N	*ANATOMIE*	=VEN, <ANGI, <VASCUL
	trad	vein	Vene	veine	vena	vena			
ANGI	réal	angio	Angio	angio	angio	angio	N	*ANATOMIE*	=VASCUL, ~VAS
	trad	blood vessel	Blutader	vaisseau sanguin	vaso sanguigno	vaso sanguíneo			
ECTOMI	réal	ectomy	ektomie	ectomie	ectomia	ectomía	N	*ACTE MEDICAL*	~TOMI, ~STOMI
	trad	ablation	Ablation	ablation	ablazione	ablación			

Figure 2 : Table multilingue des Eléments de Formation (échantillon)

Les observations (2) et (3) du §2 conduisent tout naturellement à la conception d'une table réunissant l'ensemble des quelques 900 EFs utilisés dans le vocabulaire biomédical, et dont la Fig. 2 donne un échantillon. A chaque représentation abstraite d'un EF (col.1) correspondent

[6] Les frontières base-affixe sont marquées '#'

[7] Les réalisations indiquées possèdent des variantes allomorphiques codées également dans la Table quand elles reflètent des situations morphologiquement pertinentes. Ainsi, *algo* et *algés* sont des variantes de *algie* ne pouvant occuper que la position Y dans un composé.

sa catégorie grammaticale (col.3), la tête de chapitre SNOMED où il apparaît (col.4), et les relations lexicales de base (col.5) dont les symboles sont expliqués en §2, et que l'EF abstrait entretient avec d'autres EFs abstraits présents dans la table. Par exemple, GASTR est synonyme de STOMAC, appartient à ABDOMIN, et a à voir avec HEPAT (*foie*), ENTER (*intestin*) et PANCREAT (*pancréas*). Enfin, la col.2 décrit les instances de l'EF pour chaque langue prise en compte : chaque instance couple la réalisation du symbole abstrait, avec sa traduction. Ainsi, ALGI est instancié par *algia/algy*$_{EN}$: 'pain', *algie*$_{DE}$: 'Schmerz', *algie*$_{FR}$: 'douleur', *algia*$_{IT}$: 'dolore', *algia*$_{ES}$: 'dolor'. L'ajout d'une nouvelle langue dans le système suppose donc uniquement l'insertion d'une nouvelle sous-colonne dans la col.2.

3.3 Règles indépendantes de la langue pour le calcul des relations lexicales

La projection des relations lexicales entre EFs (Fig.2), sur les noms et adjectifs composés dont l'analyse morphologique fait apparaître ces EFs, requiert l'activation de l'une des quatre règles indiquées dans la Fig. 3. Ces règles sont totalement indépendantes de la langue. Chacune est décrite formellement dans la col. 1, et exemplifiée dans les colonnes suivantes. La règle **R2**, par exemple, établit que tout couple de composés A et B dont les constituants X_A et X_B sont synonymes, entretiennent la même relation R que celle établie entre Y_A et Y_B, sauf si R est la relation de méronymie : si Y_A est une partie de Y_B en effet, A est hyponyme de B. A titre d'exemples, la synonymie entre MORT et THANAT (*mort*) se propage entre les adjectifs *mortifero*$_{IT}$ et *tanatogeno*$_{IT}$, la relation d'hyponymie entre *apivore*$_{FR}$ et *entomophage*$_{FR}$ provient de celle entre API (*abeille*) et ENTOMO (*insecte*), alors que celle entre *Enterodynie*$_{DE}$ et *Abdominalgie*$_{DE}$ résulte de la méronymie entre ENTER (*intestin*) et ABDOMIN (*abdomen*). Enfin, l'approximation entre BACTERI et BACILL entraîne celle entre *bacilliform*$_{EN}$ et *bacterioid*$_{EN}$. La règle **R4** est symétrique à **R2**, en ce que Y_A et Y_B y sont synonymes, et la relation entre A et B dépend alors de celle qu'entretiennent X_A et X_B. Enfin **R1** (resp. **R3**) est la version simplifiée de **R2** (resp. **R4**), où A et B partagent le constituant X (resp. Y) [8].

Règle	Exemple		
	Y	X	$[Y_AX_A]$ R $[Y_BX_B]$
R1 **A = $[Y_AX]$ et B=$[Y_BX]$** Si $Y_A \leftarrow Y_B$ alors A < B sinon si Y_A R Y_B et R est { =, <, ~} alors A R B	PROCTO ← COLO LEUCO ← HEMATO ABDOMIN=LAPAR ALBUMIN<PROTEIN XER ~SCLER	RRAGIE GRAMME SCOPIE EMIE OPHTALMIE	**EN:** proctorrhagia < colorrhagia **DE:** Leukogramm < Hämatogramm **FR:** abdominoscopie = laparoscopie **IT:** albuminemia < proteinemia **ES:** xerophthalmia ~sclerophthalmia
R2 **A = $[Y_AX_A]$ et B=$[Y_BX_B]$ et $X_A = X_B$** si $Y_A \leftarrow Y_B$ alors A < B sinon si Y_A R Y_B et R est { =, <, ~} alors A R B	ENTER←ABDOMIN MORT = THANAT API < ENTOMO BACILL ~BACTERI	**$X_A = X_B$** ALGIE = ODYNIE FERE = GENE VORE = PHAGE FORME = OÏDE	**DE:** Enterodynie < Abdominalgie **IT:** mortifero = tanatogeno **FR:** apivore < entomophage **EN:** bacilliform ~bacterioid
R3 **A = $[YX_A]$ et B=$[YX_B]$** Si X_A R X_B et R est {=, <, ~} alors A R B	BACTER OTO ARTHR	OÏDE = FORME RRAGIE < RRHEE ALGIE ~ITE	**FR:** bactérioïde = bactériforme **DE:** Otorrhagie < Otorrhö **ES:** artralgia ~artritis
R4 **A = $[Y_AX_A]$ et B=$[Y_BX_B]$ et $Y_A = Y_B$** si X_A R X_B et R est {=, <, ~} alors A R B	**$Y_A = Y_B$** ORTHO = RECTI METR = HYSTER LIP = ADIP	DONTE = DENT RRAGIE < RRHEE MATOSE ~OME	**FR:** orthodonte = rectident **FR:** métrorragie < hystérorrée **EN:** lipomatosis ~adipoma

Figure 3 : Règles de Calcul des Relations Lexicales

L'intéraction entre un analyseur morpho-sémantique, la table multilingue des EFs et les règles

[8] Une version monolingue des règles et de la table des EF est présentée dans (Namer, Zweigenbaum 2004).

de calcul des relations lexicales résulte en une chaîne de traitement qui conduit à l'appariement des mots composés savants au moyen des relations lexicales de synonymie =, hyponymie < et approximation ~. L'analyseur décompose le lexème d'entrée, en identifiant s'il y a lieu, les EFs qui le constituent[9]. Ces EFs servent à alimenter le système des règles de calcul des relations lexicales : pour chaque EF, rapporté à sa structure abstraite, l'ensemble des relations lexicales de base définies dans la table est collecté. Les règles **R1** à **R4** sont activées, et prédisent toutes les relations potentielles abstraites avec l'input. La dernière tâche à effectuer consiste alors à filtrer les relations correspondant à des mots inexistants dans le corpus dans lequel les appariements sont calculés. C'est cet enchaînement, réalisé en français sur un lexique de grande taille, qui fait l'objet du prochain paragraphe.

4 Résultats pour le français

L'approche décrite ci-dessus a été implémentée en français. Les résultats ont été obtenus à partir d'un lexique totalisant 29000 noms, adjectifs et verbes du vocabulaire spécialisé, collectés à partir de diverses sources, librement accessibles en ligne, ou mises à la disposition des projets UMLF et VumeF [10]. La réalisation de la chaîne de traitement en français est rendue possible avant tout par l'existence de l'analyseur morpho-sémantique DériF ("Dérivation en français"). L'analyse par DériF d'un lexème catégorisé adapte les hypothèses théoriques avancées à l'origine dans (Corbin 1987). Basé sur l'application d'un système ordonné de règles, le mécanisme est récursif et permet la gestion des ambiguïtés, se réappliquant sur chaque (liste de) résultat obtenu précédemment. L'analyse morphologique d'un lexème construit sur une base elle-même construite est donc est hiérarchisée. Le résultat est un triplet, la première partie retrace sous forme crochetée l'historique des étapes d'analyse, la seconde réunit les lexèmes résultats obtenus à chaque étape, et la troisième est constituée d'une formulation en langue naturelle de la relation morphologique liant l'input à son (ses) constituant(s) immédiat(s). Les néologismes sont analysés et pseudo-définis comme des mots régulièrement construits (ce qui est généralement le cas). Quand il analyse un mot composé, enfin, DériF fournit une représentation linéaire Y/X de la décomposition de celui-ci en constituants. Le fonctionnement ainsi résumé de DériF est illustré par l'analyse de *gastralgie*$_{NOM}$, dans les 3 premières lignes de la Fig.4. On note que la définition calculée pour *gastralgie* mobilise la table des EFs qui fournit la traduction, respectivement de *gastr* (estomac) et *algie* (douleur). Les représentations abstraites de Y et X (**'Constituants'**, Fig.4, ligne 4) sont transmises au système de calcul des relations lexicales. Comme cela a été mentionné en §3.3, les quatre règles **R1** à **R4** sont activées pour produire les relations lexicales candidates de l'input (i.e. dans l'exemple, *gastralgie*). Tout d'abord, (**R1**) X est conservé, et Y est remplacé par tous les EFs trouvés dans la table avec lesquels Y est en relation ; ceux-ci sont restitués sous leur forme de réalisation en français, et la relation potentielle est calculée selon **R1** (e.g. dans la Fig. 4 `eql:stomach/algie`[11]) ; ensuite, (**R2**), X est remplacé par chacun de ses synonymes dans la Table des EFs, et l'opération de substitution de Y est identique à ce qui se passe avec **R1** (e.g. `isa:abdomin/odynie`) ; puis les rôles de Y et X sont inversés, lors de l'activation de **R3** (e.g. `see:gastr/ite`) et de **R4** (e.g. `see:stomach/ite`).

---

[9] Selon le type d'analyseur, l'analyse morphologique de l'input fournit éventuellement aussi une pseudo-définition, sous-forme de relation entre l'input et ses composants.

[10] Pour ne mentionner que quelques sources : les versions françaises de la CIM-10, du MesH et le dictionnaire en ligne BIOTOP (URL : http://georges.dolisi.free.fr/Terminologie/Menu/terminologie__medicale_menu.htm)

[11] L'affichage par DériF des relations lexicales possibles est de la forme `'R:Y/X'`; R symbolise la synonymie = par `'eql'`, l'hyponymie < par `'isa'` et l'approximation ~ par `'see'`.

```
      gastralgie/NOM==> [ [ gastr N* ] [ algie N* ]  NOM ]
      (gastralgie/NOM, algie/N*)
      " douleur (du -- liée au) estomac "
Constituants = /gastr/algie/
Type = maladie
Relations possibles = (eql:gastr/algo, eql:gastr/algés, eql:gastr/odyn,
eql:stomac/algie, eql:stomac/algo, eql:stomac/algés, eql:stomac/odyn,
eql:stomach/algie, eql:stomach/algo, eql:stomach/algés, eql:stomach/odyn,
isa:abdomin/algie, isa:abdomin/algo, isa:abdomin/algés, isa:abdomin/odyn,
see:entéro/algie, see:entéro/algo, see:entéro/algés, see:entéro/odyn,
see:gastr/ite, see:hépat/algie, see:hépat/algo, see:hépat/algés,
see:hépat/odyn, see:pancréat/algie, see:pancréat/algo, see:pancréat/algés,
see:pancréat/odyn, see:stomac/ite, see:stomach/ite)
```

Figure 4 : Relations lexicales candidates pour *gastralgie*$_{NOM}$

A partir de cet ensemble de relations candidates, le système ne garde que les relations concernant les termes attestés. Pour *gastralgie,* et étant donné le contenu du lexique de 29000 entrées du français, on s'attend à ce que seuls les éléments soulignés correspondent à des lexèmes "réels". Les autres sont soient morphologiquement impossibles (`gastr/algés`, par exemple ne peut pas se réaliser, car *algés* est une forme que l'on ne trouve qu'en position Y), soit non attestés (dans le corpus du moins) : c'est par exemple le cas de `pancréat/odyn`, car *pancréatodynie* n'est pas dans notre lexique. Etant donné un composé A (ex. *gastralgie*) l'identification de ses relations lexicales 'réelles' s'effectue au moyen du couple Y/X, calculé par DériF pour chaque entrée B du lexique et consigné en valeur du trait 'Constituants'. Quand pour un input B donné, Y/X s'identifie à l'un des candidats de la listes des relations possibles de A, B est ajouté à la liste des relations attestées de A. A la fin de cette étape, chaque input A du lexique de travail se voit associé sa famille lexicale, regroupant l'ensemble des composés du corpus avec lesquels A entretient l'une des relations de synonymie, hyponymie et approximation. La Fig. 5 reproduit la famille lexicale de *gastralgie*.[12]

```
(11565) gastralgie/NOM (maladie)    "  douleur (du -- liée au) estomac  "
       gastralgie/NOM: synonym of  gastrodynie/NOM, stomacalgie/NOM, stomacodynie/NOM,
stomachodynie/NOM, (gastralgique/ADJ)
       gastralgie/NOM: subtype of  abdominalgie/NOM
       gastralgie/NOM: see also  entéralgie/NOM, entérodynie/NOM, gastrite/NOM,
hépatalgie/NOM, hépatodynie/NOM, pancréatalgie/NOM
```

Figure 5 : Famille lexicale de *gastralgie*

Les modules d'analyse de DériF implémentent à ce jour divers procédés morphologiques, que ce soit la suffixation, la préfixation, la conversion ou la composition savante. DériF est actuellement à même d'analyser comme complexes 17240 des 29000 lexèmes du corpus de travail[13]. La chaîne de traitement enfin produit plus de 3000 familles lexicales à partir des lexèmes composés du corpus, générant au total des liens entre 7438 ADJs et/ou NOMs distincts.

5 Discussion, perspectives

L'utilisation de la morphologie des mots composés dans le but d'optimiser la recherche d'information en biomédecine a déjà fait l'objet d'expérimentations, entre autre par (Schulz et

[12] D'autres termes sont ajoutés à la famille suivant des crières morphologiques. Notamment, un adjectif relationnel (e.g. *gastralgique*) est considéré comme 'synonyme' de son nom base.

[13] Les lexèmes complexes non analysés sont ceux formés suivant des patrons constructionnels non encore (complètement) intégrés dans DériF.

al. 1999), et (Hahn et al. 2001), qui se servent également d'EFs (qu'ils appellent 'subwords'). Cependant, contrairement à ce qui est présenté ici, ils n'exploitent pas les relations lexicales entre les EFs (donc ne calculent pas de relations lexicales), et leur analyse morphologique est réduite à un simple découpage linéaire, qui ne permet pas d'associer une définition à l'input. En contrepartie, bien entendu, notre approche présente un inconvénient majeur, qui est celui de tout système basé sur l'utilisation de contraintes linguistiques, et demandant la gestion des exceptions. Il nécessite une validation humaine à trois niveaux au moins : pour vérifier la pertinence des analyses, pour valider les pseudo-définitions et surtout pour contrôler les relations lexicales de base dans la table des EFs. Notamment, il faut éviter des annotations trop spécifiques sur des EFs polyréférentielles en médecine. Ainsi, étiqueter LABI (*lèvre*) comme partie-de BUCC (*bouche*) entraînerait un codage pour le moins curieux d'adjectifs comme *inguino-labial*, relatif à la gynécologie.[14]

Les améliorations prioritaires de la démarche présentée (en dehors de l'évolution de DériF, qui ne concerne que le français) passent tout d'abord par l'ajout de nouvelles règles de calcul des relations lexicales, qui s'appliquent aux termes préfixés et/ou suffixés. Elles permettront par exemple d'identifier *gastrique* comme une propriété synonyme de *stomacal*, et hyponyme d'*abdominal*. Dans la table des EFs, certaines relations d'approximation pourraient se spécialiser. Certains EFs constituent en effet des pôles opposés d'une même propriété : e.g. BRACHY *court*, versus DOLICHO *long*. Enfin, mais cela conduira à ajouter un nouveau module monolingue au système, on pourrait générer automatiquement (dans la langue de son choix) les termes correspondant aux relations lexicales possibles d'un terme A, absents du corpus de travail et morphologiquement plausibles: pour le français, cela reviendrait, par exemple (Fig. 4), à générer *abdominodynie, pancréatodynie, stomac(h)ite* qui sont non seulement absents du corpus de travail mais aussi introuvables sur Internet.

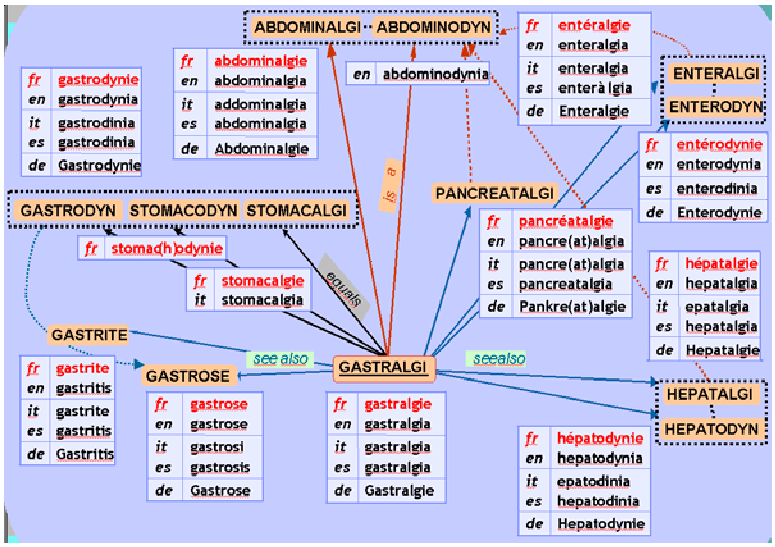


Figure 5 : Une Famille Lexicale Abstraite: celle de GASTRALGI

La réalisation dans d'autres langues que le français[15] de l'approche présentée ne nécessite pratiquement que de disposer d'un parseur morphologique pour chaque nouvelle langue. En un premier temps, celui-ci peut être extrêmement rudimentaire (e.g. un simple raciniseur)

[14] Plus généralement, les EFs ambigus nécessitent un traitement par DériF qui exploite des listes d'exception : e.g. *péd* signifie pied (*pédologue, pédicure, pédestre*) ou enfant (*pédagogue, pédophile, pédiatre*).
[15] L'ébauche de ce travail d'extension est actuellement en cours pour l'anglais.

dans la mesure où sa tâche fondamentale est d'identifier les composants X et Y d'un composé savant. Une fois cet analyseur disponible e.g. pour les cinq langues qui nous ont servi à illustrer notre démarche, la chaîne de traitement (à l'exception de l'analyseur) ne manipule plus que des données abstraites. On obtient alors un ensemble de familles lexicales abstraites à l'image de ce qu'illustre la Fig. 5. Dans chacune d'elles, les noms partageant la même structure et les mêmes composants dans différentes langues sont identifiés par une étiquette abstraite ; ce sont ces étiquettes qui sont alors reliées entre elles par les relations lexicales selon le même mécanisme que celui que nous avons décrit pour le français au §4.

6 Conclusion

Nous avons décrit une méthode permettant de regrouper les noms et adjectifs composés savants du langage biomédical selon des liens sémantico-lexicaux, grâce à une classification multilingue de base (la table des EFs) établie à partir des terminologies internationales du domaine médical. Quelques règles indépendantes de la langue servent à propager ces relations de base sur les composés qui contiennent ces EFs, pour calculer les relations lexicales qu'entretiennent les composés entre eux. Les résultats obtenus en français sont utilisés pour étendre le système d'extraction de variantes terminologiques à de nouveaux liens simples : *maladie du foie / maladie hépatique* mais aussi plus complexes : *traitement contre la douleur à l'estomac / traitement antigastralgique*. Nous testons également la réutilisabilité des règles de calcul des relations lexicales pour établir des liens de synonymie, hyponymie et approximation entre les termes polylexématiques. L'idée est de vérifier la validité des appariements du type : *douleur à l'estomac* < *douleur au ventre*. On pourrait également envisager une utilisation en analyse du discours des relations d'hyponymie (*abdominalgie* < *douleur*) pour la recherche des liens anaphoriques[16].

Les applications multilingues de la démarche présentée (selon Fig. 5) sont pour la plupart immédiatement concevables : question-réponse multilingue, recherche d'information, enrichissement de bases de connaissances translinguistiques… Une autre utilisation est la traduction par voisinage, qu'illustre la Fig.5. Chaque étiquette abstraite y regroupe les noms qui ont été effectivement rencontrés dans les corpus spécialisés de chaque langue. On note, à ce sujet, que *abdominodynia*$_{EN}$ tout comme *stomachodynie*$_{FR}$ sont des structures qui ne se rencontrent que dans une langue. Cependant, leur traduction est calculable immédiatement via le lien de synonymie qui part de leur étiquette abstraite ; les traductions indirectes de e.g. *stomachodynie*$_{FR}$ sont donc : *stomacalgia*$_{IT}$, *Gastrodynie*$_{DE}$, *gastrodinìa*$_{ES}$, *gastrodynia*$_{EN}$. Enfin, on voit comment les relations d'hyponymie peuvent être exploitées de manière similaire, pour concevoir des classes lexicales translinguistiques.

Références

BOURIGAULT, D., JACQUEMIN, C., et al. 2001. *Recent Advances in Computational Terminologies*. Amsterdam/Philadelphia: John Benjamins.

CHIAO, Y-C., ZWEIGENBAUM, P. 2003. The effect of a general lexicon in corpus-based identification of French-English medical word translations. *Actes MIE*, Amsterdam:.

CORBIN, D. 1987. *Morphologie dérivationnelle et structuration du lexique.* Lille: PUL.

[16] Merci au relecteur anonyme pour cette suggestion.

DAILLE, B. 2001. L'identification en corpus d'adjectifs relationnels: une piste linguistique pour l'extraction automatique de terminologie. In *T.A.L*, 42/3, Paris, Hermès:815-832.

DARMONI, S. J., JARROUSSE, E., et al. 2003. VumeF: Extending the French part of the UMLS. *Proceedings of the AMIA Symposium*, Washington, DC:824

FRADIN, B. 2000. Combining forms, blends and related phenomena. In *Extragrammatical and Marginal Morphology*, München: Lincom Europa: 11-59

GAUSSIER, E. 2001. General Considerations on Bilingual Terminology Extraction. In *Recent Advances in Computational Terminology*. Amsterdam/Philadelphia: John Benjamins: 167-184

GRABAR, N., ZWEIGENBAUM, P. 2000. A general method for sifting linguistic knowledge from structured terminologies. *Journal of AMIA* 7(suppl):310-314.

HAHN, U., HONECK, M., et al. 2001. Subword segmentation: Leveling out morphological variations for medical document retrieval. *Journal of AMIA* 8(suppl):229-233.

HASPELMATH, M. 2002. *Understanding Morphology*. London: Arnold.

HATHOUT, N. 2003. L'analogie, un moyen de croiser les contraintes et les paradigmes. Acquisition de connaissances à partir de dictionnaires de synonymes, *RIA*. 17(5-6),:923-934.

IACOBINI, C. 2003. Composizione con elementi neoclassici. In *La formazione delle parole in italiano*, Tübingen: Niemeyer: 69-96

JACQUEMIN, C., TZOUKERMANN, E. 1999. NLP for term variant extraction: A synergy of morphology, lexicon, and syntax. In *NLP and Information Retrieval*, Boston, Kluwer: 25-74.

LOVIS, C., MICHEL, P.A., et al. 1995. Word segmentation processing: a way to exponentially extend medical dictionaries. *8th World Congress on Medical Informatics:* 28-32.

LOVIS, C., BAUD, R., et al. 1998. Medical dictionaries for patient encoding systems: a methodology. *Artificial Intelligence in Medecine* 14:201-214.

NAMER, F. 2003. Automatiser l'analyse morpho-sémantique non affixale: le système DériF. In *Cahiers de Grammaire*. Toulouse: ERSS: 31-48.

NAMER, F., ZWEIGENBAUM P., 2004 Acquiring meaning for French Medical Terminology: contribution of Morphosemantics. in *11th MEDINFO*. 2004. San Francisco, CA:535-539.

SCHULZ, S., ROMACKER, M., et al. 1999. Towards a multilingual morpheme thesaurus for medical free-text retrieval. *Proceedings of MIE*, Ljubliana, Slovenia: 891-894.

TRAN, T-D., BURGUN, A., et al. 2003. Acquisition semi-automatique de terminologie bilingue en biologie moléculaire à partir des corpus comparables. *TIA*, Strasbourg: 166-175

WARREN, B. 1990. The importance of combining forms. In *Contemporary Morphology*, Berlin, New York: Mouton - Walter de Gruyter: 111-132.

ZWEIGENBAUM, P., BAUD, R., et al.. 2003. Towards a unified medical lexicon for French. In *Actes MIE*, Amsterdam: IOS Press: 415-420.

Extraction semi-supervisée de couples d'antonymes grâce à leur morphologie

Didier Schwab, Mathieu Lafourcade et Violaine Prince
LIRMM
Laboratoire d'informatique, de Robotique
et de Microélectronique de Montpellier
MONTPELLIER - FRANCE.
{schwab,lafourca,prince}@lirmm.fr
http ://www.lirmm.fr/ ~{schwab, lafourca, prince}

Mots-clefs : antonymie, morphologie, antonymie complémentaire, antonymie scalaire, antonymie duale, répartition statistique

Keywords: antonymy, morphology, complementar antonymy, scalar antonymy, dual antonymy, statistic distribution

Résumé Dans le cadre de la recherche sur la représentation du sens en Traitement Automatique des Langues Naturelles, nous nous concentrons sur la construction d'un système capable d'acquérir le sens des mots, et les relations entre ces sens, à partir de dictionnaires à usage humain, du Web ou d'autres ressources lexicales. Pour l'antonymie, il n'existe pas de listes séparant les antonymies complémentaire, scalaire et duale. Nous présentons dans cet article une approche semi-supervisée permettant de construire ces listes. Notre méthode est basée sur les oppositions de nature morphologique qui peuvent exister entre les items lexicaux. À partir d'un premier ensemble de couples antonymes, elle permet non seulement de construire ces listes mais aussi de trouver des oppositions morphologiques. Nous étudions les résultats obtenus par cette méthode. En particulier, nous présentons les oppositions de préfixes ainsi découvertes et leur validité sur le corpus puis nous discutons de la répartition des types d'antonymie en fonction des couples opposés de préfixes.

Abstract In the framework of meaning representation in Natural Language Processing, we focus on enabling a system to autonomously learn word meanings and semantic relations from user dictionaries, web contents and other lexical resources. For antonymy, as a lexical semantic relation, no resource provides distinctions between complementary, scalar and dual antonymies. In this paper, we present a semi-supervized method to collate such lists, based on operating morphological opposition holding between lexical items. The approach presented here starts from a bootstraped initial list. It is able to augment such lists but also to find out morphological oppositions. We scrutinize the obtained results and discuss the distribution of antonymy types.

1 Introduction

Dans le cadre de la recherche sur la représentation du sens en Traitement Automatique des Langues Naturelles, nous nous concentrons sur la construction d'un système capable d'acquérir le sens des mots, et les relations entre ces sens, à partir de dictionnaires à usage humain, du Web ou d'autres ressources lexicales. Pour le Français, qui constitue la langue de référence de notre expérimentation, des bases de données gratuites regroupant les principales fonctions lexicales n'existent pas ou ne correspondent pas exactement à nos attentes. En effet, les dictionnaires d'antonymes, regroupent indifféremment les trois types d'antonymie connus : *antonymie complémentaire*, *antonymie scalaire* et *antonymie duale*. Ces trois types seront expliqués dans le prochain chapitre. Afin de gagner en précision, nous cherchons donc à construire des listes d'antonymes basées sur cette typologie.

Nous présentons dans cet article une approche semi-supervisée permettant de construire ces listes. Notre méthode est basée sur les oppositions de nature morphologique qui peuvent exister entre les items lexicaux. Ainsi, les préfixes *ante-* et *post-* s'opposent sur une idée de durée. Cette méthode, à partir d'un premier ensemble de couples connus comme antonymes, extrait les préfixes susceptibles de s'opposer puis, cherche, dans un corpus constitué par les entrées de notre base lexicale, d'autres termes susceptibles d'être antonymes. Un expert valide les couples ainsi extraits et, par là même, les couples de préfixes qui leur correspondent. Une recherche automatique de nouveaux préfixes opposés est effectuée dans le corpus à partir de ces nouveaux couples. La méthode est itérée jusqu'au moment où il n'y a plus de préfixes candidats.

Nous étudions les résultats obtenus par cette méthode. En particulier, nous présentons les oppositions de préfixes ainsi découvertes et leur validité dans le corpus puis nous discutons de la répartition des types d'antonymie en fonction des couples opposés de préfixes.

2 L'antonymie

On considère que « *deux items lexicaux sont en relation d'antonymie si on peut exhiber une symétrie de leurs traits sémantiques par rapport à un axe.* » (Schwab *et al.*, 2002). La symétrie se décline alors de différentes manières, selon la nature de son support. Il se dégage ainsi trois types d'antonymie : l'antonymie *complémentaire*, l'antonymie *scalaire* et l'antonymie *duale*, cette dernière regroupant les *conversifs* et les *duals propres*. Nous notons l'antonymie par un signe d'équivalence ayant subit une rotation de 90 degrés (‹*riche*›III_c‹*pauvre*›). Ce signe rappelle à la fois le signe marquant la synonymie, relation considérée comme opposée à l'antonymie, chez Polguère (Polguère, 2003) et la symétrie axiale existant entre les deux termes antonymes.

2.1 Antonymie complémentaire

Cette antonymie concerne les couples tels que ‹*absent*›III_c‹*présent*›, ou ‹*existence*›III_c‹*inexistence*›.

il est présent ⇒ il n'est pas absent	il n'est pas absent ⇒ il est présent
il est absent ⇒ il n'est pas présent	il n'est pas présent ⇒ il est absent

En termes de logique, nous avons :

$\forall x \quad P(x) \Rightarrow \neg Q(x)$	$\forall x \quad \neg P(x) \Rightarrow Q(x)$
$\forall x \quad Q(x) \Rightarrow \neg P(x)$	$\forall x \quad \neg Q(x) \Rightarrow P(x)$

Nous reconnaissons ici une relation de disjonction exclusive. Dans ce cadre, l'affirmation d'un des termes implique nécessairement la négation de l'autre. Sur le plan de la symétrie, l'antonymie complémentaire présente deux types de symétrie : (1) une symétrie de valeur dans un système à deux valeurs seulement, comme dans l'exemple précédent, (2) une symétrie par rapport à l'application d'une propriété : le *noir* est l'absence de couleur, il est donc "opposé" à toute couleur, et à toute combinaison de couleurs.

2.2 Antonymie scalaire

Les antonymes scalaires (ou gradables) concernent les systèmes échelonnés comme la taille (‹*grand*›III$_s$‹*petit*›) ou la température (‹*chaud*›III$_s$‹*froid*›). La symétrie se réalise par rapport à une valeur de référence du système qui n'est pas toujours représentée par un mot. Par exemple, pour ‹*grand*›III$_s$‹*petit*›, nous avons :

Cet homme est grand	⇒ Cet homme n'est pas petit
Cet homme est petit	⇒ Cet homme n'est pas grand
Cet homme n'est pas grand	⇒ Cet homme est petit ∨ cet homme est de taille moyenne
Cet homme n'est pas petit	⇒ Cet homme est grand ∨ cet homme est de taille moyenne

Cet homme est « *ni grand ni petit* » qui désigne en général la taille moyenne, mais qui ne signifie pas dans le cas présent (comme dans le cas de ‹*vivant*›III‹*mort*›) que la propriété ne s'applique pas. C'est simplement qu'il existe ici une "valeur neutre" à partir de laquelle les autres s'échelonnent. En logique classique, on pourrait l'exprimer, si R est la propriété ayant la valeur de référence (neutre ou médiane), par

$$\forall x P(x) \Rightarrow \neg Q(x) \wedge R(x)$$

$$\forall x Q(x) \Rightarrow \neg P(x) \vee R(x) \qquad \forall x \neg Q(x) \not\Rightarrow P(x)$$
$$\forall x, R(x) \Rightarrow \neg Q(x) \wedge \neg P(x) \qquad \forall x \neg P(x) \not\Rightarrow Q(x)$$

La valeur de référence peut ne pas être la seule valeur possible, mais un des éléments remarquables de l'échelle (pour des propriétés multi-valuées par exemple). L'usage de termes gradables implique toujours une évaluation et donc une comparaison. Celle-ci peut être explicite : « *Jean est plus petit/grand que Pierre* », « *il avance/recule* » (le terme moyen étant *immobile*[1]). Elle peut aussi être implicite et renvoyer à des normes tacitement admises par l'individu ou la communauté à laquelle il appartient : « *il fait chaud* » dit par un habitant d'un pays équatorial ne se réfèrera pas à la même idée de chaleur (donc à la même valeur de référence) qu'un habitant des fjörds de Norvège.

2.3 Antonymie duale

Les antonymes *duals* sont composés de deux sous-familles : les antonymes conversifs et les duals propres. Ils correspondent au troisième type de symétrie, celui que l'usage et la nature même des objets peut introduire.

2.3.1 Conversifs

Les conversifs (appelés aussi réciproques) sont des couples tels que ‹*mari*›III$_d$‹*femme*›, ‹*acheter*›III$_d$‹*vendre*›, ‹*prêter*›III$_d$‹*emprunter*›, ‹*avant*›III$_d$‹*après*›, ‹*père*›III$_d$‹*fils*›. De nombreux linguistes comme Igor Mel'čhuk (Mel'čuk *et al.*, 1995) ne les considèrent pas comme des antonymes. La fonction *anti* de son DEC désigne en réalité les antonymes complémentaires et les antonymes scalaires. Il dédie aux conversifs une autre fonction lexicale *conv*. Cependant, dans la mesure où pour nous, la modélisation de l'antonymie correspond à une étude compléte des mécanismes de symétrie, nous avons considéré les conversifs comme un cas particulier de symétrie, et les avons naturellement associés à un processus antonymique "étendu".

Pierre est le père de Marc ↔ Marc est le fils de Pierre.

Ce qui s'exprime, en terme de logique, par :

$$\forall x, y P(x,y) \leftrightarrow Q(y,x)$$

[1] On peut remarquer que le neutre d'un type d'antonymie peut être opposable dans une autre antonymie. Ici, par exemple, ‹*mobile*›III$_c$‹*immobile*›

Dans le cas des conversifs, si on remplace dans une phrase un terme x par son réciproque y, on peut systématiquement rétablir la synonymie entre les deux phrases à condition de permuter les arguments syntaxiques mis en relation par x comme le montre la formule. Ainsi, pour les conversifs, il y a symétrie par rapport à la place des arguments (P est réciproque de Q).

2.3.2 Duals

Les duals propres sont une notion d'antonymie que nous introduisons pour rendre compte d'un effet particulier de mise en relation de termes où la symétrie porte cette fois-ci sur des fonctions culturelles (symétrie consacrée par l'usage) et spatio-temporelles (propriétés particulières de l'espace-temps). Les duals sont des mots que la culture associe comme ‹*soleil*›III_d‹*lune*›, ou qui ne vont pas, à priori, l'un sans l'autre comme ‹*question*›III_d‹*réponse*› ou alors sont l'expression d'une antonymie temporelle i.e. qui exprime le passage d'un état à un autre comme ‹*naissance*›III_d‹*décès*›. Dans ce troisième cas, on peut remarquer que ces deux évènements marquent le passage entre deux antonymes complémentaires (‹*inexistence*›III_c‹*existence*› dans le cas de ‹*naissance*›III_d‹*décès*› ou bien ‹*présence*›III_c‹*absence*› dans le cas de ‹*départ*›III_d‹*arrivée*›). L'antonymie duale propre présente naturellement une symétrie qui n'est pas relevée dans l'échange des places d'arguments puisqu'il s'agit de prédicats unaires. Elle exprime le fait que si l'un des deux prédicats est vrai, il existe une valeur pour laquelle l'autre l'est aussi nécessairement. Pour la modéliser, on écrira :

$$\exists x_0 \ P(x_0) \leftrightarrow \exists Q, \ Q(x_0)$$

avec Q dual de P qui modélise par exemple le fait que si x_0 a un début, alors il existe aussi une fin à x_0 ou :

$$\exists x_0 \ P(x_0) \leftrightarrow \exists Q, \exists y_0 \ Q(y_0)$$

avec Q dual de P qui exprime que, si x est une question, il existe un objet y et il existe un prédicat réponse, tel que y est une réponse à x.

Cette nécessité du prédicat dual peut rendre compte de certains couples de descripteurs temporels. Ainsi, ‹*avant*›III_d‹*après*› en prédicats unaires, sont linguistiquement différenciés sur le plan de la catégorie grammaticale, comme dans « *il y a un avant et un après* » à ne pas confondre avec ‹*avant*›III_s‹*après*› qui sont des scalaires avec comme valeur médiane ‹*pendant*›.

3 Construction de listes d'antonymes

3.1 Problématique

Dans le cadre de nos recherches sur la représentation du sens en TALN, notre modèle est basé sur un apprentissage effectué à partir de diverses sources comme des dictionnaires à usage humain, le Web, ou des dictionnaires de relations lexicales (Schwab *et al.*, 2004). S'il existe depuis quelques années pour le Français des dictionnaires de synonymes sur le web (celui du CRISCO[2] mais aussi une partie d'EuroWordNet), les dictionnaires d'antonymes étaient eux inexistants. Seul, depuis quelques mois, le dictionnaire du CRISCO fournit des antonymes. Toutefois, il ne considère pas les différents types d'antonymies présentées dans la section 2. C'est pourquoi, dans le but d'améliorer nos représentations nous avons cherché à étudier des méthodes qui nous permettraient de construire le plus automatiquement possible des listes de couple d'antonymes classées suivant leur type. La méthode choisie va utiliser les oppositions de nature morphologique entre les termes.

[2]urlhttp ://elsap1.unicaen.fr/cgi-bin/cherches.cgi

3.2 Morphologie et antonymie

Depuis Charlemagne, l'habitude a été prise de créer des termes à partir de racines latines mais aussi grecques (Walter, 1988). C'est pourquoi en français il n'est pas rare de trouver des termes dits *populaires* c'est-à-dire ayant subi des déformations normales dans une langue (‹*mère*›, ‹*ciel*›) à côté de termes dits *savants* qui eux sont directement construits à partir de morphèmes issus du latin ou du grec (‹*maternel*›, ‹*céleste*›).

Les morphèmes sont les unités minimales significatives qui constituent les mots. Par exemple, le mot "fleurs" est composé de deux morphèmes : le radical (ou base) correspondant à l'item ‹*fleur*› et le suffixe marquant le pluriel *s*. Il existe deux types de morphèmes : (1) les *morphèmes lexicaux* qui correspondent aux items lexicaux ou à une légère variante ; (2) les *morphèmes grammaticaux*, autrement appelés *affixes*. Situés avant le radical, un affixe est dit *préfixe*, après, *suffixe*, dans le radical, *infixe*.

Les morphèmes sont porteurs de sens. Par exemple, le préfixe latin *bi-* correspond à *deux* (‹*binaire*›, ‹*bisexuel*›), le préfixe latin *semi-* et le préfixe grec *hémi-* à une idée de milieu (‹*semi-conducteur*›, ‹*semi-rigide*›, ‹*hémisphère*›) et le préfixe grec *péri-* à ‹*tour*› (‹*périmètre*›, ‹*péricarde*›).

De nombreux mots "*savants*" ont donc été (et sont encore) créés "de toute pièce" en utilisant des préfixes ou des suffixes marquant une idée négative par rapport à la racine ou alors provenant de mots opposés en Latin, en Grec et aujourd'hui en Français. Ce sont ces deux types d'affixes que nous allons utiliser pour construire automatiquement des listes d'antonymes. Ainsi, les préfixes *poly-* (‹*plusieurs*›) et *mono-* (‹*un*›) s'opposent sur le *nombre*, *hyper-* (superlatif) et *hypo-* (au-dessous) s'opposent par rapport à une valeur de *référence* tandis que *non-* ou *més-* marquent la négation. Pour les suffixes, on remarquera l'opposition *-phobe*|||*-phile* (‹*homophobe*›|||‹*homophile*›) ou l'opposition *-dynamique*|||*-statique* (‹*hydrostatique*›|||‹*hydrodynamique*›). Notre article n'étudie ici que les préfixes mais la méthode choisie pour extraire les suffixes est la même.

Il existe des études comme (Béchade, 1992) sur les préfixes du Français et leur signification. Elles permettent de faire une première étude du comportement des préfixes mais ne permettent pas de façon rigoureuse d'opposer deux préfixes et encore moins de chercher à savoir de quel type d'antonymie ils sont marqueurs. Ainsi, une vérification en corpus s'avère impérative. C'est pour cette raison que nous avons mis au point un processus semi-automatique de construction de listes d'antonymes qui permet aussi d'extraire les préfixes opposés.

4 Processus

4.1 Principe

Notre méthode est proche de celle utilisée par (Morin, 1999) pour l'acquisition de schémas lexico-syntaxiques mais elle s'en différencie sur deux points principaux :

- Notre problématique première n'est pas de récupérer des morphèmes opposés mais bien de **construire**, le plus automatiquement possible, des listes d'antonymes. Nous avons tout de même conservé les informations concernant ces préfixes afin de savoir s'ils caractérisent plus particulièrement telle ou telle antonymie. Nous présentons ces résultats dans la section 5.
- Dans (Morin, 1999), les experts valident directement les schémas. Ici, les schémas sont validés indirectement si des couples d'items antonymes, caractérisés par ces schémas, le sont. Cette méthode semble augmenter le nombre de cas à examiner mais il nous paraît difficile d'éliminer des couples de préfixes sans observer leur comportement. Un filtrage automatique basé sur le nombre de couples validés permet de limiter ces cas.

La figure 1 présente le processus de construction de ces listes qui est composé de sept étapes. La méthode utilisant les suffixes est identique.

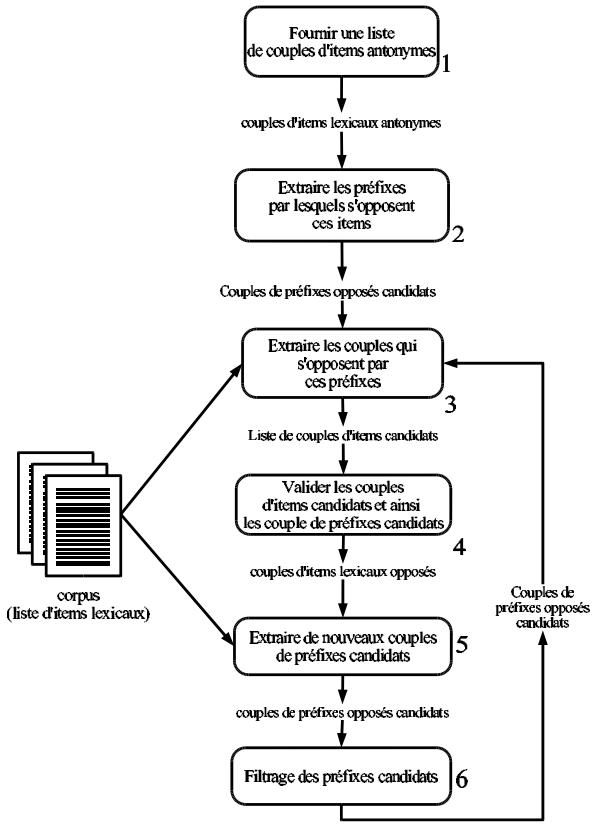


FIG. 1 – Processus d'acquisition de préfixes et de termes antonymes.

1. *Fournir une liste de couples d'items lexicaux antonymes* : cette liste peut être fournie par des dictionnaires ou bien spécifiées manuellement. Elle va permettre d'avoir un noyau de référence pour amorcer le processus.
2. *Extraire les préfixes par lesquel s'opposent ces items* : la méthode consiste à enlever aux mots le plus long suffixe commun aux deux. Ainsi, avec ‹*monosémique*› et ‹*polysémique*›, on enlève le suffixe *sémique*. On obtient ainsi une liste de couples de préfixes candidats.
   Ces deux premières étapes sont facultatives. On peut directement fournir une liste de préfixes censés s'opposer.
3. *Extraire les couples qui s'opposent par ces préfixes* : on extrait du corpus ces couples d'items. On obtient ainsi une liste de couples d'items candidats.
4. *Valider les couples candidats* : on vérifie manuellement les termes candidats et on ne conserve que ceux qui sont effectivement des antonymes. Cette phase est réalisée par un "expert". Si au moins un des couples validés est caractérisé par un des couples de préfixes candidats, ce dernier est validé.
5. *Extraire des préfixes candidats* : parmi la liste des couples retenus, on extrait les suffixes pour trouver dans la base de nouveaux préfixes marquant l'opposition. Le principe consiste à sortir du corpus l'ensemble des termes ayant ce suffixe et d'extraire les préfixes de chacun.
6. *Filtrer les préfixes candidats* : il s'agit de ne retenir que les couples de préfixes qui caractérisent au moins n couples d'items lexicaux dans le corpus.

Le choix de n est important puisque si n est grand, on élimine un grand nombre de préfixes à vérifier manuellement ce qui facilite la tâche de l'expert mais ne garantit pas l'extraction de l'ensemble des oppositions de préfixes. En revanche, le choix d'un n trop petit multiplie le nombre de couples à vérifier et s'avère difficile à concevoir (un choix de $n = 2$ sur la liste des "a" privatifs entraîne la vérification de 3251 couples de préfixes !). Notre choix s'est porté sur un compromis de $n = 5$ qui élimine un nombre suffisant de candidats (Dans notre expérience, nous n'avons plus alors que 110 vérifications à effectuer).

4.2 Déroulement

Considérons le corpus très réduit suivant : ‹*acyclique*›, ‹*anticyclique*›, ‹*anticyclone*›, ‹*antimoine*›, ‹*bisémique*›, ‹*biphonie*›, ‹*cyclique*›, ‹*moine*›, ‹*monophonie*›, ‹*monosémique*›, ‹*polyphonie*›, ‹*polysémique*›, ‹*souris*›. Du fait de la taille du corpus, ici $n = 2$.

1. *étape 1* : On prend une liste de couples que l'on considère comme antonymes :
   ‹*monosémique*›ıı‹*polysémique*›, ‹*acyclique*›ıı‹*cyclique*›
2. *étape 2* : On extrait les préfixes par lesquels ces couples s'opposent :
   *mono-*ıı*poly-*, *a-*ııɛ
3. *étape 3* : On extrait les couples qui s'opposent par ces préfixes :
   pour *mono-*ıı*poly-*, ‹*monosémique*›ıı‹*polysémique*›, ‹*monophonie*›ıı‹*polyphonie*›
   pour *a-*ııɛ, ‹*acyclique*›ıı‹*cyclique*›
4. *étape 4* : L'expert valide les paires extraites :
   ‹*monosémique*›ıı‹*polysémique*› et ‹*acyclique*›ıı‹*cyclique*› sont validées, les deux couples de préfixes sont validés.
   ‹*monophonie*›ıı‹*polyphonie*› est rejetée, cela n'entraîne aucune conséquence sur la validité des préfixes.
5. *étape 5* : On extrait de nouveaux préfixes candidats grâce aux paires validées et au corpus. ‹*monosémique*›ıı‹*polysémique*› possèdent en commun le suffixe *sémique*, on recherche dans le corpus les termes qui ont cette même caractéristique. On extrait ainsi ‹*polysémique*›, ‹*monosémique*› et ‹*bisémique*›. En excluant l'opposition déjà considérée *mono-*ıı*poly-*, ces couples de termes nous permettent comme nouveaux couples de préfixes candidats *mono-*ıı*bi-* et *poly*ıı*bi-*. De même, ‹*acyclique*›ıı‹*cyclique*› permet d'obtenir *anti-*ııɛ
6. *étape 6* : On filtre les couples de préfixes :
   – *mono-*ıı*bi-* apparaît dans deux couples (‹*monosémique*›ıı‹*bisémique*›, ‹*monophonie*›ıı‹*biphonie*›), il est donc conservé.
   – *poly*ıı*bi-*, en revanche, n'apparaît qu'une fois dans le corpus, il est donc supprimé. Dans notre expérience, ces préfixes candidats n'ont pas été rejetés par le filtrage automatique mais par les experts qui ont invalidé les dix couples d'items extraits pour ce couple de préfixes. Il ne semble donc pas y avoir d'exception pour les préfixes *bi-* (marquant une idée de ‹*deux*›) et *poly* (marquant une idée de ‹*plusieurs*›) dont les idées sont incluses l'une dans l'autre.
   – *anti-*ııɛ apparaît lui dans trois couples ‹*cyclone*›ıı‹*anticyclone*›, ‹*cyclique*›ıı‹*anticyclique*› et ‹*moine*›ıı‹*antimoine*›. En pratique la validation ne cherche bien sûr pas tous les couples et s'arrête dès qu'elle en a trouvé deux.

   On réitère l'étape 3
7. *étape 3* : On extrait les couples qui s'opposent par ces préfixes :
   pour *mono-*ıı*bi-*, ‹*monosémique*›ıı‹*bisémique*›, ‹*monophonie*›ıı‹*biphonie*›
   pour *anti-*ııɛ, ‹*cyclone*›ıı‹*anticyclone*›, ‹*cyclique*›ıı‹*anticyclique*› et ‹*moine*›ıı‹*antimoine*›.
8. *étape 4* : L'expert valide les paires extraites :
   ‹*monosémique*›ıı‹*bisémique*›, ‹*monophonie*›ıı‹*biphonie*› ‹*cyclone*›ıı‹*anticyclone*›, et ‹*cyclique*›ıı‹*anticyclique*› sont validées, les couples de préfixes *mono-*ıı*bi-* et *anti-*ııɛ sont validés.
   ‹*moine*›ıı‹*antimoine*› est rejetée, cela n'entraîne aucune conséquence sur la validité des préfixes.

9. *étape 5* : On cherche à extraire de nouveaux préfixes candidats grâce aux paires validées et au corpus.
   Il n'y en a plus, le processus s'arrête.

5 Résultats

Le corpus que nous avons utilisé est constitué de 79 220 items lexicaux issus de notre base lexicale sémantique (Schwab *et al.*, 2004). Ces termes correspondent globalement aux entrées hors noms propres de dictionnaires sous forme électronique : dictionnaires classiques (Larousse, 2004) (Robert, 2000), dictionnaires de synonymes (CRISCO[3]), thésaurus (Larousse, 1992).

Lors de cette expérience, pour de simples raisons pratiques, les auteurs de cette publication ont joué le rôle d'expert validateur. On peut estimer le temps de réalisation de l'expérience à une cinquantaine d'heures utilisées à plus de 99% par la phase de validation (étape 4).

Cette méthode nous a permis d'extraire 49 couples de préfixes opposés. Le tableau de la figure 2 en présente quelques uns.

préfixe 1	préfixe 2	exemple	contre-exemple
a-	ε	‹*chromatique*›∷‹*achromatique*›	‹*afin*›∦‹*fin*›
an-	ε	‹*aérobie*›∷‹*anaérobie*›	‹*anatomiste*›∦‹*atomiste*›
anti-	ε	‹*communiste*›∷‹*anticommuniste*›	‹*moine*›∦‹*antimoine*›
dés-	ε	‹*accord*›∷‹*désaccord*›	‹*avouer*›∦‹*désavouer*›
in-	ε	‹*imaginable*›∷‹*inimaginable*›	‹*incas*›∦‹*cas*›
il-	ε	‹*licite*›∷‹*illicite*›	‹*illustre*›∦‹*lustre*›
pré-	post-	‹*préface*›∷‹*postface*›	ε
hyper-	hypo-	‹*hyperonymie*›∷‹*hyponymie*›	‹*hyperstatique*›∦‹*hypostatique*›
syno-	anto-	‹*synonymie*›∷‹*antonymie*›	ε
méro-	holo-	‹*méronymie*›∷‹*holonymie*›	ε
mono-	poly-	‹*polysémique*›∷‹*monosémique*›	‹*monophonie*›∦‹*polyphonie*›
mono-	stéréo	‹*monophonie*›∷‹*stéréophonie*›	‹*monotype*›∦‹*stéréotype*›

FIG. 2 – Exemples de préfixes antonymes extraits

La figure 3 présente le pourcentage de paires validées par l'expert, c'est-à-dire le nombre de paires considérées comme valides pour chaque couple de préfixes antonymes candidats. On peut constater que si le taux de validation est très important pour la plupart des couples de préfixes il est, en revanche, très faible pour le *a privatif* (*a-* et *an-*).

Finissons par la typologie des paires d'antonymes extraites. La figure 4 présente les résultats obtenus. On constate que globalement la morphologie permet de relativement bien connaître le type d'antonymie. Ainsi, on peut donner quelques indications sur les couple de préfixes suivant leur sémantique :

– *temporels* (*anté-*∷ε, *post-*∷ε, *anté-*∷*post-*, *pré-*∷*post-*) : Tous sont scalaires sauf le couple dual ‹*christ*›∷$_d$‹*antéchrist*›. La définition de (Larousse, 2004) nous donne « *Imposteur qui, suivant l'Apocalypse, doit venir quelque temps avant la fin du monde pour essayer d'établir une religion opposée à celle de Jésus-Christ.* ». Dans ce cas, l'opposition sémantique s'explique

[3]urlhttp ://elsap1.unicaen.fr/cgi-bin/cherches.cgi

préfixe 1	préfixe 2	nombre de paires extraites	nombre de paires validées	nombre de paires invalidées	pourcentage
a-	ε	288	71	217	24,6%
an-	ε	61	15	41	24,6%
anti-	ε	156	147	9	94,2%
dés-	ε	195	179	16	91,7%
in-	ε	616	539	77	87,5%
il-	ε	31	18	13	58%
anté-	ε	15	10	5	80%
post-	ε	37	33	4	89%
anté-	post-	3	2	1	66,6%
pré-	post-	11	11	0	100%
hyper-	hypo-	25	24	1	96%
syno-	anto-	4	4	0	100%
méro-	holo-	2	2	0	100%
mono-	poly-	28	25	3	89,2%
mono-	stéréo-	5	4	1	80%

FIG. 3 – Extraits des résultats d'extraction

par l'opposition ‹*dieu*›III‹*démon*› tandis que la construction de l'opposition morphologique exprime l'idée de l'arrivée du démon avant le retour du christ.

– *médicaux* (*hyper-*III*hypo-*) : Ils caractérisent des mesures qui sont donc au-dessus ou au-dessous de la normale (‹*hyperthyroïdie*›III‹*hypothyroïdie*›).
– *nombre* (*bi-*III*ε*, *tri-*III*ε*, *mono-*III*poly-*, *mono-*III*stéréo-*) : Ils s'opposent par une propriété possédée une fois (*mono*) ou plusieurs (*bi-*, *tri-*, *poly*, *stéréo-*). Ils sont complémentaires.
– *absence de propriété* : *il-*III*ε* (‹*illicite*›III‹*licite*›, ‹*illimité*›III‹*limité*›), *a-*III*ε* (‹*typique*›III‹*atypique*›, ‹*sociabilité*›III‹*asociabilité*›), *an-*III*ε* (‹*anencéphale*›III‹*encéphale*›) ils sont tous complémentaires sauf ‹*anion*›III‹*ion*› qui relève plutôt de l'antonymie duale.
– *opposition culturelle ou produit permettant de lutter contre quelque chose* : *anti-*III (‹*anticlérical*›III‹*clérical*›, ‹*antiviral*›III‹*viral*›). Ils sont duals.

6 Conclusion

Dans cet article, nous avons présenté la méthode semi-supervisée qui nous a permis de construire le plus efficacement possible trois listes de couples d'antonymes, chacune correspondant à un des trois types : complémentaire, scalaire et dual. Cette méthode est basée sur les oppositions de nature morphologique qui peuvent exister entre les items lexicaux et utilise un corpus constitué des entrées de notre base lexicale. À partir d'un premier ensemble de couples antonymes, cette méthode permet à la fois de construire ces listes et de trouver de nouveaux couples de préfixes antonymes. Nous avons ainsi pu étudier la distribution des différents types d'antonymie en fonction du couple de préfixe.

préfixe 1	préfixe 2	complémentaires	scalaires	duals
a-	ε	71 (100 %)		
an-	ε	14 (92,8%)		1 (7,2%)
anti-	ε			147 (100 %)
dés-	ε		3 (1,6 %)	176 (98,3 %)
in-	ε	539 (100 %)		
il-	ε	18 (100%)		
anté-	ε		9 (90%)	1 (10%)
post-	ε		23 (100 %)	
anté-	post-		2 (100%)	
pré-	post-		11 (100%)	
hyper-	hypo-		24 (100 %	
mono-	poly-	25 (100%)		
mono-	stéréo-	4 (100 %)		

FIG. 4 – Répartition des schémas suivant le type

Références

Hervé BÉCHADE. *Phonétique et morphologie du français moderne et contemporain*. Presses Universitaires de France, 1992.

LAROUSSE, *Thésaurus Larousse - des idées aux mots, des mots aux idées*. Larousse, 1992.

LAROUSSE, *Le Petit Larousse Illustré 2004*. Larousse, 2004.

Igor MEL'ČUK, André CLAS, et Alain POLGUÈRE. *Introduction à la lexicologie explicative et combinatoire*. Duculot, 1995.

Emmanuel MORIN. « *Extraction de liens sémantiques entre termes à partir de corpus techniques* ». Thèse de doctorat, Université de Nantes, 1999.

Alain POLGUÈRE. *Lexicologie et sémantique lexicale*. Les Presses de l'Université de Montréal, 2003.

Le ROBERT, *Le Nouveau Petit Robert, dictionnaire alphabétique et analogique de la langue française*. Éditions Le Robert, 2000.

Didier SCHWAB, Mathieu LAFOURCADE, et Violaine PRINCE. « Vers l'apprentissage automatique, pour et par les vecteurs conceptuels, de fonctions lexicales. L'exemple de l'antonymie ». Dans les actes de *TALN 2002*, volume 1, Nancy, Juin 2002.

Didier SCHWAB, Mathieu LAFOURCADE, et Violaine PRINCE. « Hypothèses pour la construction et l'exploitation conjointe d'une base lexicale sémantique basée sur les vecteurs conceptuels ». Dans les actes de *JADT 2004*, Louvain-La-Neuve, Belgique, Mars 2004.

Henriette WALTER. *Le Français dans tous les sens*. Livre de poche, 1988.

Utilisation de corpus de spécialité pour le filtrage de synonymes de la langue générale*

Natalia Grabar[1, 2], Pierre Zweigenbaum[1, 2]
(1) INSERM, U729, 75006 Paris ;
(2) INALCO, CRIM, 75343 Paris Cedex 07 ;
(3) Assistance Publique - Hôpitaux de Paris, STIM/DSI, 75674 Paris Cedex 14
{ngr,pz}@biomath.jussieu.fr

Mots-clefs : Langue de spécialité, langue générale, structuration de terminologies, synonymes, portabilité, filtrage

Keywords: Specialized language, general language, terminology structuring, synonyms, portability, filtering

Résumé Les ressources linguistiques les plus facilement disponibles en TAL ressortissent généralement au registre général d'une langue. Lorsqu'elles doivent être utilisées sur des textes de spécialité il peut être utile de les adapter à ces textes. Cet article est consacré à l'adaptation de ressources synonymiques générales à la langue médicale. L'adaptation est obtenue suite à une série de filtrages sur un corpus du domaine. Les synonymes originaux et les synonymes filtrés sont ensuite utilisés comme une des ressources pour la normalisation de variantes de termes dans une tâche de structuration de terminologie. Leurs apports respectifs sont évalués par rapport à la structure terminologique de référence. Cette évaluation montre que les résultats sont globalement encourageants après les filtrages, pour une tâche comme la structuration de terminologies : une amélioration de la précision contre une légère diminution du rappel.

Abstract General language resources are often more easily available for NLP applications. When using them to process specialized texts it might be useful to adapt them to these texts. This paper describes experiments in adapting general language synonymous resources to the medical domain. A set of filtering methods through a domain corpora is applied. Original and filtered synonyms are then used for normalizing term variation in a terminology structuring task. Their relative contributions are evaluated in comparison with the original structure of the reference terminology. This evaluation shows that the overall results are encouraging, as for the terminology structuring task : improvement of precision while recall is slightly decreased.

*Nos expériences en structuration de terminologies ont été présentées dans (Grabar & Zweigenbaum, 2004). Cet article est plus spécifiquement consacré à l'adaptation de ressources linguistiques générales aux textes de spécialité et à l'influence de cette adaptation sur les résultats.

1 Introduction

Les ressources linguistiques les plus facilement disponibles en TAL ressortissent généralement au registre général d'une langue : lexiques flexionnels, dictionnaires généraux, synonymes, etc. Par définition, la contrainte de spécialisation domaniale est absente de ces ressources. Pour obtenir de meilleurs résultats lors de leur utilisation dans un domaine de spécialité, une adaptation à ce domaine peut être utile. Le but de ce travail consiste ainsi à adapter un ensemble de synonymes généraux au domaine médical grâce aux filtrages effectués sur un corpus médical. Nous utilisons les synonymes originaux et filtrés, à côté d'autres ressources et traitements, pour la normalisation de variantes de termes médicaux. Ces normalisations s'avèrent importantes dans de nombreuses applications (indexation et recherche d'information, questions-réponses, codage dans des terminologies contrôlées, acquisition terminologique, traduction, etc.). Avec des objectifs similaires, (Jacquemin, 1999) applique des règles de transformation morpho-syntaxique pour apparier les termes d'un corpus avec une terminologie contrôlée. Dans le domaine médical, (McCray *et al.*, 1994) exploitent en plus d'autres niveaux de normalisation pour mettre en correspondance des termes provenant de différentes terminologies. En structuration de terminologies, (Hamon *et al.*, 1998) utilisent des synonymes simples pour la détection de liens de synonymie entre termes complexes, grâce à la compositionnalité sémantique. Dans les expériences présentées ici, nous effectuons différents types de normalisation de termes (traitements au niveau de caractères et de l'ordre de mots, ressources morphologiques et synonymiques) : des termes potentiellement proches par leur sens peuvent ainsi être appariés. Nous appliquons ces différentes normalisations en structuration de terminologies, à travers l'hypothèse d'inclusion lexicale qui nous permet d'induire des relations hyperonymiques entre termes.

Dans la section 2, nous présentons les synonymes généraux provenant du Petit Robert ; dans la section 3, les méthodes pour le filtrage de ces synonymes et pour leur évaluation. L'évaluation est faite au sein d'une tâche de structuration de terminologie, où la structure induite est comparée avec la structure originale de ces mêmes termes. Dans la section 4, nous présentons et analysons les résultats. Nous terminons avec une conclusion et des perspectives (sec. 5).

2 Synonymes de la langue générale : Le Robert

L'innovation du dictionnaire Le Robert a été le dépassement de l'organisation alphabétique des lexèmes grâce à l'ajout des rapports analogiques établis sur la base des étymologies, définitions, enchaînements syntaxiques, liens de synonymie et d'antonymie (Robert, 1967). L'ensemble de ces rapports permet aux usagers de saisir plus aisément le sens des lexèmes. Il a fourni également les séries de synonymes que nous utilisons : plus de 140 000 paires de synonymes simples, comme par exemple {*culot*, *fond*}. Elles semblent correspondre à une édition des années 70 de ce dictionnaire[1]. Plusieurs questions peuvent se poser lors de l'utilisation de ressources synonymiques en TAL. Nous discutons ici de leur symétrie, transitivité et spécialisation.

La synonymie, reliant des lexèmes ou expressions contextuellement interchangeables (Cruse, 1986, p. 88), est souvent considérée comme une relation symétrique. Les ressources synonymiques du Robert se présentent sous forme {*entrée dictionnairique*, *famille de synonymes*} ou {*entrée dictionnairique*, *synonyme*}. Nous considérons alors l'*entrée dictionnairique* comme

[1]Nous remercions l'INaLF et Didier Bourigault d'avoir rendu ces ressources disponibles, Thierry Hamon de nous les avoir fournies nettoyées et formatées et Jean-Luc Manguin de nous avoir communiqué leur datation.

canon vers lequel sa famille ou ses synonymes peuvent être « normalisés ». Par contre, nous n'autorisons pas la normalisation dans le sens opposé, ni donc la symétrie. Cela semble raisonnable, comme le montre l'exemple de la famille *boulimie* :

> *boulimie* : *cynorexie*, *hyperorexie*, *hyperphagie*, *sitiomanie*, *faimcalle*, *appétit*, *avidité*,

qui reçoit, si nous considérons la synonymie comme une relation symétrique, *faim*, *fringale* et *frénésie*. De la même manière, *ventre* passe de 11 synonymes à 19 et *trouble* de 22 à 64.

Lors de l'utilisation de synonymes, pour ne pas générer trop de bruit, nous ne calculons pas de fermeture transitive complète. Par contre, nous autorisons une transitivité « locale » : au sein d'une famille, deux synonymes (comme *cynorexie* et *hyperorexie*) peuvent être normalisés vers leur entrée (*boulimie*), et donc considérés eux-mêmes comme synonymes.

Et enfin, la spécialisation de ces ressources. D'une part, les sens spécifiques peuvent manquer, ce qui peut être cause de silence. Par exemple, la famille *culot* :

> *culot* : *fond*, *dépôt*, *résidu*, *benjamin*, *aplomb*, *assurance*, *audace*, *effronterie*, *toupet*, *estomac*.

ne contient pas l'acception médicale (« *Amas d'érythrocytes tassés au fond du récipient de conservation après centrifugation du plasma sanguin* ») proche de *transfusion*. D'autre part, le mélange de registres et donc l'instabilité sémantique des lexèmes, surtout dans un contexte de spécialité, peuvent être cause de bruit. Rappelons que les rapports analogiques du Robert peuvent correspondre en réalité à plusieurs relations, généralement non distinguées (synonymes, analogies, thèmes d'expressions, variantes orthographiques, sous-entrées, hyperonymes, etc.) et mettent ensemble des données fondamentalement hétérogènes (Marcus, 2003). Pour toutes ces raisons, il peut être utile de filtrer les synonymes de la langue générale, surtout lorsqu'ils doivent être utilisés dans des traitements automatiques appliqués à un domaine de spécialité.

3 Méthodes

3.1 Méthode d'adaptation des synonymes généraux

Pour l'adaptation de synonymes généraux aux textes de spécialité, nous utilisons un corpus médical d'environ 8,5 millions d'occurrences. Ce corpus contient des documents hospitaliers (lettres, comptes rendus hospitaliers) et des documents Web collectés à travers le portail médical CISMeF (Darmoni *et al.*, 2001)[2]. Nous supposons ainsi que les synonymes les plus pertinents doivent appartenir au même registre. Les tests utilisés se basent sur le fait que ces synonymes apparaissent dans les mêmes textes et « pas trop loin » (*cf.* Recherche d'associations), et plus spécifiquement qu'il existe des constructions syntaxiques qui sont des indices forts de synonymie (*cf.* Repérage d'indices de synonymie).

Recherche d'associations. La recherche de mots associés (cooccurrences, collocations, etc., voir (Manning & Schütze, 1999)) est une méthode courante en TAL « à base de corpus ». Nous utilisons la mesure du « rapport de vraisemblance » (*log likelihood ratio*) comme dans (Zweigenbaum *et al.*, 2003), dont nous avons repris et adapté les programmes. Nous recherchons des

[2] `http://www.chu-rouen.fr/cismef/`

paires de synonymes qui cooccurrent plus souvent que le hasard dans une fenêtre de 2*150 mots *pleins*. Cela permet de confirmer des paires de synonymes comme :

{*abcès*, *phlegmon*}, {*biopsie*, *ponction*} et {*dernier*, *culot*}, {*signal*, *appel*}.

Avec cette approche de filtrage, une première sélection est faite à travers le corpus : les synonymes doivent apparaître dans la fenêtre de mots fixée. Une deuxième sélection est faite grâce au classement : les associations de synonymes les plus stables ont un meilleur classement.

Repérage d'indices de synonymie en corpus. Les patrons lexico-syntaxiques (Séguéla & Aussenac-Gilles, 1999) et les marqueurs de coordination (Lame, 2002, sec.6.1) sont utilisés par les auteurs pour la structuration de termes. Ils sont projetés sur les corpus et permettent de mettre au jour des relations sémantiques recherchées entre les termes `X` et `Y` :

« `X appelé Y` », « `X est défini comme 1-MOT Y` »,
« `X est confondu avec Y` », « `X n'est autre que 1-MOT Y` »,
« `X ou Y` », « `X ni Y` », « `X et Y` », « `pas de X, pas de Y` ».

La projection de patrons de synonymie et marqueurs de coordination sur notre corpus permet de valider des paires de synonymes comme :

- {*gonflement*, *œdème*} : « *L'œdème* `est défini comme` *un gonflement palpable produit par l'expansion du volume interstitiel liquidien.* »
- {*rhinopharynx*, *cavum*} : « *Le rhinopharynx* `appelé` *cavum est situé sous la base du crâne, en arrière des fosses nasales, au-dessus de l'oropharynx et en avant des 2 premières vertébres cervicales.* »
- {*syndrome*, *affection*} : « *Cette affection* `est appelée` *le syndrome hépato-rénal qui est défini comme une augmentation progressive de la créatinine plasmatique, sans cause évidente autre chez un patient atteint de maladie hépatique avancée.* »
- {*repos*, *sommeil*} : « *Trop souvent, repos* `est confondu avec` *sommeil et activité avec éveil.* »
- {*bruit*, *souffle*} : « *Examen cardiaque : bruits bien frappés aux 4 foyers sans souffle* `ni` *bruit surajoutés.* »
- {*orthopnée*, *dyspnée*} : « *Examen cardio-vasculaire :* `pas` *de dyspnée,* `pas` *d'orthopnée, présence d'œdèmes des membres inférieurs avec un godet positif.* »

3.2 Méthodes d'évaluation du filtrage des synonymes

L'évaluation des synonymes originaux et des synonymes filtrés est faite à travers une tâche de structuration de termes (Grabar & Zweigenbaum, 2004). Les synonymes, à côté d'autres traitements et ressources, sont utilisés pour la normalisation de la variation des termes du thesaurus MeSH (NLM, 2001)[3]. Nous recourons à ces normalisations en effectuant des calculs d'inclusions lexicales pour la détection de relations hiérarchiques. Nous utilisons ensuite la structure originale du MeSH comme référence pour évaluer le rappel et la précision des relations induites.

Détection d'inclusions lexicales. L'inclusion lexicale est naturellement utilisée dans la formation de termes (Kleiber & Tamba, 1990) :

acides gras / *acides gras indispensables*.

Nous exploitons ce fait pour la détection de relations hyperonymiques entre termes. Pour ceci, nous vérifions si tous les mots d'un terme sont inclus dans un autre terme. Si c'est le cas, le terme inclus est supposé être le père hiérarchique P, et le terme incluant, son fils F. Nous effectuons les tests sur les termes segmentés en mots, d'abord sur leurs formes brutes et ensuite avec une série de normalisations :

[3] `http://www.nlm.nih.gov/mesh/meshhome.html`

- normalisation de base : conversion en caractères minuscules, suppression des accents, de la ponctuation, des nombres et des mots « vides » ;
- normalisations avec des ressources morphologiques flexionnelles de la langue médicale et générale, des ressources dérivationnelles et allomorphiques ;
- normalisations avec des synonymes qui sont ceux de la langue médicale, soit ceux de la langue générale, soit les synonymes de la langue générale filtrés sur les corpus médicaux.

Les ressources linguistiques utilisées se présentent sous forme de paires de mots {*canon*,*forme*}. Lors des traitements, si un mot d'un terme correspond à une *forme*, il est normalisé en son *canon* correspondant.

Nous effectuons la mise en minuscules et la suppression d'accents parce que dans la version 2001 du MeSH les termes étaient encore écrits en majuscules non accentuées (*EPITHELIOMA SQUIRRHEUX*), tandis que nos ressources linguistiques sont en minuscules accentuées {*carcinome*, *épithélioma*}. À côté de cela, les termes du MeSH, qui correspondent à un langage d'indexation artificiel, peuvent comporter des virgules (*VIRUS A ADN, INFECTIONS*) ou faire omission des mots grammaticaux (*LESION REPERFUSION MYOCARDIQUE*), ce qui explique que nous appliquons une suppression automatique de la ponctuation, des mots « vides » et que nous ignorons l'ordre des mots. L'abstraction de l'ordre des mots et des mots « vides » est aussi utile lorsque les termes à apparier comportent des dérivations (*sténose de l'aorte* et *aorte sténosée*, *TRANSPLANTATION CARDIAQUE* et *TRANSPLANTATION COEUR-POUMON*). Notons également que nous n'effectuons pas d'analyse syntaxique. L'ensemble de nos normalisations peut ainsi conduire vers des appariements non pertinents (*AGE DENTAIRE / SOINS DENTAIRES SUJET AGE*). On peut supposer qu'une analyse syntaxique des dépendances dans les termes permettrait d'en éliminer un certain nombre.

Évaluation par rapport au référentiel existant. Pour évaluer les relations induites, nous les comparons avec les relations qui existent dans la structure originale du thesaurus MeSH. Nous cherchons ainsi à savoir si les relations hiérarchiques du MeSH pouvent être induites avec l'hypothèse d'inclusion lexicale. Nous calculons alors le rappel et la précision (r_{MeSH} est le nombre de relations dans le MeSH (95 815), r_e le nombre de relations induites existant dans le MeSH, et r_n le nombre de relations induites hors-MeSH) : $R = \frac{r_e}{r_{MeSH}}$; $P = \frac{r_e}{r_e + r_n}$

4 Apport du filtrage des synonymes : résultats et analyse

Nous avons mis en œuvre les méthodes de filtrage présentées en 3.1 aux synonymes du Robert (sec. 2) à l'aide du corpus médical. Les synonymes originaux et filtrés ont été utilisés dans la tâche de structuration de terminologies (sec. 3.2). Nous présentons ici les résultats du filtrage des synonymes et leur impact relatif sur la tâche de structuration : évolution quantitative des relations, utilisation effective des ressources linguistiques, évaluation de ces relations par rapport à la structure originale du MeSH et analyse de quelques relations.

Filtrage des synonymes. Le calcul d'associations permet de valider le nombre le plus élevé de synonymes : 15 589 paires en gardant 60 % des meilleures associations ; les marqueurs de coordination en valident 1 736 et les patrons lexico-syntaxiques 46 paires. Le filtrage complet (union) nous donne un ensemble de 16 154 paires de synonymes, soit une réduction de 88,5 % par rapport aux 140 141 paires d'origine. Il est intéressant de remarquer qu'aucune instanciation

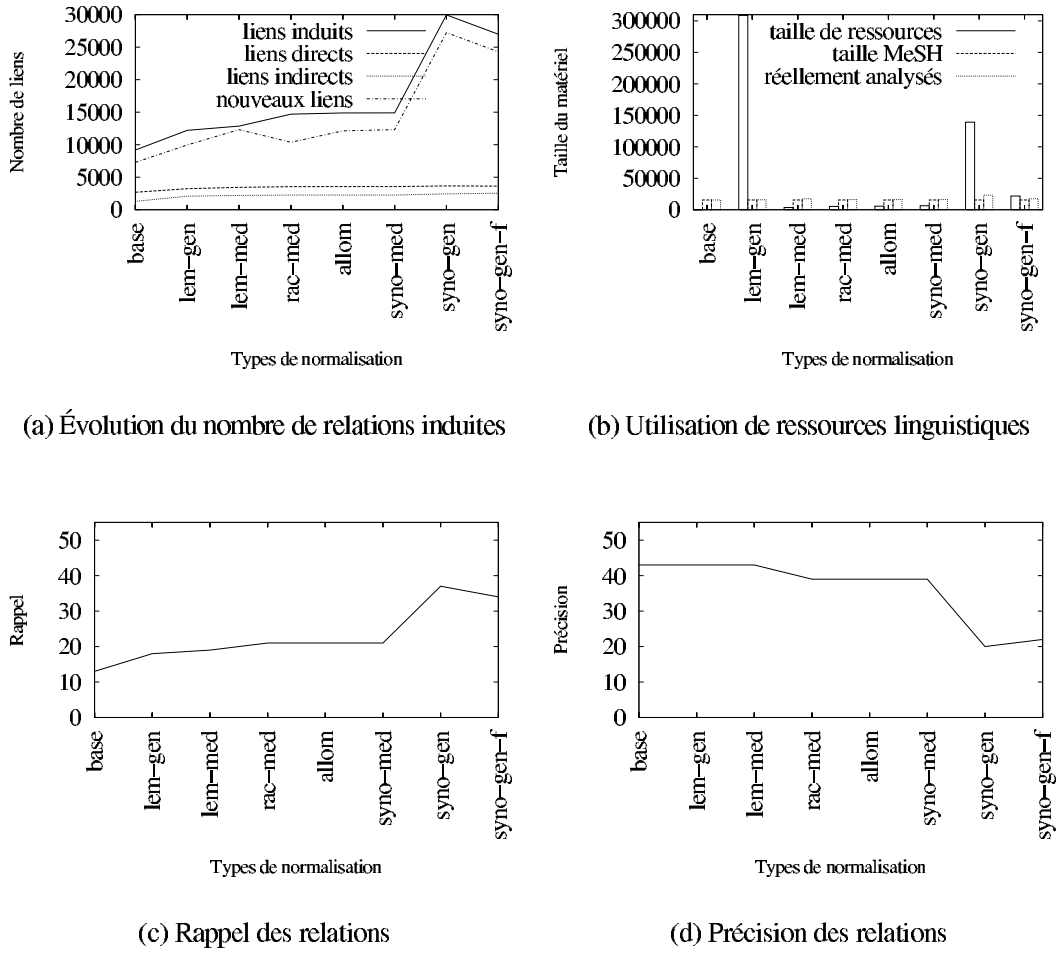
de patrons de (Séguéla & Aussenac-Gilles, 1999) n'a été relevée dans la partie *documents hospitaliers* du corpus. Cela semble raisonnable : ces documents s'adressent à des spécialistes avec, comme but principal, la transmission d'informations sur les patients. Par contre dans les documents du Web, destinés souvent à un public plutôt non averti, les reformulations et le recours à la synonymie sont fréquents. Il apparaît en outre que les marqueurs de coordination relient souvent non des synonymes mais des co-hyponymes ({*bruit*, *souffle*}, {*orthopnée*, *dyspnée*} dans les exemples de la sec. 3.1). Ce caractère des marqueurs déjà noté dans (Pearson, 1998) est accentué ici par la nature hétérogène des relations qui constituent les rapports analogiques dans Le Robert. Notons que l'ensemble de filtrages utilisé exploite les relations syntagmatiques entre les mots. Il serait intéressant d'étudier en plus la piste paradigmatique, par exemple les calculs distributionnels (Nazarenko *et al.*, 2001).

Évolution quantitative des relations. Le calcul d'inclusion lexicale a été appliqué à une liste « à plat » de 19 638 termes du MeSH (écrits à l'époque en majuscules non accentuées). La figure 1(a) montre le nombre de relations induites à chaque étape des normalisations : normalisation des caractères et suppression des mots vides (*base*), application des ressources flexionnelles de la langue générale (*lem-gen*) et de la langue médicale (*lem-med*), des ressources dérivationnelles (*rac-med*) et allomorphiques (*allom*), des synonymes de la langue médicale (*syno-med*), de la langue générale (*syno-gen*) et des synonymes de la langue générale filtrés (*syno-gen-f*). Chaque ressource synonymique est utilisée alternativement, en plus des normalisations antérieures (*base*, *lem-med*, *rac-med* et *allom*). Nous analysons ici les deux dernières étapes : *syno-gen* (synonymes de la langue générale) et *syno-gen-f* (synonymes de la langue générale filtrés). L'analyse de l'impact des synonymes médicaux *syno-med* constitue une perspective. La figure 1(a) décompose le nombre total des relations induites en relations correctes (directes et indirectes du MeSH[4]) et les nouvelles relations (hors-MeSH). Les ressources *syno-gen*, comportant un nombre très élevé de paires de synonymes, doublent le volume des relations induites à l'étape précédente : de 14 884 relations *allom* nous passons à 29 969 avec *syno-gen*, ce qui fait 15 085 relations en plus. Avec les synonymes généraux filtrés le volume est un peu moins important : 26 986 relations (12 102 de plus qu'avec *allom*).

Utilisation effective des ressources linguistiques. La figure 1(b) montre l'utilisation effective de ressources linguistiques. Pour chaque normalisation, nous indiquons le nombre de mots dans la ressource, le nombre de mots dans les termes du MeSH (15 446) et le nombre de mots réellement traités. Nous pouvons ainsi voir que la totalité des ressources est rarement utilisée, surtout avec les ressources de la langue générale (*lem-gen* et *syno-gen*). Cela ne ralentit pas les calculs car seules les ressources pertinentes sont chargées. Par contre, cela montre *a posteriori* que le recouvrement entre les deux ensembles (les données à traiter et les ressources) est faible. Il serait intéressant de pouvoir évaluer ce recouvrement avant les traitements et choisir les ressources qui conviennent le mieux, comme proposent de le faire (Ninova *et al.*, 2005).

Évaluation des relations par rapport à la structure du MeSH. La comparaison des relations induites avec la structure originale du MeSH permet d'évaluer les rappel et précision (fig. 1(c) et 1(d)). Nous pouvons ainsi voir que le rappel augmente de manière générale avec l'injection de connaissances linguistiques supplémentaires. De 13,7 % à l'étape *base* et 21,6 %

[4]Les relations directes existent entre un père et un fils hiérarchiques. Les relations indirectes correspondent à toute relation d'un terme vers l'un de ses ancêtres.



(a) Évolution du nombre de relations induites

(b) Utilisation de ressources linguistiques

(c) Rappel des relations

(d) Précision des relations

FIG. 1 – Illustration des différentes étapes du calcul des inclusions lexicales avec et sans le filtrage des synonymes de la langue générale.

à l'étape *allom* il atteint un sommet avec les synonymes de la langue générale (37,6 %), ces derniers présentant beaucoup de candidats synonymes. Suite aux filtrages, le rappel diminue : il perd 3 % et se fixe à 34,6 %. Face à ces valeurs de rappel très faibles, n'oublions pas que nous testons une seule approche de structuration de termes à travers les relations hyponymiques, basée sur un type d'indices lexicaux (inclusions lexicales). Sa combinaison avec d'autres approches devrait donner une structuration plus complète (Kavanagh, 1995). Comme c'est souvent le cas, l'évolution de la précision est opposée : l'injection de connaissances supplémentaires apporte plus de « risque » dans la génération de relations incorrectes. Par rapport à la configuration de *base* (43,3 %) et d'*allom* (39 %), la précision est de 20,4 % avec *syno-gen* et 22,9 % avec *syno-gen-f*, avec une amélioration de 2,5 % suite aux filtrages. L'apport relatif des ressources synonymiques adaptées ou non au domaine semble alors être similaire à celui des ressources spécialisées ou générales (voir par exemple (Hamon *et al.*, 1998)) : les ressources générales prises dans leur ensemble augmentent le rappel, tandis que leur adaptation augmente la précision. Le choix de filtrer ou non doit donc être fait selon qu'on privilégie le rappel ou la précision. Dans une tâche comme la structuration de terminologies, qui demande une intervention humaine lors de la validation des inductions automatiques, il peut ainsi être plus utile d'avoir une meilleure précision. Une autre piste pour l'amélioration de la précision est liée sans doute à

	Relation hyponymique	Canon	Synonyme(s)
Liens directs	{*adénocarcinome*, *épithélioma squirrheux*}	*carcinome*	*adénocarcinome*, *épithélioma*
	{*céramiques*, *porcelaine dentaire*}	*céramique*	*porcelaine*
	{*famille*, *filiation illégitime*}	*filiation*	*famille*
	{*gravitation*, *modification pesanteur*}	*attraction*	*gravitation*, *pesanteur*
	{*immunisation*, *rappel vaccination*}	*immunité*	*immunisation*, *vaccination*
	{*mouvement*, *activité motrice*}	*vie*	*mouvement*, *activité*
	{*rhinite*, *coryza spasmodique*}	*rhume*	*rhinite*, *coryza*
	{*thérapeutique*, *traitement combiné*}	*thérapeutique*	*traitement*
Liens indirects	{*anesthésie*, *hypnose dentisterie*}	*narcose*	*anesthésie*, *hypnose*
	{*assainissement*, *drainage sanitaire*}	*assèchement*	*assainissement*, *drainage*
	{*calculs*, *lithiase rénale*}	*calcul*	*lithiase*
	{*épithélioma*, *carcinome lobulaire*}	*carcinome*	*épithélioma*
	{*personnalité*, *concept soi*}	*personnalité*	*soi*
	{*thérapeutique*, *traitement par art*}	*thérapeutique*	*traitement*
	{*vascularite*, *artérite temporale*}	*angéite*	*vascularite*, *artérite*

TAB. 1 – Exemples de relations directes et indirectes du MeSH perdues suite aux filtrages.

notre décision d'autoriser la transitivité « locale » des synonymes au sein d'une famille, comme le montre l'exemple {*abcès*, *volume sanguin*} dans la suite de cette section. Ce fait est amplifié d'une part parce que les termes du MeSH ont pour vocation de couvrir tout le domaine médical. La cohésion sémantique de l'ensemble de ces termes est donc moins importante que ce qu'elle pourrait être dans un texte. D'autre part, nous ne vérifions pas l'existence de termes intermédiaires qui servent à l'appariement (*grosseur* et *grosseur sanguine* pour l'exemple ci-dessus). Notons que lors de la détection de relations de synonymie entre termes complexes, (Hamon *et al.*, 1998) bloquent la transitivité « locale ».

De manière générale, nous constatons que ce sont les synonymes généraux qui présentent une rupture avec l'évolution des courbes du rappel et de la précision par rapport aux étapes précédentes. Il serait ainsi intéressant d'analyser de plus près ce que nous obtenons avec l'application de la transitivité « locale » et de compléter ces comparaisons en analysant également l'apport relatif des synonymes généraux (filtrés ou originaux) et des synonymes spécifiques au domaine.

Analyse de relations « filtrées ». Nous examinons ici des relations que nous n'induisons plus suite aux filtrages des synonymes. Elles correspondent aux relations correctes du MeSH et à des relations hors-MeSH, donc potentiellement incorrectes.

Dans le tableau 1, nous présentons des relations correctes du MeSH, directes et indirectes. La première colonne contient ces relations, la deuxième contient le *canon* et la troisième les synonymes normalisés vers le *canon* lors des traitements. Nous induisons 69 relations directes du MeSH et 106 indirectes en moins. Ainsi dans le premier exemple, la famille de *carcinome*, qui avait deux éléments à l'origine (*adénocarcinome*, *épithélioma*), n'en garde plus que le premier. Ce qui empêche d'induire la relation {*adénocarcinome*, *épithélioma squirrheux*}, pourtant correcte, après les filtrages. Dans le deuxième exemple {*céramiques*, *porcelaine dentaire*}, les termes sont appariés à travers une lemmatisation {*céramiques*, *céramique*} et une relation de synonymie {*porcelaine*, *céramique*}.

Nous avons analysé environ 5 % des 3 000 relations hors-MeSH que nous n'induisons plus. Sur cet ensemble, les filtrages semblent montrer pour la plupart leur efficacité. Par exemple, la famille *grosseur*, de 19 éléments à l'origine, parmi lesquels *abcès*, *volume* et *obésité*, permet d'induire les paires comme {*abcès*, *volume sanguin*} et {*obésité*, *volume sanguin*}, manifestement erronées. Cette famille étant réduite lors des filtrages à trois éléments, ces paires ne sont plus générées. De la même manière, les paires comme {*absorption*, *sidération myocarde*}, {*acétylène*, *infusion goudron*}, {*autoritarisme*, *gouvernement États Unis*}, {*crime*, *faute professionnelle*} ou {*émeute*, *lutte contre moustique*} n'apparaissent plus suite à la réduction des familles *anéantissement* (46 éléments à l'origine), *combustible* (27), *gouvernement* (38), *crime* (12) et *révolte* (16), respectivement.

Dans (Grabar & Zweigenbaum, 2004), nous soulignons que des relations hors-MeSH peuvent souvent être pertinentes, par exemple la paire hors-MeSH {*adénocarcinome*, *épithélioma mixte*} perdue suite à la réduction de la famille *carcinome*. Bien qu'étant hors-MeSH, cette paire est probablement correcte car elle suit le même modèle que la paire directe du MeSH {*adénocarcinome*, *épithélioma squirrheux*} (tab. 1). D'autres paires, comme {*agraphie*, *alexie pure*} ou {*agraphie*, *aphasie anomique*} (famille *aphasie* : *agraphie*, *alexie*), sont aussi des relations potentiellement pertinentes, mais non hiérarchiques.

Parmi les relations hors-MeSH analysées ici nous avons beaucoup d'erreurs, mais aussi des paires potentiellement pertinentes, qu'elles correspondent aux relations hyperonymiques ou non. La décision de les encoder dans une terminologie relève d'une décision humaine, que les approches automatiques ne peuvent pas induire.

5 Conclusion et perspectives

Les expériences et analyses présentées avaient pour but de montrer que lorsque des ressources spécifiques à un domaine de spécialité ne sont pas disponibles il est possible de recourir aux ressources de la langue générale. Pour que leur utilisation soit plus bénéfique, il peut être utile de les adapter à ce domaine. Nous avons ainsi testé les ressources synonymiques de la langue générale (Le Petit Robert) sur des données provenant du domaine médical. Ces ressources ont été utilisées en structuration de terminologie à côté d'autres ressources et traitements pour la normalisation des variantes de termes. L'intérêt de l'utilisation des synonymes lors des normalisations est dû au fait qu'ils permettent d'accéder à des variations qui ne sont pas accessibles avec d'autres ressources. L'adaptation est effectuée à travers des filtrages sur des corpus médicaux. Trois approches sont utilisées : recherche d'associations, marqueurs de coordination et patrons lexico-syntaxiques. L'ensemble des filtrages permet de réduire les synonymes originaux d'environ 90 %. La comparaison de l'apport de ressources synonymiques filtrées et originales montre une amélioration de la précision de 2,5 % et une perte du rappel de 3 %, ce qui nous semble bénéfique pour une tâche comme la structuration de terminologie où le bruit est important.

Parmi les perspectives de ce travail, notons essentiellement l'exploration de la piste paradigmatique, par exemple distributionnelle (Nazarenko *et al.*, 2001), pour le filtrage de synonymes ; une analyse plus poussée de l'impact de la transitivité « locale » ; la comparaison de l'apport relatif des synonymes généraux et spécialisés ; et l'application d'une analyse syntaxique des dépendances dans les termes. Par ailleurs, il serait intéressant d'observer l'influence de l'adaptation de synonymes, et de ressources linguistiques en général, dans d'autres contextes applicatifs.

Références

CRUSE D. A. (1986). *Lexical Semantics*. Cambridge : Cambridge University Press.

DARMONI S. J., THIRION B., LEROY J.-P., DOUYÈRE M., LACOSTE B., GODARD G., RIGOLLE I., BRISOU M., VIDEAU S., GOUPY E., PIOT J., QUÉRÉ M., OUAZIR S. & ABDULRAB H. (2001). A search tool based on 'encapsulated' MeSH thesaurus to retrieve quality health resources on the Internet. *MIIM*, **26**(3), 165–178.

GRABAR N. & ZWEIGENBAUM P. (2004). Lexically-based terminology structuring. In *Terminology*, volume 10, p. 23–54.

HAMON T., NAZARENKO A. & GROS C. (1998). A step towards the detection of semantic variants of terms in technical documents. In *International Conference on Computational Linguistics (COLING-ACL'98)*, p. 498–504, Université de Montréal, Montréal, Quebec, Canada.

JACQUEMIN C. (1999). Syntagmatic and paradigmatic representations of term variation. In *37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, p. 341–348, University of Maryland.

KAVANAGH J. (1995). *The Text Analyser : A Tool for Extracting Knowledge From Text*. Master of computer science thesis, University of Ottawa, Ottawa, Canada.

KLEIBER G. & TAMBA I. (1990). L'hyperonymie revisitée : inclusion et hiérarchie. *Langages*, **98**, 7–32. L'hyponymie et l'hyperonymie (dir. Marie-Françoise Mortureux).

LAME G. (2002). *Construction d'ontologies à partir de textes. Une ontologie du droit dédié à la recherche d'information sur le Web*. Thèse de doctorat en informatique temps réel, robotique et automatique, École des Mines de Paris, Paris.

MANNING C. D. & SCHÜTZE H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA : MIT Press.

MARCUS A. (2003). *Dictionnaires électroniques et hypertextualité. Analyse critique des renvois doubles du Grand Robert*. Mémoire de DESS, CRIM/INaLCO. Sous la direction de David Piotrovsky.

MCCRAY A. T., SRINIVASAN S. & BROWNE A. C. (1994). Lexical methods for managing variation in biomedical terminologies. In *Proceedings of the Annual SCAMC*, p. 235–239.

NAZARENKO A., ZWEIGENBAUM P., HABERT B. & BOUAUD J. (2001). Corpus-based extension of a terminological semantic lexicon. In *Recent Advances in Computational Terminology*, p. 327–351. John Benjamins.

NINOVA G., NAZARENKO A. & HAMON T. (2005). Comment mesurer la couverture d'une ressource terminologique pour un corpus ? In *TALN 2005*. À paraître.

NLM (2001). *Medical Subject Headings*. National Library of Medicine, Bethesda, Maryland. `http://www.nlm.nih.gov/mesh/meshhome.html`.

PEARSON J. (1998). *Terms in Context*, volume 1 of *Studies in Corpus Linguistics*. Amsterdam/Philadelphia : John Benjamins.

ROBERT P. (1967). *Préface du Petit Robert*. Paris : Dictionnaires Le Robert. Première édition.

SÉGUÉLA P. & AUSSENAC-GILLES N. (1999). Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine. In *Actes d'Ingénierie des Connaissances (IC)*, p. 79–88, Palaiseau, France.

ZWEIGENBAUM P., HADOUCHE F. & GRABAR N. (2003). Apprentissage de relations morphologiques en corpus. In B. DAILLE, Ed., *Actes de TALN 2003 (Traitement automatique des langues naturelles)*, p. 285–294, Batz-sur-mer : ATALA IRIN.

Combiner analyse superficielle et profonde : bilan et perspectives

Philippe Blache
Laboratoire Parole et Langage
CNRS & Université de Provence
pb@lpl.univ-aix.fr

Mots-clefs : Analyse syntaxique, analyse superficielle, analyse profonde

Keywords: Parsing, shallow and deep parsing

Résumé L'analyse syntaxique reste un problème complexe au point que nombre d'applications n'ont recours qu'à des analyseurs superficiels. Nous faisons dans cet article le point sur les notions d'analyse superficielles et profondes en proposant une première caractérisation de la notion de complexité opérationnelle pour l'analyse syntaxique automatique permettant de distinguer objets et relations plus ou moins difficiles à identifier. Sur cette base, nous proposons un bilan des différentes techniques permettant de caractériser et combiner analyse superficielle et profonde.

Abstract Deep parsing remains a problem for NLP so that many applications has to use shallow parsers. We propose in this paper a presentation of the different characteristics of shallow and deep parsing techniques relying on the notion of operational complexity. We present different approaches combining these techniques and propose a new approach making it possible to use the output of a shallow parser as the input of a deep one.

1 Introduction

Le problème de l'analyse syntaxique reste une question complexe à la fois du point de vue théorique et computationnel. La solution généralement adoptée pour traiter des masses de données volumineuses ou des entrées non standard consiste à recourir à des analyseurs superficiels, robustes et efficaces, mais ne construisant que des informations partielles. Il existe un certain nombre d'études proposant de combiner les techniques d'analyse superficielle profonde permettant soit d'améliorer l'efficacité des analyseurs profonds en leur offrant un meilleur contrôle des processus, soit de proposer une approche permettant de choisir le type d'analyse désiré en fonction des besoins. Cet article dresse un bilan de ces différentes techniques en caractérisant les notions d'analyse profonde et superficielle. Ces caractéristiques sont données non seulement d'un point de vue opérationnel, mais également en introduisant la notion de complexité des phénomènes syntaxiques à analyser. Il s'agit d'une première tentative de classification distinguant les phénomènes faciles à analyser de ceux plus complexes.

On distingue généralement analyse de surface et analyse profonde en fonction de la précision de l'information linguistique construite par un analyseur. Les techniques utilisées sont habituellement différentes : on retrouve plutôt les techniques probabilistes du côté des analyseurs superficiels tandis que les analyseurs profonds utilisent plutôt des approches symboliques. Cette caractérisation doit être complétée par la prise en compte de la finalité de l'application utilisant l'analyseur. Il convient pour cela d'identifier précisément les besoins en termes morpho-syntaxiques ou sémantiques pour identifier le niveau d'analyse requis (chunks pour les systèmes de synthèse de la parole, repérage d'objets nominaux pour les applications de recherche d'information, etc.). Cependant, certaines applications nécessitent, même ponctuellement, des informations plus détaillées concernant les relations syntaxiques ou les effets de sens pour une construction donnée. Nous avons ainsi d'une part une distinction en termes d'efficacité (les analyseurs superficiels sont plus rapides et plus robustes que les analyseurs profonds) et de l'autre une distinction de finalité.

La question du *déterminisme* est à prendre en compte de façon distincte. Si les analyseurs superficiels sont déterministes, les analyseurs profonds traitent quant à eux l'ambiguïté : toutes les possibilités sont prises en compte pendant l'analyse et le système fournit plusieurs solutions lorsque l'ambiguïté ne peut être levée. Une façon de réduire la complexité d'un analyseur profond sans le ramener à un analyseur superficiel consiste à le rendre déterministe. A un premier niveau, l'entrée elle-même peut être déterminisée par l'utilisation d'un étiqueteur désambiguïsant. La déterminisation de l'analyse consiste alors à éliminer des constructions en cours. Les propriétés de coupure utilisées peuvent être de type très différents : probabilistes (par exemple en utilisant des informations syntaxiques associées à des poids), topologiques (propriétés formelles des structures construites, par exemple profondeur des arbres, taille des constituants, etc.), ou encore cognitives (préférences de catégorisation, de rattachement, etc.). Ces techniques permettent de prendre des décisions de façon incrémentale en cours d'analyse. Elles peuvent être associées à des techniques de retardement consistant à repousser certains choix et maintenir plusieurs solutions en parallèle, par exemple en les factorisant. On peut donc à ce stade donner quelques critères distinctifs entre les deux approches :

- analyseur superficiel : rapide et robuste, il fournit une structuration simple en termes d'unités non récursives ainsi que des relations portant sur ces unités
- analyseur profond : fournit une description couvrante des constructions de la langue en indiquant les relations syntaxiques ou syntactico-sémantiques entre ses constituants.

Il existe plusieurs approches permettant de combiner ces approches, la section suivante en propose une présentation. Nous reviendrons ensuite sur une caractérisation de la complexité des phénomènes à analyser avant de décrire, dans la dernière partie, une technique hybride permettant à un analyseur profond de tirer parti d'une analyse superficielle.

2 Approches combinant analyse superficielle et analyse profonde

Il existe un certain nombre de travaux proposant d'utiliser simultanément les techniques d'analyse superficielle et profonde. Un workshop a été récemment consacré à l'étude de ce problème (cf. [Hinrichs04]) et a permis de faire un tour d'horizon de la situation. Dans la plupart des cas, la technique consiste à utiliser l'analyse superficielle en tant que *pré-traitement* d'une analyse

profonde. Par analyse superficielle, on entend surtout ici formatage de l'entrée visant la désambiguïsation de l'*étiquetage* morpho-syntaxique, le traitement des mots inconnus et pouvant aller jusqu'à l'analyse d'unités entières comme les entités nommées par exemple à l'aide de *grammaires locales*. Ce type d'approche peut s'avérer très efficace et offre l'avantage de réutiliser voire d'adapter des composants différents : on trouve par exemple dans [Grover01] une description de la réutilisation d'outils originellement prévus pour l'analyse en GPSG. Dans ce type d'approche, le contrôle de l'analyse profonde se fait donc en limitant l'espace de recherche de l'analyseur grâce à une réduction du nombre d'étiquettes à prendre en compte. De plus, des parties entières peuvent être pré-analysées, ce qui réduit d'autant le nombre de structures à construire. L'intérêt majeur de ce type d'approche réside dans le fait l'analyseur n'a pas à être modifié : il est donc possible de traiter avec le même système une entrée brute ou pré-traitée.

Un second type d'approche, relativement peu répandu, consiste à utiliser les résultats d'un analyseur syntaxique superficiel. L'entrée de l'analyseur profond est la sortie de l'analyseur superficiel, ce qui nécessite l'adaptation de l'analyseur profond. Une première technique consiste à modifier (on emploie également le terme de *lifter*) les informations construites par l'analyseur superficiel. Cela concerne les unités lexicales comme les groupes syntaxiques. Dans le premier cas, il s'agit de transformer une unité simple en une structure enrichie adaptée au format de l'analyseur profond, par exemple par des patterns de filtrage (cf. [Blache95]). Les chunks ou les unités syntaxiques construites peuvent également être transformés à l'aide de règles adaptées utilisant là encore des patterns. Une telle approche est décrite dans [Marimon02] qui transforme ainsi les unités lexicales et les chunks en structures attribut-valeurs du type HPSG comme décrit dans l'exemple suivant :

$$\texttt{rule(}\left[\begin{array}{l}\text{SYNSEM} \mid \text{LOCAL}\left[\begin{array}{l}\text{MORPH}\left[\begin{array}{l}\text{LEMME } \boxed{2}\\ \text{MORPHEME } \boxed{1}\\ \text{AGR } \textit{fem, sing}\end{array}\right]\\ \text{CAT} \mid \text{HEAD}\left[\text{NCLASS } \textit{common}\right]\end{array}\right]\end{array}\right]\texttt{, [Pos='Ncfs-', Lemma=}\boxed{2}\texttt{],}\boxed{1}\texttt{).}$$

On trouve dans la même perspective une technique consistant à enrichir directement la structure construite à l'aide de techniques spécifiques. C'est le cas de [Johnson02]) qui décrit comment, à partir d'arbres syntaxiques simples, créer des arbres complexes à nœud vide. Il s'agit dans ce cas d'une opération d'adjonction qui s'appuie sur des schémas d'arbre spécifiant les endroits où ces nœuds peuvent être insérés et la valeur des arguments qu'ils doivent prendre. Le contrôle du processus se fait grâce à une hiérarchisation de ces schémas. Un troisième type d'approche, défendu dans [Uszkoreit02], propose l'utilisation en parallèle d'une analyse superficielle et d'une analyse profonde. Cette approche (cf. [Crysmann02] ou [Frank03]) consiste à exploiter les informations de l'analyseur superficiel pour contrôler l'analyseur profond et réduire son espace de recherche. Les informations de contrôle fournies par l'analyseur superficiel portent dans cet exemple sur la structure topologique de la phrase en allemand. L'idée est de repérer les champs topologiques par différentes techniques (cf. [Neumann00]) et guider ainsi la construction de la structure par l'analyseur profond. Le dernier type d'approche repose sur la possibilité de régler la finesse de l'analyse en fonction des objectifs. On peut distinguer deux cas selon que les ressources utilisées sont identiques ou pas. Un premier type d'approche consiste simplement à faire varier la grammaire en entrée. L'utilisation d'une grammaire simple, peu ambiguë et n'utilisant que des constituants de forte granularité permettra d'obtenir une analyse grossière d'un énoncé. Dans ce cas, nous pouvons parler de superficialisation d'un analyseur profond par l'utilisation d'une grammaire superficielle (cf. [Puvert04]). Mais il est également

Type	*Caractéristiques*	*Exemple*
Pré-traitement	Etiqueteur désambiguïsant, grammaires locales	[Grover01]
Pré-analyse	Analyse superficielle = input de l'analyseur profond	[Marimon02],[Johnson02]
Contrôle	L'analyseur profond est guidé par l'analyseur superficiel	[Crysmann02], [Frank03]
Granularité variable	Même analyseur, le type de sortie est une option	[Blache02]

Figure 1: Différentes techniques de combinaison d'analyseurs

possible de proposer des techniques permettant d'exploiter des ressources identiques en termes de lexique et de grammaire. Ce type d'approche nécessite la possibilité pour l'analyseur de construire des structures partielles, limitées à un certain type de constituant (par exemple les *SN* dans le cas de systèmes de recherche d'information). De même, ce type de système doit pouvoir construire une segmentation de l'input (par exemple sous la forme de chunks). Mais le même analyseur doit pouvoir à l'autre bout de la chaîne construire également une structure détaillée. Un exemple de ce type d'approche est décrit dans [Blache02]. Il s'appuie sur une représentation décentralisée de l'information sous la forme de contraintes. Le réglage de la granularité d'analyse s'opère en faisant varier la tolérance de l'analyseur par un seuil de contraintes qu'il est possible de relâcher. Le choix de la structure construite en sortie se fait quant à lui en spécifiant le type de contraintes à satisfaire.

3 Les difficultés syntaxiques

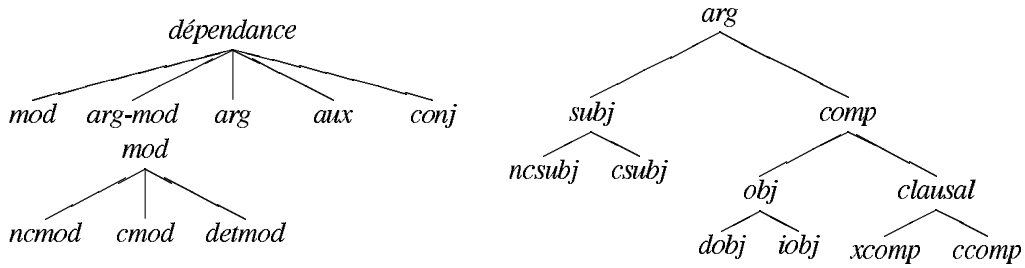
Une analyse précise des difficultés rencontrées par les analyseurs syntaxiques, en dehors des problèmes purement computationnels, reste à établir. Il serait en effet très utile de distinguer les phénomènes faciles à analyser de ceux qui ne le sont pas et d'en expliquer les raisons. Il s'agit d'un exercice difficile, ce problème ne recoupant pas toujours la notion de complexité linguistique : certaines constructions peuvent être facilement interprétables, mais présenter des difficultés en termes d'implantation. C'est le cas par exemple des phénomènes d'extraction qui ne présentent que peu d'ambiguïté d'interprétation mais pour lesquels les systèmes ont des difficultés d'analyse. Réciproquement, les enchâssements de syntagmes peuvent être complexes mais ne présentent pas en soi de difficultés pour un analyseur. Il est intéressant de proposer une ensemble de constructions ou phénomènes qu'un analyseur doit pouvoir traiter. L'article de synthèse [Abeillé00] en fournit une première liste établie de façon tout à fait empirique sur la base d'une analyse des capacités des analyseurs existants au moment de la rédaction de l'article (voici 5 ans, ce domaine a bien entendu beaucoup évolué depuis) :

- dépendances locales: accord, sous-catégorisation des prédicats, expressions semi-figées, restrictions modifieur-modifié, clitiques, etc.
- dépendances moyennes : pronominalisation, contrôle des infinitives, association négative, quantifieurs flottants, etc.
- dépendances à distance : questions, relatives, constructions disloquées, etc.
- alternances syntaxiques : passif, impersonnel, causatives, etc.
- phénomènes de coordination et de comparaison

Cette liste comporte des phénomènes variés et dont la complexité de traitement dépend de la finesse de l'analyse qu'on veut en donner, ainsi que du type de représentation de l'information choisi. Les phénomènes d'accord par exemple sont généralement faciles à traiter pour le français ou l'anglais à condition de disposer dans la grammaire ou le lexique d'un codage

explicite de l'information. Pour ce qui concerne les dépendances à distance, on observe des situations très différentes. Les *relatives* font partie des constructions souvent faciles à traiter, y compris du point de vue de la structure sémantique. Les constructions *disloquées* posent en revanche plus de problèmes. Elles sont assez faciles à repérer, mais la relation sémantique entre l'élément disloqué et le reste de l'énoncé est assez complexe à traiter, même en présence d'un pronom résomptif. Il convient dans ce cas tout d'abord d'identifier ce pronom, celui-ci pouvant apparaître dans des positions très variées, de vérifier les compatibilités morpho-syntaxiques entre l'antécédent et le pronom, mais aussi les compatibilités sémantiques entre l'antécédent et la structure régissant le pronom. Les constructions *clivées* présentent le même type de problème : elles sont en français faciles à identifier, mais le repérage de leur site d'attachement est généralement complexe. Il faut souligner que cette complexité de traitement ne se traduit pas par une difficulté d'interprétation par un humain : les clivées sont au contraire dans la plupart des cas très facile à interpréter (ce type de problème est signalé dans [Puver04]). Pour une même construction, certaines informations sont donc plus complexes à obtenir que d'autres. Si l'on prend en compte le critère de facilité d'analyse pour caractériser un analyseur superficiel, on peut alors dire que l'identification de la présence d'une dépendance à distance peut être obtenue facilement notamment grâce à des marques morphologiques fortes.

Par ailleurs, il faut distinguer d'un côté la structure elle-même (la hiérarchie des objets) et de l'autre les relations existant entre ces objets. Dans le cas d'une approche syntagmatique par exemple, un analyseur devra produire un arbre, mais également indiquer les relations syntaxiques ou sémantiques existant entre les constituants. Un certain nombre de propositions ont été faites pour cela dans le cadre de l'évaluation des analyseurs syntaxiques (cf. [Carroll01], [Briscoe02] ou [Carroll03]). Ce paradigme propose de recenser un certain nombre de relations servant de base à la comparaison et l'évaluation des analyseurs syntaxique. L'ensemble des relations (adapté aux besoins du français par rapport à la proposition de [Carroll01]) est décrit dans la figure 3 et s'organise selon la hiérarchie suivante :



On propose dans le tableau suivant une répartition entre relations faciles et difficiles à identifier. Ce jugement est ici établi sur une base empirique. On essaie de donner quelques arguments justifiant ce classement, mais il conviendrait d'en faire une description plus systématique.

Nom	Description
dépendance	Relation de dépendance générique entre une tête et un dépendant
mod	Relation entre une tête et son modifieur. Le type est le mot introduisant la dépendance
ncmod	Modificateur lexical (non propositionnel)
cmod	Modificateurs propositionnels
detmod	Relation déterminants / noms
arg-mod	Relation tête/argument, celui-ci étant réalisé comme un modifieur (par exemple un SP complément du verbe)
arg	Relation générique tête/argument (plutôt de type complément)
subj	Relation prédicat/sujet
ncsubj	Sujet lexical (non propositionnel)
csubj	Sujets propositionnels (par exemple infinitive sujet)
comp	Relation tête/complément
obj	Relation tête/objet
dobj	Relation prédicat/objet direct (premier complément non propositionnel)
iobj	Relation prédicat/complément non propositionnel introduit par une préposition
clausal	Relation tête/complément propositionnel
xcomp	la proposition complément n'a pas de sujet réalisé
ccomp	la proposition complément a un sujet réalisé

Figure 2: Description des relations

Faciles		**Difficiles**	
Relation	*Caractéristique*	*Relation*	*Caractéristique*
Ncmod	les relations adj/n, sp/n, sp/v, sa/n sont juxtaposées	*n/sp*	séparé par d'autres éléments
Cmod rel	marque morphologique et généralement adjacence	*Cmod*	ambiguïté de rattachement (p. ex. sp/V vs. sp/n)
Detmod	linéarité, adjacence	*Arg-mod*	doit tenir compte de la forme verbale et de la sémantique du mod
Ncsubj	ordre, marque morpho (accord)	*Csubj*	repérage de la proposition, complexité potentielle, identification du verbe tête de la prop sujet, de la tête de la phrase.
Xcomp	marques morphologiques et ordre (eg attribut, infinitif antéposé, etc.)	*Ccomp*	repérage de la subordonnée, identification des têtes verbales
Dobj	forme du complément (nonclausal), ordre (premier) et adjacence avec le verbe Iobj mais ambiguïté rattachement	*Conj*	difficulté de distinguer les conjonctions simples (coordonnés de même type) des autres
Aux	marque morphologique, adjacence	*Contrôle*	Cette relation n'est pas exprimée directement mais par doublement de la relation subj. Dépend du type du verbe et du type du complément

Cette observation rapide permet de dégager quelques éléments de caractérisation de la complexité opérationnelle (et non pas théorique) de l'analyse syntaxique. D'une façon générale en effet, les relations les plus faciles à analyser sont celles profitant d'une conjonction de plusieurs sources d'information stables : faible taux d'ambiguïté des constituants entrant en jeu dans la relation (par exemple seule la relation *Detmod* peut relier un déterminant et un nom), marque morphologique régulière (pronom relatif, conjonctions, construction "*c'est ... que*", etc.), ordre linéaire strict, etc. De plus, le niveau de la relation dans la hiérarchie influe également sur sa complexité : une relation générique sera plus facile à repérer qu'un de ses sous-type (par exemple la relation *Comp* est plus facile à indiquer que *Ccomp*).

A l'inverse, les relations complexes sont celles nécessitant d'accéder à des informations locales spécifiques (par exemple des traits lexicaux), dépendant de la forme des constituants non lexicaux reliés (par exemple type du verbe ou de la préposition dans le *SV* ou le *SP*) ou encore reposant sur des phénomènes sémantiques de restriction. Le niveau sémantique présente

d'ailleurs des caractéristiques similaires : il est par exemple plus facile de traiter le rôle d'un modifieur que la portée de la quantification.

Il n'est donc pas possible de distinguer simplement, comme nous l'avons vu en première partie, un analyseur superficiel d'un analyseur profond sur de simples critères d'efficacité. Mais il ne semble pas non plus pertinent de distinguer les deux approches sur la base du type d'information construit en sortie. Le parenthésage d'un énoncé est une tâche globalement facile si on se contente de constituants non récursifs. Elle devient nettement plus difficile si l'on cherche à décrire le niveau propositionnel ou les dispositifs complexes. De même, comme nous venons de le voir, certaines relations syntaxiques peuvent être plus faciles à identifier que d'autres. Il est donc intéressant d'introduire deux nouveaux critères pour la distinction entre types d'analyse : *niveau opérationnel* (un analyseur profond est non déterministe) et *niveau formel* (un analyseur superficiel ne construit que des informations simples). Les critères de déterminisme et de type d'information peuvent bien entendu être combinés. On pourra par exemple trouver des analyseurs déterministes pouvant construire des informations complexes. Il est possible dans ce cas de parler d'analyseurs intermédiaires.

4 Une stratégie d'analyse hybride

Ainsi que nous venons de le voir, l'analyse profonde a fréquemment recours à des techniques d'analyse superficielle, notamment grâce à une désambiguïsation de l'entrée. Par ailleurs, les résultats obtenus pour la construction d'une analyse superficielle par un analyseur superficiel et un analyseur profond ne sont pas très différents, quelque soit la forme de l'input. La comparaison des résultats obtenus par deux analyseurs sur un même ensemble de corpus dans le cadre de la campagne *Easy* (ces résultats seront présentés lors du workshop *Easy* associé à TALN) montre en effet une forte convergence, aussi bien pour le traitement de corpus de langue écrite que de langue parlée.

Nous avons donc un certain nombre d'arguments qui militent en faveur de systèmes mixtes permettant de fournir comme résultat, en fonction des besoins,aussi bien une analyse superficielle qu'approfondie. Plus précisément, nous proposons une architecture à deux niveaux permettant de réutiliser une analyse superficielle comme entrée d'un analyseur profond. Il ne s'agit pas de modifier la structure superficielle construite (à la différence de l'approche proposée par [Johnson02]), mais bien de construire une représentation plus riche utilisant les objets construits par la superficielle pour construire des objets plus complexes. Il est pour cela nécessaire de définir les objets "superficiels" comme pouvant être des constituants pour l'analyse détaillée. Un parenthésage classique sous forme de chunks ne serait pas pertinente dans cette approche, un chunk ne pouvant être une unité constitutive d'un groupe syntaxique de niveau supérieur.

L'objectif d'une telle approche est tout d'abord de combiner des outils différents, en ne déclenchant éventuellement une analyse détaillée qu'en fonction des besoins. Mais elle permet également d'envisager l'analyse superficielle comme outil de contrôle de l'analyse détaillée. Dans ce cas, toutes les informations construites par l'analyseur superficiel sont susceptibles d'être utilisées par l'analyseur profond. Ces informations sont de deux types : il s'agit d'une part de groupes de mots (donc des informations de parenthésage) et d'autre part des relations entre des formes ou des groupes. Il convient donc de proposer la construction de groupes qui soient à la fois pertinents pour une analyse superficielle, mais également utilisables par un analyseur détaillé. Ces groupes sont nécessairement de premier niveau (i.e. sans constituants emboîtés),

ils ne contiennent que des éléments lexicaux. L'objectif est de définir des groupements très simples et peu ambigus. La grammaire suivante donne une idée du type de groupes pouvant être construits :

```
GV ::= [Adv[neg]] (Clit) [Aux] (Adv) V
GN ::= Det [Adv] [Adj] N[c] | [Det] N[p] | [Det] [Adj] N[p] | Pro[p]
GP ::= Prep Det [Adv] [Adj] N[c] | Prep N[p] | Prep V[ppres] | Prep V[inf]
GA ::= [Adv] Adj | [Adv] V[ppas]
Gadv ::= Adv*
```

Bien entendu, cette grammaire est largement incomplète, et de nombreuses catégories ne sont pas prises en compte, ce qui ne perturbe pas le comportement d'un analyseur superficiel. De même, d'autres règles complétant la description de ce qu'on peut considérer comme étant des syntagmes noyaux peuvent être ajoutées. Enfin, une représentation sous forme syntagmatique ne préjuge pas non plus du formalisme choisi. On peut par exemple décrire cette même information sous forme de dépendances ou de contraintes. Signalons qu'une grammaire de ce type a été utilisée lors de la campagne d'évaluation Easy (cf. [Vilnat04]). De leur côté, les relations pouvant être établies par un analyseur superficiel sont relativement générales et déterminées sur la base d'informations simples, en particulier l'ordre linéaire. Le tableau suivant propose quelques relations avec leur sémantique opérationnelle. Chaque relation est caractérisée par des propriétés qu'il est possible d'extraire de la liste des groupes précédemment construite. Nous utiliserons dans ce qui suit les notations suivantes : GX^+ pour indiquer que le groupe GX fait déjà partie d'une relation et X pour indiquer une suite quelconque d'objets.

Sujet	$(GN \prec X \prec GV) \wedge (\nexists\ GN^+ \in X)$
Aux	$(Aux \prec X \prec V[ppas]) \wedge (\nexists\ V \in X)$
Objet	$(GV \prec X \prec GN) \wedge (\nexists\ GN^+ \in X)$
Conj	$(GX \prec X \prec Conj \prec GX) \wedge (\nexists\ GX \in X)$

Les relations telles qu'elles sont définies ne permettent de spécifier qu'une partie des relations. Par exemple, la relation sujet ne prend pas en compte les inversions, de même que la relation de coordination ne permet que les coordinations simples. Le problème essentiel de ce type de relation est la surgénération. Il est cependant possible d'ajouter un niveau de contrôle spécifique, notamment concernant le type de relation possible pour une catégorie donnée ou encore en ayant recours à des informations lexicales. Une analyse intermédiaire ou détaillée tirant parti d'une analyse superficielle de ce type vise donc la construction de groupes syntaxiques de niveau supérieur ainsi que de relations complexes. Le principe consiste à utiliser en entrée les objets construits par l'analyseur superficiel. Les constituants des unités syntaxiques détaillées sont donc soit des groupes soit des catégories lexicales. Dans les deux cas, l'analyseur superficiel peut associer à ces groupes et catégories des indications en termes de probabilités permettant ainsi de contrôler le processus d'analyse profonde en réduisant l'espace de recherche de l'analyseur. Il est possible de définir les bases d'une analyse intermédiaire. Les règles de la grammaire correspondante utilisent les groupes et les relations construits par l'analyseur superficiel, ce qui permet de préciser ou d'exclure certains constituants en fonction de leur propriétés syntaxiques. Ces relations sont indiquées entre chevrons. Nous obtenons ainsi des règles de la forme :

```
SN ::= GN [GA] [Rel] [GP] <∄ mod(GP,GV)>
SV ::= GV [GN] [GP] <∄ mod(GN,GV)>
Rel ::= Pro[rel] [GN1] GV [GN2] <∄ suj(GN2,GV)>
```

```
SX ← pile_synt[en_cours]
si GC ∈ SX
    ajouter(GC, SX)
finsi sinon
    répéter
        fermer(SX);
        en_cours-;
    tant que ((GC ∉ pile_synt[en_cours]) et (ouvert(pile_synt[en_cours]))
            et (en_cours ≥ 0))
    si en_cours ≠ 0
        ajouter(GC, pile_synt[en_cours])
    finsi
pile_synt[top++] ← GC
```

Figure 3: Algorithme intermédiaire

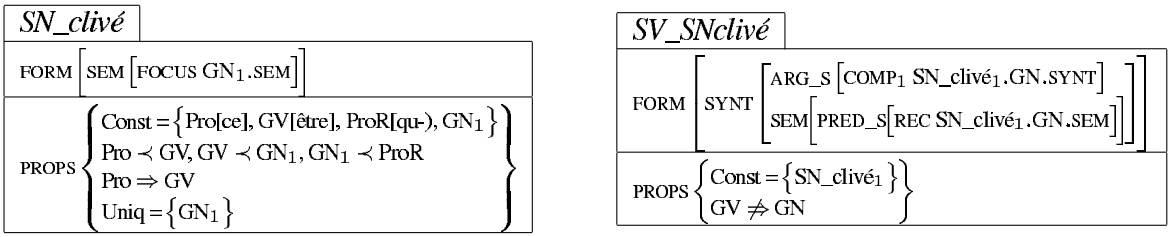


Figure 4: Description du SN clivé en GP

Un algorithme simple (cf. figure 3) consiste pour chaque groupe à vérifier s'il peut appartenir à un syntagme. On utilise pour cela une pile des syntagmes (notée *pile_synt*) construits et deux pointeurs : l'un pointant sur le sommet de la pile (noté *top*), l'autre sur le syntagme en cours (noté *en_cours*). On indique par *GC* le groupe courant, on se dote d'une fonction *ajouter_constituant(Const, SX)* permettant d'ajouter const la liste des constituants de *SX* ainsi que d'une fonction *fermer(SX)* clôturant la liste de constituants de *SX* et d'une fonction booléenne *ouvert(SX)* indiquant si *SX* est ouvert ou fermé.

Il est donc possible d'obtenir à faible coût un analyseur intermédiaire construisant des objets dont les constituants sont des groupes fournis par l'analyse superficielle. La figure (4) présente l'exemple d'une description du clivage du *SN* dans le formalisme des grammaires de propriétés (cf. [Blache05]). Cette analyse s'appuie sur deux constructions : la première décrivant le site de l'extraction, et la seconde les relations avec le site duquel l'élément a été extrait. On y constate, de la même façon que pour l'analyseur intermédiaire, l'utilisation des groupes et des relations en tant que source d'information élémentaire, tout en la complétant avec des informations propres au formalisme choisi.

5 Conclusion

L'analyse syntaxique est un problème dont les facteurs de complexité doivent être précisés. Il est pour cela nécessaire de distinguer précisément les différents types d'analyse (superficielle ou profonde) avant de proposer une caractérisation des phénomènes influant sur cette complexité.

Nous proposons dans cet article une première approche de ce problème qui permet d'envisager une coopération entre ces différentes approches. La technique proposée permet à l'analyse détaillée de s'appuyer sur les résultats de la superficielle, ce qui permet de réduire l'espace de recherche en fournissant en entrée non plus des objets atomiques, mais des informations complexes.

Références

Abeillé A. & P. Blache. (2000). "Grammaires et analyseurs syntaxiques", Traité IC2, Volume Ingénierie des langues, Hermès.

Blache P. & M. Delpui (1995) "Outil d'intégration de bases de connaissances lexicales aux analyseurs syntaxiques", in actes des Journées "Lexicomatique et Dictionnairique".

Blache P., J.-M. Balfourier & T. van Rullen (2002), "From Shallow to Deep Parsing Using Constraint Satisfaction", in proceedings of *COLING-2002*

Blache P. (2005) "Property Grammars: A Fully Constraint-Based Theory", in *Constraint Satisfaction and Language Processing*, H. Christiansen & al. (eds), Springer-Verlag LNAI 3438.

Briscoe, E., J. Carroll, J. Graham & A. Copestake (2002) "Relational evaluation schemes", in proceedings of the *Beyond PARSEVAL Workshop, LREC-02*.

Carroll J. & T. Briscoe (2001) "High Precision Extraction of Grammatical Relations", in proceedings of *IWPT-01*.

Carroll J., G. Minnen & T. Briscoe (2003) "Parser Evaluation. Using a Grammatical Relation Annotation Scheme", in A. Abeillé (ed) *Treebanks: Building and Using Syntactically Annotated Corpora*, Kluwer.

Crysmann B. A. Frank, B. Kiefer, S. Müller, G. Neumann, J. Piskorski, U. Schäfer, M. Siegel, H. Uszkoreit, F. Xu, M. Becker & H. Krieger (2002) "An Integrated Architecture for Shallow and Deep Processing", in proceedings of *ACL-02*.

Frank A., M. Becker, B. Crysmann, B. Kiefer & U. Schäfer (2003) "Integrated Shallow and Deep Parsing: TopP meets HPSG", in proceedings of *ACL-03*.

Grover C. & A. Lascarides (2001) "XML-Based Data Prepartion for Robust Deep Parsing", in proceedings of *ACL/EACL-01*.

Hinrichs E. & K. Simov eds.(2004) Proceedings of the Workshop *"Combining Shallow and Deep Processing for NLP", ESSLLI-04*.

Johnson M. (2002) "A Simple Pattern-Matching Algorithm for Recovering Empty Nodes and their Antecedents", in proceedings of *ACL-02*.

Marimon M. (2002) "Integrating Shallow Linguistic Processing into a Unification-Based Spanish Grammar", in proceedings of *COLING-02*.

Neumann G., C. Braun & J. Piskorski (1999) "A Divide and Conquer Strategy for Shallow Parsing of German Free Texts", in proceedings of *ANLP-00*.

Puver M. & R. Kempson (2004) "Incremental Parsing or Incremental Grammar? ", in proceedings of the workshop *Incremental Parsing: Bringing Engineering and Cognition Together, ACL-04*.

Uszkoreit H. (2002) "New Chances for Deep Linguistic Processing", in proceedings of *COLING-02*.

Vilnat A., L. Monceaux, P. Paroubek, I. Robba, V. Gendner, G. Illouz & M. Jardino (2004) "Annoter en constituants pour évlauer des analyseurs syntaxiques", in actes de *TALN-04*.

Chaînes de traitement syntaxique

Pierre Boullier, Lionel Clément, Benoît Sagot, Éric Villemonte de La Clergerie
INRIA - Projet Atoll
Domaine de Voluceau, Rocquencourt, B.P. 105,78153 Le Chesnay (France)
{Benoit.Sagot,Eric.De_La_Clergerie}@inria.fr
Lionel.Clement@lefff.net

Mots-clefs : Analyse syntaxique, évaluation

Keywords: Parsing, Evaluation

Résumé Cet article expose l'ensemble des outils que nous avons mis en œuvre pour la campagne EASy d'évaluation d'analyse syntaxique. Nous commençons par un aperçu du lexique morphologique et syntaxique utilisé. Puis nous décrivons brièvement les propriétés de notre chaîne de traitement pré-syntaxique qui permet de gérer des corpus tout-venant. Nous présentons alors les deux systèmes d'analyse que nous avons utilisés, un analyseur TAG issu d'une méta-grammaire et un analyseur LFG. Nous comparons ces deux systèmes en indiquant leurs points communs, comme l'utilisation intensive du partage de calcul et des représentations compactes de l'information, mais également leurs différences, au niveau des formalismes, des grammaires et des analyseurs. Nous décrivons ensuite le processus de post-traitement, qui nous a permis d'extraire de nos analyses les informations demandées par la campagne EASy. Nous terminons par une évaluation quantitative de nos architectures.

Abstract This paper presents the set of tools we used for the EASy parsing evaluation campaign. We begin with an overview of the morphologic and syntactic lexicon we used. Then we briefly describe the properties of our pre-syntactic processing that allows us to deal with real-life corpus. Afterwards, we introduce the two parsers we used, namely a TAG parser based on a meta-grammar and an LFG parser. We compare these parsers, showing their common points, e.g., the extensive use of tabulation and compact representation techniques, but also their differences, concerning formalisms, grammars and parsers. We then describe the post-processing that allowed us to extract from our analyses the data required by the EASy campaign. We conclude with a quantitative evaluation of our architectures.

1 Introduction

L'objectif pour les participants de la campagne nationale EASy pour l'Évaluation des Analyseurs Syntaxiques était d'analyser, automatiquement et en moins d'une semaine, environ 35000 phrases. Les analyses devaient être rendues dans le format défini dans le Guide d'annotation (Gendner & Vilnat, 2004). Ce format regroupe une annotation (obligatoire) en constituants et une annotation (faculative) en dépendances syntaxiques, que l'on pouvait rendre sous une forme ambiguë ou désambiguïsée. Bien que nos analyseurs soient non-deterministes, nous avons choisi de fournir à la fois des constituants et des dépendances désambiguïsées.

Les corpus à analyser étaient des corpus réels, non retravaillés, mais segmentés en tokens et en phrases, principalement à des fins d'alignement des résultats des participants. Ils couvraient différents styles, avec environ 6 000 phrases de corpus généraux (journalistiques, législatifs), 8 000 phrases de corpus littéraires, près de 8 000 phrases de corpus de courrier électronique (avec tout le bruit que l'on peut imaginer dans un tel corpus), plus de 2 000 phrases de corpus médicaux, 7 000 phrases de corpus de transcription d'oral (avec les marques spécifiques à de tels corpus, comme les hésitations, les reprises, les répétitions, etc.), et 3 500 phrases de corpus de questions (issus de concours de questions-réponses).

Il nous a donc fallu développer un certain nombre d'outils permettant de transformer ces corpus en entrées acceptables par nos analyseurs. Par ailleurs, nous avons développé un lexique morphologique et syntaxique à large couverture, une méta-grammaire TAG et une grammaire LFG, et des mécanismes permettant de désambiguïser nos analyses et d'en extraire les constituants et dépendances définis par le guide d'annotation. Ces composants ont dû être articulés harmonieusement, construisant ainsi deux chaînes complètes d'analyse syntaxique.

2 Lexique

Le lexique que nous avons utilisé est en cours de développement au sein de l'équipe (Sagot *et al.*, 2005). Il s'agit d'un lexique morphologique et syntaxique à large couverture, dont l'architecture repose sur une structure hiérarchique avec héritage. En effet, le lexique morphologique et syntaxique est construit en deux phases à partir d'informations élémentaires factorisées. La première phase, morphologique, construit un fichier de formes fléchies associées à leur lemme et leur étiquette morphologique à partir d'un fichier de lemmes, d'un fichier décrivant les différentes flexions, et d'un fichier d'exceptions. La seconde phase, syntaxique, construit le lexique final à partir du fichier de formes fléchies, d'un fichier associant les lemmes à des patrons syntaxiques et d'un fichier décrivant ces patrons au sein d'une structure d'héritage.

Le lexique comporte aujourd'hui 404 366 formes fléchies distinctes représentant 600 909 entrées dont certaines sont factorisées. Le développement de ce lexique met en œuvre différentes technique d'acquisition, de complétion et de correction. Outre la récupération de ressources libres de droits, des techniques d'apprentissage automatique de lexiques morphologiques ont été utilisées. Elles ont donné naissance à la première version du *Lefff* (Clément *et al.*, 2004; Clément & Sagot, 2004), qui est un lexique des verbes français présents dans un gros corpus journalistique. Par ailleurs, un des points faibles des lexiques est souvent le manque de couverture pour les multi-mots (tels que *pomme de terre* ou *un peu*). Nous avons donc expérimenté des techniques d'acquisition de multi-mots (cf. (Sagot *et al.*, 2005)).

Notre lexique est encore récent et comporte un certain nombre d'erreurs et de manques. Pour le compléter et le corriger, d'autres techniques ont été employées (cf. (Sagot *et al.*, 2005)). Notre module de correction orthographique permet de détecter automatiquement les mots pour lesquels il n'existe pas de correction à faible coût. Il s'agit le plus souvent de mots manquants à rajouter manuellement. Nous avons également appliqué des méthodes de détection automatique des entrées syntaxiquement incorrectes. L'idée est qu'un mot apparaissant principalement dans des phrases non-analysables a des chances d'être syntaxiquement incomplet ou erroné dans le lexique. Enfin, certaines informations spécifiques (associations verbe-préposition, verbes supports et leurs noms prédicatifs, ...) peuvent être acquises semi-automatiquement moyennant des techniques statistiques simples sur gros corpus. D'autres méthodes sont aujourd'hui envisageables, par exemple des méthodes stochastiques sur des sorties d'analyse syntaxique de corpus avec des grammaires robustes sur-génératrices (cadres de sous-catégorisation très souples, etc.).

3 Traitements pré-syntaxiques

3.1 Description

Nous avons eu à traiter des corpus bruts et donc bruités, bien loin des phrases de linguistes ou des jeux de tests, impliquant le traitement de divers types d'entités nommées[1] (Maynard *et al.*, 2001), des adresses aux « smileys », la correction de fautes d'orthographe, la délimitation des phrases et des mots, et la gestion des particularités de certains corpus oraux ou de transcriptions de sites internet. La segmentation des corpus en phrases et tokens fournie par les organisateurs était parfois soit partielle soit incompatible avec nos outils. Cette segmentation devant être celle des résultats rendus, notre chaîne de traitement pré-syntaxique (décrite plus en détail dans (Sagot & Boullier, 2005)) a été adaptée pour garder en permanence un lien entre une unité morphosyntaxique manipulée par nos outils (unité que nous appelerons *mot*) et le ou les tokens d'entrée (issus de la segmentation fournie) qui lui correspondent. Ainsi, pendant tout le processus, les tokens d'entrée sont conservés dans des *commentaires* (entre accolades et complétés par leur position dans la chaîne d'entrée) qui sont immédiatement suivis du mot associé[2]. Par exemple[3],

contactez-moi␣au␣1␣av.␣Foch,␣75016␣Paris,␣ou␣par␣e-mail␣à␣my.name@my-email.com.

deviendra, si on laisse de côté les ambiguïtés[4]

{contactez$_{0..1}$} contactez {-moi$_{1..2}$} moi {au$_{2..3}$} à {au$_{2..3}$} le {1 av. Foch, 75016 Paris$_{3..9}$} _ADDRESS {,$_{9..10}$} , {ou$_{10..11}$} ou {par$_{11..12}$} par {e-mail$_{12..13}$} e-mail {à$_{13..14}$} à {my.name@my-email.com$_{14..15}$} _EMAIL {.$_{15..16}$} . {.$_{15..16}$} _SENT_BOUND.

[1] Nous utilisons ce terme dans un sens légèrement plus large, en y incluant toutes les séquences de tokens de ce type, y compris celles qui ne sont généralement pas considérées comme des entités nommées (p.ex. les nombres).

[2] Nous utilisons les conventions suivantes : un mot artificiel (par exemple un identifiant d'entité nommée) commence par un « _ » ; dans le corpus, les caractères « _ », « *{* » et « *}* » sont remplacés par les mots artificiels *_UNDERSCORE*, *_O_BRACE* et *_C_BRACE*, qui sont donc des mots du lexique. Ainsi, ces trois caractères sont disponibles comme méta-caractères.

[3] Dans cet article, le symbole « ␣ » représente de manière plus visible un espace, et donc une frontière de tokens ou de mots.

[4] On notera que le même token peut être utilisé plusieurs fois de suite, pour gérer les agglutinées (ainsi *au*$_{2..3}$). Par ailleurs, le token spécial *_SENT_BOUND* indique une frontière de phrase.

Par ailleurs, pour pouvoir prendre en compte certaines ambiguïtés, le résultat de notre chaîne de traitement pré-syntaxique, et donc l'entrée de nos analyseurs n'est pas une séquence de mots mais un treillis (DAG) de mots.

L'architecture de notre chaîne de traitement pré-syntaxique est la suivante :

Grammaires locales sur texte brut : reconnaissance d'un certain nombre d'entités nommées (et autres expressions apparentées) avant la phase de correction orthographique (adresses électroniques, URL, dates, numéros de téléphone, horaires, adresses, nombres en chiffres, smileys, mots entre guillemets, ponctuations et artefacts de transcription de l'oral),

Segmentation en phrases et identification des tokens inconnus : regroupement de deux phrases (au sens de la segmentation EASy) en une seule phrase, ou à l'inverse découpage d'une phrase en plusieurs (nous avons adapté pour cela notre segmenteur, qui étend les idées simples proposées p. ex. par (Grefenstette & Tapanainen, 1994)) ; puis identification des tokens non analysables comme mots du lexique ou combinaison de mots du lexique[5],

Grammaires locales concernant les tokens inconnus : reconnaissance d'entités nommées mettant en jeu des tokens inconnus à l'aide des résultats de la phase précédente : acronymes avec leur expansion, noms propres avec titres, séquences en langues étrangères[6],

Correction orthographique et segmentation : transformation de tout token inconnu (c.-à-d. ne faisant pas partie d'une entité nommée reconnue) en un ou plusieurs mots du lexique par correction orthographique[7], segmentation des tokens et regroupement de tokens adjacents, à l'aide du correcteur orthographique SXSPELL (Sagot & Boullier, 2005),

Grammaires locales sur mots connus : entités nommées composées de mots du lexique (nombres, y compris les ordinaux, et dates écrits en toutes lettres),

Traitement non-déterministe : cette phase, qui produit un treillis de mots du lexique, permet de reconnaître les multi-mots (comme *pomme de terre*) et les agglutinées (comme *au*) tout en préservant toutes les ambiguïtés possibles, mais aussi de représenter différentes alternatives pour gérer les erreurs d'accentuation ou de majuscule initiale[8].

À titre d'illustration, la figure 1 montre la sortie de cette chaîne pour la phrase unique *Jean␣abite␣en␣outre␣au␣1␣,␣rue␣de␣la␣Pompe*, où une espace correspond à une frontière de tokens au sens de la segmentation fournie par EASy. Les notations y sont allégées, et seuls les cas où il n'y a pas correspondance exacte entre un token et un mot sont indiqués : le ou les tokens

[5]Par *combinaison de mots du lexique* nous entendons des tokens tels que *parle-m'en* ou *anti-Bush-né*.

[6]Ces grammaires reposent sur la méthode suivante. Soit $w_1 \dots w_n$ une phrase dont les mots sont les w_i. Nous définissons une fonction d'étiquetage t qui associe (grâce à des expressions régulières) une étiquette $t_i = t(w_i)$ à chaque mot w_i, où les t_i sont pris dans un petit ensemble fini d'étiquettes possibles (respectivement 9 et 12 pour les deux grammaires locales concernées). Ainsi, une séquence d'étiquettes $t_1 \dots t_n$ est associée à $w_1 \dots w_n$. Ensuite, un (gros) ensemble de transducteurs finis transforme $t_1 \dots t_n$ en une nouvelle séquence d'étiquettes $t'_1 \dots t'_n$. Si dans cette dernière la sous-séquence $t'_i \dots t'_j$ correspond à un certain patron, la séquence de mots correspondante $w_i \dots w_j$ est considérée comme reconnue par la grammaire locale.

Soit par exemple l'énoncé *Peu␣après␣,␣le␣Center␣for␣irish␣Studies␣publiait* ..., où *Center*, *irish* et *Studies* ont été identifiés comme mots inconnus. On associe à cet énoncé les étiquettes suivantes : `cnpNEEucn`...(`c` correspond à *initiale en majuscule*, `n` à *probablement français* (cas par défaut), `p` à *ponctuation*, `N` à *connu comme français*, `E` à *connu comme étranger* et `u` à *inconnu*). Ces étiquettes sont transformées en la nouvelle séquence `cnpNeeeen`..., où *e* correspond à *étranger* : *Center␣for␣irish␣Studies* est reconnu comme une séquence en langue étrangère.

[7]Si la correction orthographique est impossible ou trop coûteuse, deux mots du lexique représentant les mots inconnus sont utilisés, l'un correspondant aux mots à initiale majuscule, l'autre à ceux à initiale minuscule.

[8]Nous essayons aussi de corriger les composants de multi-mots qui n'existent pas isolément mais qui ne prennent pas part à leur multi-mot. Par exemple, *brac* n'existe que comme composant du multi-mot *bric␣à␣brac*. Ainsi, *un␣brac* n'a pas été corrigé précédemment, mais est corrigé en *un␣bras*.

sont alors entre accolades, le mot associé étant indiqué derrière. On notera que *Jean*, en tant que premier mot, peut aussi désigner une catégorie de pantalon, que la faute d'orthographe sur *abite* est corrigée, la reconnaissance de l'adresse et le traitement du multi-mot et de l'agglutinée.

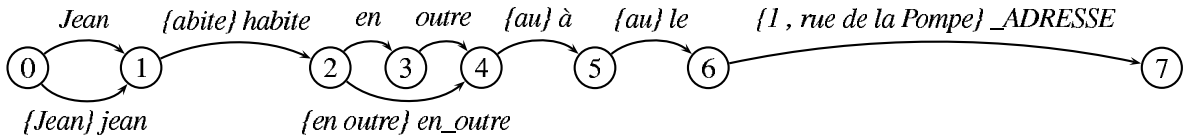


FIG. 1 – DAG associé à *Jean abite en outre au 1 , rue de la Pompe.*

Nos expériences montrent l'importance cruciale pour l'analyse syntaxique d'une telle chaîne de traitement pré-syntaxique, en particulier pour ceux des corpus d'EASy qui sont les plus éloignés du français écrit standard : les corpus de courrier électronique et de transcriptions d'oral.

3.2 Évaluation

L'évaluation d'une telle chaîne est difficile car nous ne disposons pas d'un corpus de référence approprié. Cependant, on peut en avoir un aperçu grâce à des tests préalablement menés sur un corpus journalistique de 1,1 million de mots. Tout le processus prend 13 minutes 01 seconde, soit environ 1400 tokens/sec[9]. Le tableau 1 indique les taux de détection de quelques catégories d'entités nommées manuellement validées.

Classe d'entités nommées	Occurrences	Précision	Rappel
URL	174	100%	100%
adresses (physiques)	35	100%	100%
Expressions en langue étrangère[10]	42	83%	88%

TAB. 1 – Évaluation partielle de la reconnaissance d'entités nommées.

L'évaluation de la segmentation en phrases nécessite une annotation manuelle. Nous l'avons effectuée sur les 400 premières phrases du corpus, ce qui donne un taux de précision de 100% et un taux de rappel de 100%. C'est très satisfaisant, compte tenu du fait que ce corpus journalistique est rempli de citations, de notes de bas de page, de références bibliographiques et de méta-informations qui rendent la détection des frontières de phrases assez difficile.

L'évaluation du correcteur orthographique est délicate. La phase de correction orthographique et de segmentation en mots étant réalisée par un composant qui fait appel au correcteur SXSPELL tout en gérant les phénomènes de segmentation et de majuscules, il y a deux sous-composants à évaluer : le correcteur SXSPELL et le segmenteur-correcteur qui l'utilise. De plus, il faut isoler leurs performances des qualités du lexique et du corpus considérés. Pour ce faire, nous avons identifié automatiquement parmi les 1,1 million de tokens tous ceux qui ne sont pas reconnus par le correcteur-segmenteur comme mots connus ou combinaisons valides de mots connus. Nous avons alors identifié parmi ces tokens inconnus ceux qui devraient être corrigés en des

[9]Le test a été réalisé sur une architecture AMD Athlon ? XP 2100+ (1.7 GHz) et les résultats peuvent paraître lents, comparé, par exemple, aux quelques milliers de mots par seconde que l'on peut obtenir en faisant de l'analyse syntaxique de surface. Mais la phase de correction orthographique est algorithmiquement très coûteuse (impliquant, pour chaque mot, des intersections dynamiques d'automates à plusieurs millions d'états). Les performances que nous obtenons sont donc excellentes.

[10]Test réalisé seulement sur 2000 phrases, car une annotation manuelle est nécessaire.

mots ou combinaisons de mots présents dans le lexique, et nous les avons corrigés manuellement (en tenant compte de leur contexte). Puis nous avons comparé cette correction manuelle à celle fournie par notre système. 91% des 150 tokens concernés sont corrigés (et éventuellement segmentés) correctement. Quelques exemples sont indiqués dans le tableau 2.

Token d'entrée	*arisienne*	*barriére*	*l'intervent␣ionnisme*	*n'aspire-til*	*plrrase*
Correction	*parisienne*	*barrière*	*l'␣interventionnisme*	*n'␣aspire␣-t-il*	*phrase*

TAB. 2 – Exemples de corrections réussies effectuées par le correcteur-segmenteur.

Par ailleurs, 1846 tokens sont analysés comme combinaison de mots du lexique avec (au moins) un préfixe (1712 cas) ou un suffixe (54 cas, seuls *-né*, *-clef* et leurs variantes étant concernés) connu. Ainsi, *quasi-parti␣unique␣chrétien-libéral-conservateur* est transformée en *quasi-_␣parti␣unique␣chrétien-_␣libéral-_␣conservateur*, où « *-_* » est, par convention, la marque des préfixes. Il nous faut préciser à ce stade deux faits. Tout d'abord, le corpus considéré est de très bonne qualité (150 mots du français standard mal orthographiés parmi 1,1 million de mots). D'autre part, cette évaluation du correcteur-segmenteur nous a permis de réaliser l'incomplétude du lexique, en particulier en ce qui concerne les mots d'emprunt à des langues étrangères.

4 Analyseurs syntaxiques

Nous avons développé deux analyseurs utilisant des formalismes, des architectures et des grammaires différents. Le premier, SXLFG, est un analyseur LFG à deux passes. Le second, FRMG, est un analyseur TAG à une passe utilisant une grammaire qui est la représentation compacte d'une TAG avec structures de traits et qui est obtenue par compilation d'une méta-grammaire.

4.1 Analyseur SXLFG

Le système SXLFG (Boullier *et al.*, 2005) permet de construire des analyseurs à partir de grammaires écrites dans une variante du formalisme LFG (Lexical-Functional Grammars). Les grammaires sont donc des grammaires non-contextuelles (CFG) dites *grammaires support* dont les règles dont décorées par des *équations fonctionnelles* dont la résolution repose sur l'unification. Lors d'une analyse, les équations fonctionnelles sont calculées sur une représentation compacte des arbres d'analyse provenant de la grammaire support appelée *forêt partagée*.En cas d'ambiguïté, elle partage les sous-structures communes entre plusieurs analyses.

Pour obtenir un analyseur efficace, nous effectuons les calculs d'équations fonctionnelles directement sur la forêt partagée, et non sur chaque arbre d'analyse CFG. Ceci induit la spécificité de notre variante de LFG : toute information calculée dans les structures fonctionnelles ne peut l'être que de manière *bottom-up*. En effet, puisque l'on effectue ces calculs sur la forêt d'analyse sans la modifier, la structure fonctionnelle associée à la racine d'un sous-arbre ne peut dépendre que des structures associées à ses fils. Dans le cas général, le résultat de ces calculs est un ensemble de structures fonctionnelles associées à la racine de la forêt. Si cet ensemble contient plus d'un élément, on peut par la suite appliquer des heuristiques de désambiguïsation.

Notre analyseur est un analyseur robuste, et ce à plusieurs titres. Tout d'abord, l'analyseur CFG dispose de mécanismes de rattrapage d'erreurs, permettant de traiter les cas où la phrase d'entrée est agrammaticale pour la grammaire support (on parle de phrases *non-valides pour la CFG*

support). Ensuite, en cas d'échec du calcul des équations fonctionnelles, ces équations peuvent être assouplies et donner lieu à des résultats ayant divers degrés d'imperfection. Par exemple, on peut obtenir une structure pour toute la phrase d'entrée mais qui ne respecte pas nécessairement certaines contraintes comme les cadres de sous-catégorisation (on parle d'analyse *sans vérification de cohérence*, par opposition à une analyse qui se déroule correctement jusqu'au bout, dite *avec vérification de cohérence*). En cas d'échec de cet essai, des structures fonctionnelles couvrant des portions disjointes de la phrase sont produites, qui sont appelées *structures partielles*. Au pire, la phrase d'entrée peut être *sur-segmentée*, c'est-à-dire découpée en sous-phrases (avec 5 niveaux de découpage possibles) pour essayer d'en analyser des portions correctes.

Pour la campagne d'évaluation EASy, nous sommes partis d'une grammaire LFG du français développée pour le système XLFG (Clément & Kinyon, 2001), que nous avons modifiée et complétée. Sa couverture et le degré d'ambiguïté de sa grammaire support sont encore améliorables, mais elle traite correctement un nombre respectable de phénomènes syntaxiques complexes.

4.2 Analyseur FRMG

L'analyseur FRMG s'appuie sur une grammaire d'arbres adjoints (TAG) avec décorations engendrée à partir d'un niveau plus abstrait de description, une *méta-grammaire* (MG) (Candito, 1999; Thomasset & de la Clergerie, 2005). La grammaire obtenue est très compacte avec seulement 133 arbres, car elle s'appuie sur des *arbres factorisés* utilisant des disjonctions entre nœuds, des répétitions de nœuds et, surtout, des nœuds optionnels contrôlés par des gardes. L'ancrage des arbres par les entrées lexicales se fait par unification de structures de traits appelées *hypertags*.

Un analyseur syntaxique hybride TAG/TIG[11] a été compilé à partir de la grammaire. Il peut prendre en entrée les treillis produits par la chaîne d'entrée (section 3) modulo quelques conversions pour construire les hypertags. Au démarrage de l'analyse, les arbres sont filtrés par rapport aux mots du treillis d'entrée, pour ne garder que ceux dont les nœuds d'ancrages et les nœuds lexicaux sont compatibles avec ces mots. L'analyseur utilise une stratégie d'analyse tabulaire descendante gauche-droite en une seule passe : le traitement des décorations des nœuds n'est pas repoussé dans une seconde passe, contrairement à la stratégie SXLFG. Néanmoins, les décorations ne sont pas prises en compte pour les prédictions descendantes mais seulement dans les propagations de réponses. Le parcours des arbres factorisés se fait sans expansion de ceux-ci assurant une bonne efficacité. L'analyseur retourne soit une analyse complète du treillis d'entrée, soit, en mode robuste, un ensemble d'analyses partielles couvrant au mieux ce treillis. Les analyses sont émises sous formes de forêts partagées de dérivations TAG indiquant les diverses opérations effectuées (substitution, adjonction, ancrage,...) et ensuite converties en forêts partagées de dépendances (figure 2) servant de base pour les traitements post-syntaxiques.

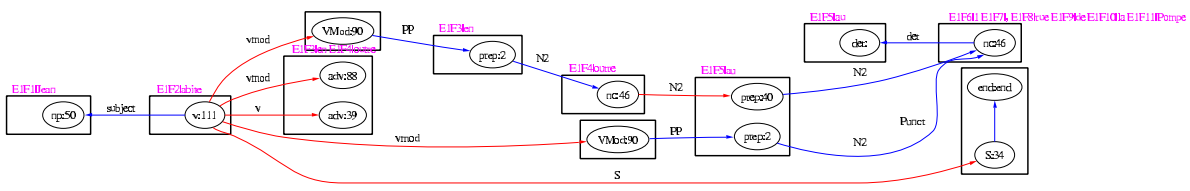


FIG. 2 – Forêt de dépendances (FRMG)

[11] Les TIG (*Tree Insertion Grammars*) sont une variantes des TAG faiblement équivalentes aux CFG.

5 Traitement post-syntaxique

Le format et la nature des informations attendus par les organisateurs de la campagne EASy (Gendner & Vilnat, 2004) ne correspondent pas nécessairement à nos propres formats et choix linguistiques (cf. figure 3). D'autre part, les techniques tabulaires de partage de calculs mises en œuvre dans nos analyseurs sont en partie motivées par le souci d'obtenir l'ensemble des analyses pour une phrase, alors que la piste d'évaluation de base pour EASy concerne des analyses syntaxiques non ambiguës. Il a donc été nécessaire de mettre en place des algorithmes de désambiguïsation et de conversion travaillant sur les structures partagées produites par nos analyseurs. Ces travaux ont été l'occasion d'explorer ce type d'algorithmes avec des approches assez différentes dans les cas de SXLFG et de FRMG. Nous avons également dû explorer diverses règles heuristiques de désambiguïsation et comprendre comment les exprimer.

GN 1	NV 2	GR 3		GP 4						
Jean	abite	en	outre	au	1	,	rue	de	la	Pompe
F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11

sujet	**verbe**
GN1	NV2

complément	**verbe**
GP4	NV2

modifieur	**verbe**
GR3	NV2

FIG. 3 – Sortie EASy fournie par SXLFG et FRMG pour la même phrase que précédemment

Dans le cas de FRMG, la désambiguïsation et la conversion s'appuient sur les forêts partagées de dépendances (section 4.2). Les arcs de dépendance se prêtent bien à l'expression d'heuristiques de désambiguïsation : chaque arc se voit attribuer un poids donné par la somme des poids élémentaires associés aux contraintes satisfaites par l'arc, avec, par exemple, un poids élevé pour une dépendance entre un verbe et un argument et moindre entre un verbe et un modifieur. Au niveau global, l'algorithme retient un ensemble d'arcs maximisant la somme de leurs poids et tels que tout nœud soit accessible par un et un seul chemin. Néanmoins, pour des raisons d'efficacité, l'algorithme a été (tardivement) complété par une notion de coût *régional* associé à un sous-ensemble d'arcs atteignables à partir d'un nœud. Une sélection bornée des meilleurs coûts régionaux est effectuée pour progressivement calculer un coût global qui n'est plus nécessairement optimal. Quoique bien plus efficace, l'algorithme reste encore trop lent dans certains cas. Une analyse plus poussée du problème (en partie aidée par l'approche suivie pour SXLFG) suggère que trop d'informations sont perdues lors de la conversion des dérivations en dépendances[12]. En particulier, le format actuel n'indique pas si deux dépendances issues d'un même mot appartiennent ou non à une même analyse, ce qui nécessite l'ajout de règles coûteuses favorisant les bonnes configurations. Nous prévoyons donc de faire évoluer notre notion de forêt partagée de dépendances. Malgré ces problèmes, nous avons pu constater l'adéquation des arcs de dépendance pour exprimer des règles de désambiguïsation ou de conversion.

Dans le système SXLFG, la phase de désambiguïsation se fait par l'application successive d'un certain nombre de règles sur les structures fonctionnelles associées à la racine de la forêt d'analyse produite par la grammaire support. Chaque règle met en œuvre un critère pour éliminer les structures fonctionnelles non optimales au sens de ce critère. La dernière règle choisit au hasard une analyse parmi celles qui restent. La forêt d'analyse est alors élaguée pour n'y laisser que l'arbre[13] support correspondant à la structure fonctionnelle choisie. L'extraction des constituants

[12] Ceci est dû au fait que nos forêts de dépendances ont initialement été conçues pour une visualisation simplifiée d'un ensemble important d'analyses.

[13] En toute rigueur, plusieurs arbres peuvent subsister s'ils correspondent à une structure fonctionnelle identique.

et des dépendances demandés par EASy se fait alors en parcourant la structure fonctionnelle et son arbre associé, à la recherche de motifs correspondant aux spécifications de la campagne. Cette phase est facilitée par le fait que l'analyse unique issue de la phase de désambiguïsation a été préalablement extraite, à l'inverse de ce qui se passe dans le système FRMG.

6 Mise en œuvre et résultats expérimentaux

Le volume de données à analyser pour EASy, le nombre d'essais que nous voulions effectuer et la complexité de la tâche étaient suffisamment conséquents pour que nous décidions de ventiler les analyses sur plusieurs machines, formant ainsi un cluster pour chaque système.

Les tableaux 3 à 5 présentent divers résultats concernant EASy mais aussi les corpus EUROTRA et TSNLP. Les nombres de phrases diffèrent selon le système, en raison d'heuristiques différentes de segmentation en phrases. Par ailleurs, le *taux d'ambiguïté moyen par mot* n'est disponible que pour FRMG, car dans SXLFG les heuristiques de désambiguïsation sont incorporées dans l'analyseur. Ce taux est défini comme le nombre moyen d'arcs de dépendance atteignant un mot moins un[14].

Corpus	#phrases	% couv.	temps d'analyse				amb.
			moy.	méd.	$\geq$ 1s	$\geq$ 10s	
EUROTRA	334	95.80%	1.81s	1.27s	61.68%	1.55%	0.7
TSNLP	1661	93.38%	0.72s	0.56s	22.03%	0.00%	0.4
EASy	34438	42.45%	5.55s	1.61s	64.41%	9.32%	0.6

TAB. 3 – Résultats pour FRMG, avec un *timeout* de 100 secondes[15]

Corpus	#phrases	couverture (sans vérif. de coh.[16])	couverture (avec vérif. de coh.)	temps d'analyse			
				moy.	méd.	$\geq$ 0.1s	$\geq$ 1s
EUROTRA	334	94.61%	84.43%	0.33s	0.02s	22.2%	6.0%
TSNLP	1661	98.50%	79.12%	0.03s	0.00s	2.8%	0.6%
EASy	40859	66.62%	41.95%	n.d.[17]			

TAB. 4 – Résultats pour SXLFG, avec un *timeout* de 15 secondes[15].

7 Conclusion

La campagne d'évaluation EASy nous a permis de mettre en évidence la différence considérable qu'il y a entre le développement d'un analyseur syntaxique et le développement d'une chaîne complète d'analyse syntaxique. En effet, outre l'importance de la qualité de la grammaire et de l'analyseur, cette campagne a montré le rôle non moins déterminant de la couverture et de la richesse du lexique, de la qualité de la chaîne de traitement, de la précision des méthodes d'exploitation des sorties des analyseurs, ainsi que la très forte interaction entre les différents composants, et en particulier entre le lexique et la grammaire.

[14]Pour une phrase non-ambiguë, chaque mot (sauf la « tête » de la phrase) est atteint par un seul arc, d'où un taux d'ambiguïté nul. Le nombre maximal d'analyses pour un taux α et une phrase de longueur n est en $O((1+\alpha)^n)$.

[16]On notera qu'un *timeout* plus élevé aurait augmenté les taux de couverture mais également les temps d'analyse.

[17]Nous n'avons pas conservé les informations permettant de donner les temps sur le corpus EASy. Toutefois, (Boullier *et al.*, 2005) donne les temps d'analyse pour les 87.51% de phrases reconnues par la CFG support.

		Corpus complet	Phrases valides pour la CFG support	
		Analyse CFG	Analyse CFG	Analyse complète
	#phrases	40859	35756	
	n_{moy} - n_{max}	20.95 - 541	19.06 - 173	
	UW_{moy} - UW_{max}	0.79 - 97	0.75 - 65	
Nombre d'analyses	med - max	32 028 - 3.10^{73}	29 582 - 5.10^{52}	1 - 1
	$\geq 10^{12}$	8.86%	7.84%	0%

TAB. 5 – Données sur les corpus[18] et nombres d'analyses pour SXLFG, avant application de l'heuristique de sur-segmentation.

Cette forte complémentarité entre les différentes phases des chaînes d'analyse syntaxique a inévitablement élargi le champ de la campagne EASy. Ce n'est pas seulement l'analyse syntaxique elle-même qui a été évaluée lors de cette campagne, mais la capacité à mettre en place des chaînes d'analyse syntaxique complètes[19]. Nous comptons exploiter le fait que nous avons déployé deux chaînes de traitement que tout sépare sauf le lexique et la chaîne pré-syntaxique. Ceci nous permettra d'effectuer des comparaisons et d'améliorer ainsi grammaires et analyseurs (en étudiant les différences entre nos résultats), mais aussi le lexique et la chaîne de traitement pré-syntaxique (en étudiant les erreurs communes).

Références

BOULLIER P., SAGOT B. & CLÉMENT L. (2005). Un analyseur LFG efficace : SXLFG. In *Actes de TALN'05*, Dourdan, France.

CANDITO M.-H. (1999). *Organisation modulaire et paramétrable de grammaires électroniques lexicalisées*. PhD thesis, Université Paris 7.

CLÉMENT L. & KINYON A. (2001). XLFG-an LFG parsing scheme for French. In *Proc. of LFG'01*.

CLÉMENT L. & SAGOT B. (2004). Site internet du Lefff (Lexique des Formes Fléchies du Français). `www.lefff.net`.

CLÉMENT L., SAGOT B. & LANG B. (2004). Morphology Based Automatic Acquisition of Large-coverage Lexica. In *Proceedings of LREC'04*, p. 1841–1844.

GENDNER V. & VILNAT A. (2004). Les annotations syntaxiques de référence PEAS. En ligne sur `www.limsi.fr/Recherche/CORVAL/easy/PEAS_reference_annotations_v1.6.html`.

GREFENSTETTE G. & TAPANAINEN P. (1994). What is a word, what is a sentence ? Problems of tokenization. In *Proceedings of the 3rd CCLTR*, Budapest, Hungary.

MAYNARD D., TABLAN V., URSU C., CUNNINGHAM H. & WILKS Y. (2001). Named entity recognition from diverse text types. In *Proceedings of RANLP 2001*, Tzigov Chark, Bulgaria.

SAGOT B. & BOULLIER P. (2005). From raw corpus to word lattices : robust pre-parsing processing. In *Actes de L&TC 2005*, Poznań, Pologne.

SAGOT B., CLÉMENT L., ÉRIC VILLEMONTE DE LA CLERGERIE & BOULLIER P. (2005). Vers un méta-lexique pour le français : architecture, acquisition, utilisation. In *Journée ATALA sur l'interface lexique-grammaire*. `http ://www.atala.org/article.php3?id_article=240`.

THOMASSET F. & DE LA CLERGERIE E. V. (2005). Comment obtenir plus des méta-grammaires. In *Actes de TALN'05*, Dourdan, France.

[18] Pour les données sur les corpus, n désigne un nombre de mots, et UW un nombre de mots inconnus.

[19] En outre, l'harmonisation des résultats des différents participants passe par une segmentation commune en phrases et en mots, différente de celle produite et utilisée par nos outils, qui a dû être conservée en permanence.

TALN 2005, Dourdan, 6-10 juin 2005

Direkt Profil : un système d'évaluation de textes d'élèves de français langue étrangère fondé sur les itinéraires d'acquisition

Jonas Granfeldt (1), Pierre Nugues (2), Emil Persson (1), Lisa Persson (2), Fabian Kostadinov (3), Malin Ågren (1), Suzanne Schlyter (1)

(1) Institut des langues romanes – Université de Lund
Box 201, S-221 00 Lund, Suède
{Jonas.Granfeldt, Malin.Agren, Suzanne.Schlyter}@rom.lu.se,
emil.person@telia.com
(2) Institut d'informatique – Université de Lund
Box 118, S-221 00 Lund, Suède
Pierre.Nugues@cs.lth.se, nossrespasil@hotmail.com
(3) Institut d'informatique – Université de Zurich
fabian.kostadinov@access.unizh.ch

Mots-clés : français langue étrangère, itinéraires d'acquisition, évaluation, annotation, analyse syntaxique partielle

Keywords: second language French, developmental sequences, evaluation, annotation, partial parsing

Résumé : *Direkt Profil* est un analyseur automatique de textes écrits en français comme langue étrangère. Son but est de produire une évaluation du stade de langue des élèves sous la forme d'un profil d'apprenant. *Direkt Profil* réalise une analyse des phrases fondée sur des *itinéraires d'acquisition*, i.e. des phénomènes morphosyntaxiques locaux liés à un développement dans l'apprentissage du français. L'article présente les corpus que nous traitons et d'une façon sommaire les *itinéraires d'acquisition*. Il décrit ensuite l'annotation que nous avons définie, le moteur d'analyse syntaxique et l'interface utilisateur. Nous concluons par les résultats obtenus jusqu'ici : sur le corpus de test, le système obtient un rappel de 83% et une précision de 83%

Abstract: *Direkt Profil* is an automatic analyzer of texts written in French as a second language. The objective is to produce an evaluation of the development stage of the students under the form of a learner profile. *Direkt Profil* carries out a sentence analysis based on developmental sequences, i.e. local morphosyntactic phenomena linked to a development in the learning of French. The paper presents the corpus that we use and briefly, the developmental sequences. Furthermore, it describes the annotation that we have defined, the parser, and the user interface. We conclude by the results obtained so far: on the test corpus the systems obtains a recall of 83% and a precision of 83%.

Jonas Granfeldt, Pierre Nugues, Emil Persson, Lisa Persson, Fabian Kostadinov, Malin Ågren, Suzanne Schlyter

1 Introduction

Les systèmes d'évaluation des compétences linguistiques et d'apprentissage des langues assisté par ordinateur (ALAO) ont peu recours aux techniques de traitement automatique des langues (TAL). Les applications commerciales existantes produisent des exercices dont la correction dépend de techniques de reconnaissance de forme. Ces techniques limitent non seulement la qualité et la nature du feedback, mais elles restreignent aussi les types d'activités possibles. Nous présentons ici un système réalisant une analyse automatique de textes produits librement. Il est fondé sur l'étude de l'acquisition des langues étrangères à l'âge adulte. Pour analyser les phrases, nous avons créé un schéma d'annotation des textes, construit un analyseur syntaxique et développé un ensemble de règles.

L'interface du programme est conçue pour que les enseignants ou les chercheurs puissent copier des textes écrits par des apprenants et les soumettre à l'analyseur. Le programme identifie les structures caractéristiques du développement grammatical du français et affiche les résultats à l'utilisateur. Le premier objectif de Direkt Profil est d'être un outil dans la recherche des stades de développement du français écrit en identifiant des telles structures. À plus long terme, le but du système est de produire une évaluation des textes d'élèves en français sous la forme d'un profil d'apprenant.

2 Le corpus CEFLE de Lund

Pour le développement et l'évaluation de notre système, nous avons utilisé le corpus CEFLE de Lund (« Corpus Écrit de Français Langue Étrangère de Lund »). Ce corpus contient environ 100 000 mots (Ågren, 2005). Les textes qui le composent sont des récits de longueur et de niveaux variés. Nous l'avons rassemblé en demandant à 85 lycéens suédois et à 22 jeunes Français de raconter par écrit, entre autres, l'histoire évoquée par les images de la Figure 1. Le but du système étant d'analyser le français langue étrangère, nous avons utilisé les textes des Français comme groupe de contrôle.

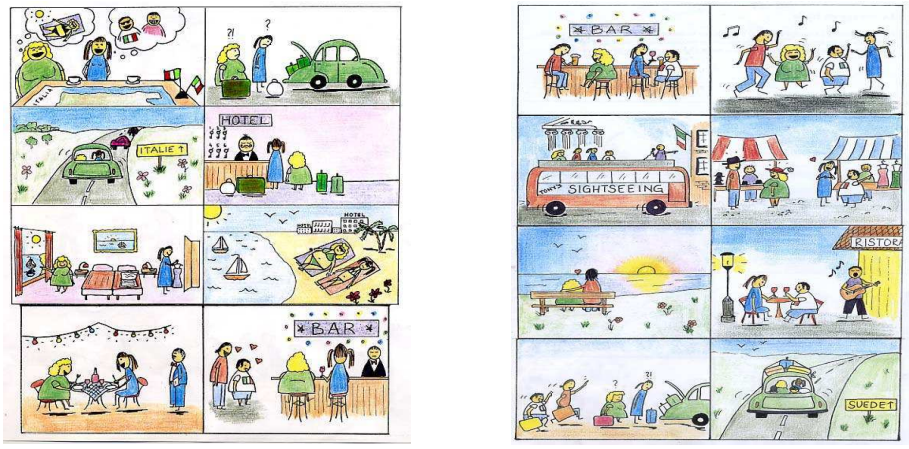


Figure 1 : *Voyage en Italie.*

Le récit qui suit est un exemple provenant d'une apprenante débutante :

> *Elles sont deux femmes. Elles sont a italie au une vacanse. Mais L'Auto est très petite. Elles va a Italie. Au l'hothel elles demande une chambre. Un homme a le clé. Le*

> *chambre est grande avec deux lies. Il fait chaud. C'est noir. Cette deux femmes est a une restaurang. Dans une bar cet deux hommes. Ils amour les femmes. Ils parlons dans la bar. Ils ont tres bien. Le homme et la femme participat a un sightseeing dans la Rome. Ils achetons une robe. La robe est verte. La femme et l'homme reste au un banqe. Ils c'est amour. La femme et l'homme est au une ristorante. es hommes va avec les femmes. L'auto est petite.*

Ce texte contient un certain nombre de constructions caractéristiques : parataxe, ordre des mots très simple, absence de pronoms objets, des formes verbales de base, fautes d'accord verbal et nominal, de genre, orthographe. Les travaux sur l'acquisition d'une langue étrangère ont montré que ces constructions (et d'autres) apparaissent avec une certaine systématicité selon le stade linguistique de l'apprenant. Ils permettent de décrire le développement de grammaires d'apprenants sous la forme d'itinéraires d'acquisition. Le corpus contient les textes bruts annotés avec leur stade de développement (Tableau 1).

3 *Direkt Profil et d'autres systèmes*

Direkt Profil est un analyseur de textes écrits en français comme langue étrangère. Il repose sur les constructions linguistiques des *itinéraires d'acquisition.* Nous avons établi une description systématique de ces itinéraires sous la forme d'une annotation et l'objectif du système est de les détecter automatiquement. L'analyseur parcourt le texte d'un apprenant en annotant et calculant les occurrences d'un phénomène particulier dans ses formes diverses. Le résultat est un profil de texte basé sur ces critères et, éventuellement, une indication du niveau du texte. L'interface présente les résultats à l'utilisateur en visualisant par des couleurs différentes les structures qu'il a détectées. Il est important de souligner que le système n'est pas un correcteur.

La plupart des outils informatiques relevant du domaine ont pour but d'aider à la rédaction. Ils identifient et parfois corrigent des fautes d'orthographe et des erreurs de grammaire. La lignée de programmes aboutissant à PLNLP (Jensen et al. 1993) et NLPWin (Heidorn 2000) est l'une des réalisations les plus notables. Le correcteur grammatical de PLNLP opère une analyse syntaxique complète. Il a été créé pour l'anglais et appliqué ensuite à d'autres langues dont le français. Il utilise des règles syntagmatiques binaires et prend en compte des relations de dépendance. PLNLP s'adresse essentiellement, mais non exclusivement, à des utilisateurs rédigeant dans leur langue maternelle.

D'autres systèmes relèvent de l'enseignement des langues assisté par ordinateur (ELAO) tels que *FreeText* (Granger et al., 2001) pour le français et *Granska* (Bigert et al. 2004) pour le suédois. *FreeText* se place dans une approche communicative à l'apprentissage des langues. Il utilise un analyseur syntaxique chomskyen pour le français. En cas d'échec, il opère à un relâchement de contraintes, par exemple sur les accords, pour diagnostiquer une erreur. *Granska*, à la différence de *FreeText*, réalise une analyse syntaxique partielle. Les auteurs justifient ce type d'analyse par une robustesse qu'ils jugent supérieure et qui permet d'accepter plus facilement des phrases incorrectes.

4 Une méthode d'analyse fondée sur les itinéraires d'acquisition du langage

Les systèmes actuels diffèrent en ce qui concerne le type d'analyse opérée : analyse complète ou partielle de la phrase. L'analyse complète et la correction d'erreurs sont difficilement

applicables aux textes d'apprenants de (très) bas niveau linguistique parce que le nombre de mots inconnus et de phrases incorrectes y sont très souvent élevés.

Dans notre corpus de test de 6 842 mots, la distribution des mots inconnus et de phrases incorrectes était la suivante. Au Stade 1, près de 100 % des phrases sont incorrectes (98,9 %) et 24,7 % des mots sont inconnus[1,2]. À ce stade, toute analyse complète des phrases nous semble très difficile. En revanche, dans le groupe de contrôle les chiffres correspondants sont de 32,7 % pour les phrases incorrectes et de 10,6% pour les mots inconnus. La Figure 2 montre aussi que la seule quantification des mesures de « mots inconnus » et « phrases incorrectes » est insuffisante pour définir le niveau linguistique des textes des apprenants. Les apprenants du Stade 3 produisent moins de phrases incorrectes que les apprenants du Stade 4 (70,5% vs. 80,2%). De plus, le pourcentage de mots inconnus chez le groupe de contrôle (natifs) est légèrement supérieur à celui des apprenants du Stade 4 (10,6% vs. 10,4%) cf. note 1 sur la définition du mot inconnu utilisé ici. Ainsi, le simple calcul des erreurs ne suffit pas pour distinguer les apprenants entre eux et les apprenants des natifs. La distinction des apprenants de différents niveaux linguistiques nécessite des analyses plus détaillées et des mesures plus fines. C'est exactement l'objet des *itinéraires d'acquisition* et de l'analyse de *Direkt Profil*.

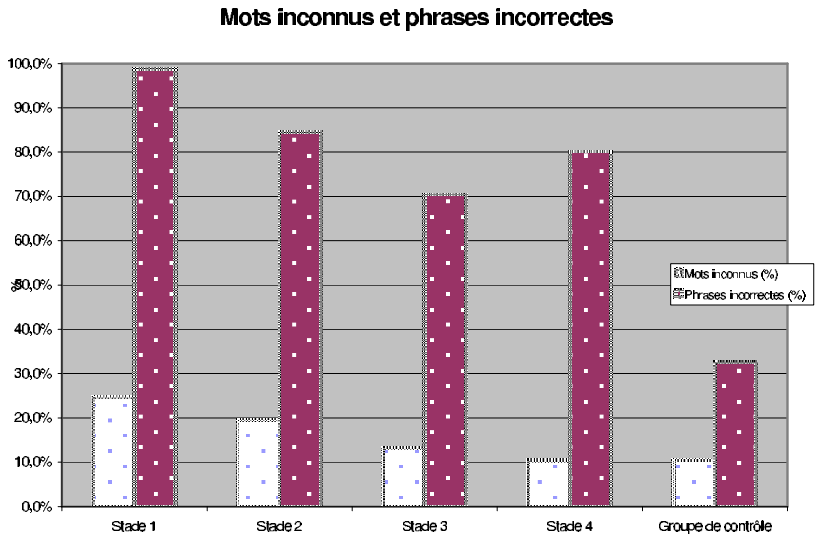


Figure 2. *Les mots inconnus et phrases incorrectes dans le corpus de test.*

[1] Est défini comme « mot inconnu » un mot qui dans sa forme actuelle n'apparaît pas dans le lexique employé par le système (en l'occurrence le lexique de l'ABU CNAM, voir ci-dessous).

[2] Est définie comme « phrase incorrecte » : toute phrase qui contient ou moins une erreur orthographique, syntaxique, morphologique ou sémantique. Pour le niveau de la sémantique, un critère d'interprétation était utilisé : une phrase était considerée sémantiquement incorrecte si elle n'avait aucun sens. Les phrases syntaxiquement etc correctes mais sémantiquement non appropriées dans le contexte (par exemple par rapport aux images, cf. section 2) étaient jugées correctes. De plus, aucune évaluation du choix de mots ou du registre des mots n'a été faite.

4.1 Les itinéraires d'acquisition

La différence la plus importante entre *Direkt Profil* et les autres systèmes actuels est la méthode sur laquelle se fonde l'analyse. *Direkt Profil* réalise une analyse des phénomènes locaux qui se sont avérés liés à un développement dans l'apprentissage du français. Ces phénomènes sont décrits sous forme d'*itinéraires d'acquisition*. Les itinéraires sont le résultat d'observations empiriques montrant que certaines constructions grammaticales sont acquises et peuvent être produites dans la langue parlée spontanée dans un ordre fixe. Cet ordre a aussi été nommé « ordre naturel », en anglais « natural order ».

Clahsen et al. (1983) ainsi que Pienemann & Johnston, (1987) ont établi des itinéraires d'acquisition pour allemand et l'anglais parlé. Pour le français parlé, Schlyter (2003) et Bartning et Schlyter (2004) proposent 6 stades de développement et des itinéraires d'acquisition couvrant plus de 20 phénomènes locaux. Ces phénomènes morphosyntaxiques sont décrits sous la forme de structures locales à l'intérieur du domaine verbal ou nominal. Le Tableau 1 en présente un sous-ensemble.

Ph	**Stades**	**1**	**2**	**3**	**4**	**5**	**6**
A.	Énoncés contenant un verbe	20-40%	30-40%	50%	60%	70%	75%
B.	Formes conjuguées (Types de verbes en opposition)	Pas d'opposition	10-20%	50%	75-100%	+	+
C.	Formes conjuguées (*je parle*) vs. (**je parlE*) (% occurrences)	50% - 75%	70-80%	80-90%	90-98%	+	+
D.	1-2-3 pers singulier (*être/avoir*)	formule sans opposition: *j'ai/ c'est*	opposition (ex. *j'ai* vs. *il a*)	erreurs isolées (ex. **je va* * *je a)*	+	+	+
E.	1re pers. pluriel (V-*ons)* (% correct)	-	70-80%	80-95%	erreurs dans des constr. complexes	+	+
F.	Sujet + *viennent, veulent, prennent*	-	- (ex. **ils prend*)	occs. isolées de V-*(n)ent*	50%	problèmes encore	+
G.	Placement des pronoms objet	-	SVO	S(v)oV	SovV apparaît	SovV productif	+
H.	Genre Article-Nom (% correct)	55-75%	60-80%	65-85%	70-90%	75-95% ?	90-100%

Tableau 1. *Itinéraires d'acquisition (exemples) et stades proposés.*
Légende : Ph = Phénomène ; occs. = occurrences ; - = pas d'occurrence / pas encore acquis ; + = acquis à 100% ; opp(osition) = deux formes différentes d'un verbe particulier dans le même enregistrement / texte ; formule = expression figées dans la langue de l'apprenant ; V-*ons*/V-*(n)ent* = Racine verbale + flexion.

L'axe horizontal indique le développement en fonction du temps d'un phénomène particulier, donc son itinéraire de développement. L'axe vertical indique l'ensemble de phénomènes grammaticaux étudiés regroupés de telle façon qu'ils présentent des caractéristiques pour l'établissement d'un stade acquisitionnel. Pour illustrer, comparons les phénomènes C (occurrences des formes verbales conjuguées) et G (pronoms objet).

Dès le Stade 1, les formes conjuguées et les formes non-conjuguées coexistent. On trouve aussi bien « je parle » (transcription de /je parl/ analysé comme une « forme conjuguée ») que « je parler » (transcription de /je parle/ analysé comme une forme « non-conjuguée »). L'estimation actuelle est qu'au Stade 1, il y a entre 50 et 75% de formes conjuguées (calculées sur les occurrences des verbes dans un contexte où ils sont normalement conjugués). Au Stade 4, le pourcentage de formes conjuguées a augmenté jusqu'à 90-98%. Pour ce phénomène morphologique, les itinéraires d'acquisition décrivent une morphologisation successive.

Le phénomène G décrit l'itinéraire d'acquisition des pronoms objets. Les premiers pronoms objets sont placés dans une position post-verbale selon le schéma Sujet-Verbe-Objet (SVO), par exemple **je vois le/la/lui* (pour *je le/la vois*). Au stade 3, les apprenants peuvent produire des énoncés selon le schéma SvoV (Sujet-verbe auxiliaire-pronom objet-Verbe), par exemple *Je veux le voir* (correct) mais aussi **j'ai le vu* (incorrect). Au stade 4, *je l'ai vu* apparaît. Pour ce phénomène syntaxique, les itinéraires d'acquisition décrivent un changement dans l'organisation linéaire des constituants concernés.

Ce tableau est soumis à une révision continuelle et les taux sont encore approximatifs. L'important est que les itinéraires d'acquisition nous fournissent un grand nombre d'hypothèses détaillées sous la forme de structures locales et qui ensemble couvrent une grande partie de la langue (domaine verbal et nominal).

5 Annotation

Le concept de groupe, nominal ou verbal, correct ou non, représente le support grammatical essentiel de notre annotation. La plupart des normes d'annotation syntaxique pour le français tiennent compte d'une manière ou d'une autre de tels groupes. Celle que propose Gendner et al. (2004) réconcilie un grand nombre de pratiques et forme une base consensuelle. Ces normes sont cependant insuffisantes pour rendre compte des constructions du Tableau 1.

Nous avons défini une annotation des textes propres au projet *Direkt Profil*. Elle repose sur l'inventaire des phénomènes linguistiques caractéristiques des itinéraires de développement et elle reprend les catégories décrites par Barting et Schlyter (2004) (Tableau 1). Nous avons représenté ces phénomènes par des arbres de décisions dont les nœuds terminaux correspondent à une catégorie d'analyse.

L'annotation de *Direkt Profil* utilise le format XML et annote les textes sur 4 niveaux. Seul le 3^{e} niveau est réellement d'ordre syntaxique.

- Le premier niveau correspond à la segmentation du texte en mots.

- Le deuxième niveau annote les multimots et les expressions figées (par exemple *je m'appelle*). Ces expressions correspondent aux phrases « par cœur » qui ont une grande importance dans les premières années d'apprentissage du français.

- Le troisième niveau correspond à une annotation syntaxique partielle du texte, restreinte aux phénomènes à identifier. Ce niveau balise simultanément chacun des mots avec sa partie du discours et les groupes verbaux et nominaux auxquels ils appartiennent. Le groupe verbal incorpore les pronoms clitiques y compris les sujets. L'élément XML `span` marque les groupes et comporte un attribut pour indiquer leur type dans la grille. Les parties du discours utilisent un élément `tag` avec des attributs pour

indiquer le lemme, la partie du discours et les traits grammaticaux. Pour le groupe verbal, la phrase *Ils parlons dans la bar* extraite du texte d'introduction reçoit l'annotation `<span class="c03"><tag pos="pro:nom:pl:p3:mas"> Ils</tag> <tag pos="ver:impre:pl:p1"> parlons </tag></span> dans la bar`. La classe `c03` s'interprète comme « verbe lexical conjugué sans accord ».

- Le quatrième niveau dénombre les types de structures caractéristiques d'un stade d'acquisition. Il utilise un élément XML `counter`, `<counter id="counter.2" counter_name="passe_compose" rule_id="participe_4b" value="1"/>`.

6 Implantation

Le programme *Direkt Profil* est pour l'instant restreint à l'analyse des groupes verbaux et des pronoms clitiques sujets. Pour chacune des catégories du Tableau 1, le programme détecte les constructions correspondantes dans un texte et les dénombre.

L'analyseur utilise des règles écrites manuellement et s'appuie sur un lexique de formes fléchies. La variété des constructions contenues dans le corpus est grande et afin de ne pas multiplier le nombre de règles, nous avons choisi une stratégie d'analyse par renforcement de contraintes. Conceptuellement, l'analyseur recherche des classes de structures syntagmatiques dont tous les traits ont été ôtés. Il identifie les structures progressivement en faisant varier les valeurs des traits. La reconnaissance des limites des groupes se fait par un ensemble de mots vides et par des heuristiques à l'intérieur des règles. Elle suit ainsi une stratégie ancienne mais robuste utilisée notamment par Vergne (1999), *inter alia*, pour le français.

Direkt Profil applique en cascade trois ensembles de règles pour produire les quatre niveaux d'annotations. Le premier ensemble segmente le texte en mots. Un ensemble intermédiaire identifie les expressions figées. Le troisième ensemble annote simultanément les parties du discours et les groupes. Finalement, le moteur crée un groupe de résultats relié au Stade de l'apprenant. Il est à noter que le moteur n'annote pas tous les mots, ni tous les segments. Il ne considère que ceux qui sont pertinents pour la détermination du Stade. Le moteur applique les règles de gauche à droite puis de droite à gauche pour résoudre certains problèmes d'accord.

Les règles représentent des structures partielles de groupe et sont divisées en une partie condition et une partie action. La partie condition contient les paramètres de recherche. Il peut s'agir d'un lemme, d'une expression régulière ou d'une classe de flexion. Le moteur parcourt le texte et applique les règles à l'aide d'un arbre de décision. Il exécute la partie condition pour identifier les séquences de mots contigus. Chaque règle produit un résultat positif (« *match* ») ou négatif (« *no match* »). Ils sont exécutés suivant le résultat de la condition et ont pour effet d'annoter le texte, de compter le nombre d'occurrences du phénomène et d'enchaîner une autre règle. En parcourant les nœuds de l'arbre, le moteur mémorise les règles qu'il a parcourues sur son chemin ainsi que les résultats des parties condition de ces règles. Quand il arrive à un nœud terminal, le moteur applique les parties action de toutes les règles.

Le moteur recherche les mots dans un dictionnaire de formes fléchies. Il ne corrige pas les fautes d'orthographe à l'exception des accents et de certains radicaux. En effet, les apprenants construisent fréquemment des participes passés erronés à partir d'une généralisation abusive

de radicaux verbaux voisins. Un exemple est le mot **prendu* (*pris*) formé sur le radical de *prendre* et du suffixe de *rendu*. Nous avons utilisé le lexique disponible au site de l'Association des Bibliophiles Universels que nous avons corrigé, transposé en XML et que nous avons enrichi des radicaux des verbes.

7 Interface

Direkt Profil fusionne l'ensemble des niveaux d'annotations dans un objet résultat. Cet objet représente le texte d'origine, l'annotation et la trace de l'application des règles et des compteurs. L'objet résultat, qui peut être enregistré, est ensuite transformé par le programme pour être présenté à l'utilisateur. L'affichage utilise les spécifications XHTML 1.1 qui peuvent être lues par un navigateur internet. *Direkt Profil* fonctionne en mode client-serveur où le serveur réalise l'annotation d'un texte et le client, intégré à un navigateur, prend en charge l'affichage et l'interaction.

La Figure 3 est une copie d'écran de l'interface graphique de *Direkt Profil* qui montre l'analyse du texte présenté dans l'Introduction. L'interface indique à l'utilisateur par une couleur différente toutes les structures que l'analyseur a détectées. Le cadre inférieur donne le code des couleurs et le dénombrement de ces structures.

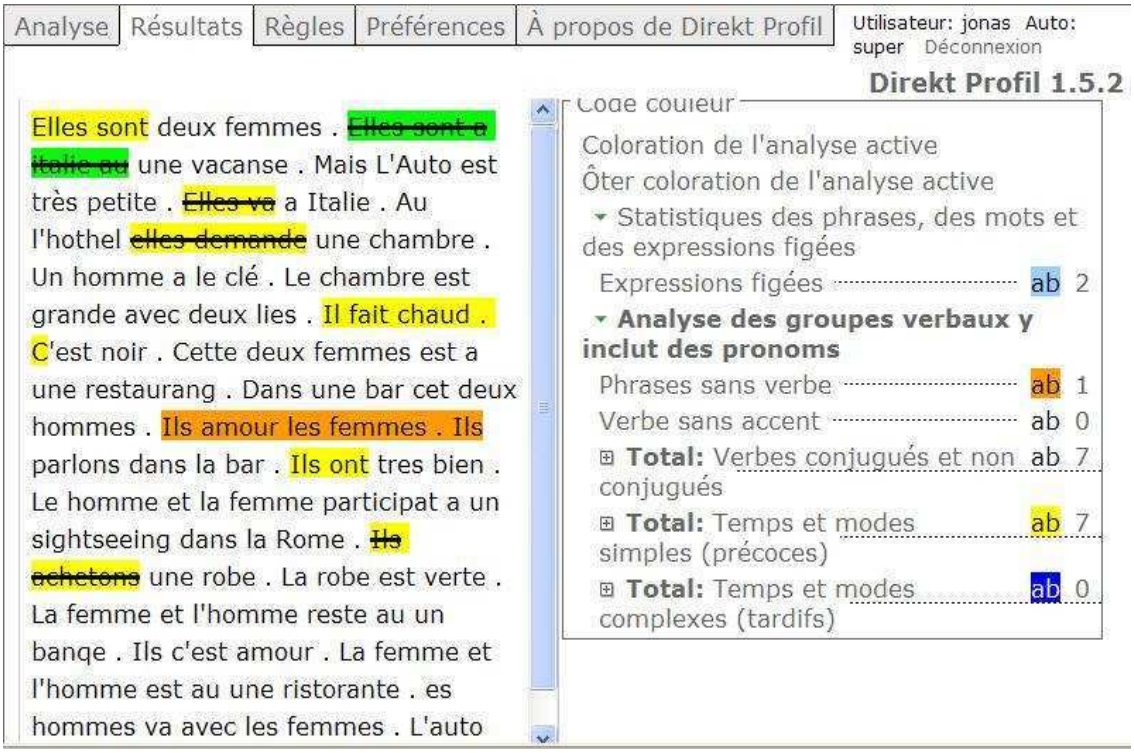


Figure 3. *L'interface graphique de Direkt Profil 1.5.*

8 Résultats et évaluation

Nous avons évalué *Direkt Profil* avec un sous-ensemble du corpus CEFLE de Lund. Nous avons choisi 20 textes au hasard répartis sur 4 Stades d'apprentissage. Nous avons aussi utilisé 5 textes provenant du groupe de contrôle. Nous n'avons pas testé dans cette version la correction des mots mal orthographiés : accent et radicaux. Le Tableau 2 présente quelques

statistiques sur la taille de textes et le Tableau 3 donne les résultats sous forme de rappel et précision.[3]

	Stade 1	Stade 2	Stade 3	Stade 4	Contrôle	Total
Nombre de textes analysés	5	5	5	5	5	25
Nombre de mots	740	1233	1571	1672	1626	6842
Nombre de phrases	85	155	166	126	107	639
Longueur moyenne des textes (mots)	148	247	314	334	325	
Longueur moyenne des phrases	8,7	7,9	9,5	13,3	15,2	

Tableau 2. *Corpus de test.*

	Stade 1	Stade 2	Stade 3	Stade 4	Contrôle	Total
Nombre de structures de référence	23	97	101	119	85	425
Nombre de structures proposées	27	98	100	112	92	429
Nombre de structures détectées correctes	15	81	89	96	73	354
Nombre de structures non détectées	5	16	12	20	11	64
Nombre de structures surdétectées	10	17	11	17	19	74
Rappel	65	84	88	81	86	83
Précision	56	83	89	86	79	83
F-measure	0,6	0,83	0,89	0,83	0,82	0,83

Tableau 3 : *Résumé des résultats de Direkt Profil 1.5.1*

À un niveau global, les résultats montrent que *Direkt Profil* parvient bien à détecter les phénomènes désirés. Il révèle aussi des différences intéressantes suivant les stades des textes. Le tableau montre que les textes du Stade 1 sont les plus difficiles à traiter (rappel de 65%). Ceci est dû en grande partie au nombre de *mots inconnus* à ce stade d'acquisition (cf. *infra* Figure 2). Le résultat en est une surdectection du phénomène « phrases sans verbes » à ce stade. Les résultats montrent aussi que *Direkt Profil* analyse mieux les textes des apprenants que les textes des étudiants français (Groupe de contrôle). Sans savoir exactement à quoi ce résultat est dû, nous pouvons constater qu'il suggère que la démarche adoptée qui vise à l'analyse des textes en français langue étrangère semble prometteuse.

9 Conclusion et travaux futurs

Nous avons présenté un système réalisant une analyse automatique de textes fondée sur les itinéraires d'acquisition dont le but est de produire un profil d'apprenant. Nous avons construit un analyseur syntaxique et développé un ensemble de règles pour annoter les textes. *Direkt Profil* est intégré dans une architecture client-serveur et dispose d'une interface permettant l'interaction avec l'utilisateur.

Les résultats montrent qu'il est possible de décrire sous forme de règles la vaste majorité des structures locales définies par les itinéraires d'acquisition. *Direkt Profil* peut ainsi les détecter

[3] Dans le corpus de test, nous n'avons pas d'exemples de textes d'apprenants des Stades 5 et 6 (cf. Tableau 1). Pour le moment, les textes du groupe de contrôle (des jeunes Français) peuvent servir d'exemples des textes de haut niveau linguistique.

et les analyser automatiquement. Nous pouvons ainsi vérifier la validité des critères d'acquisition établis pour l'oral sur notre corpus écrit.

Direkt Profil peut aussi avoir un rôle pédagogique car il sera possible d'analyser, de manière automatique et précise, le niveau grammatical d'un apprenant ou d'un élève dans la production écrite et libre. Le programme pourra être utilisé d'une part par des professeurs pour évaluer les textes de leurs élèves et d'autre part par les élèves eux-mêmes comme auto-évaluation et dans le perfectionnement de la langue.

Une version préliminaire de *Direkt Profil* est disponible en ligne à l'adresse http://www.rom.lu.se:8080/profil

Références

BARTNING, I, S. SCHLYTER (2004) « Stades et itinéraires acquisitionnels des apprenants suédophones en français L2 ». *JFLS* (Journal of French Language Studies). À paraître.

BIGERT, J., KANN, V., KNUTSSON, O., SJÖBERGH J. (2004). Grammar checking for Swedish second language learners. Chapter in CALL in the Nordic Languages, Copenhagen Studies in Language, Copenhagen Business School.

CLAHSEN, H., MEISEL, J-M., PIENEMANN, M. (1983) *Deutsch als Fremdsprache. Der Spracherwerb ausländischer Arbeiter*. Tübingen: Narr

GENDNER, V., VILNAT, A., MONCEAUX, L., PAROUBEK, P., ROBBA I. (2004) Les annotations syntaxiques de référence PEAS.
http://www.limsi.fr/Recherche/CORVAL/easy/PEAS_reference_annotations_v1.6.html

GRANGER, S., VANDEVENTER A., HAMEL M-J. (2001). « Analyse de corpus d'apprenants pour l'ELAO basé sur le TAL ». *Traitement Automatique des Langues,* 42(2), 609-621.

HEIDORN G. E., Intelligent Writing Assistance, (2000) in Robert Dale, Hermann Moisl, et Harold Somers eds, *Handbook of Natural Language Processing*, Marcel Dekker.

JENSEN, K., HEIDORN G. E., RICHARDSON S. D., (1993) *Natural Language Processing: The PLNLP Approach*, Kluwer Academic Publishers.

PIENEMANN, M, JOHNSTON, M (1987) « Factors influencing the development of second language proficiency". In D. Nunan (ed.) *Applying second language acquisition research*, 45-141. Adelaide: National Curriculum Resource Centre.

SCHLYTER, S. (2003). « Stades de développement en français L2 ». Ms. Institut d'études romanes de Lund. Université de Lund. Disponible en ligne : http://www.rom.lu.se/durs/STADES_DE_DEVELOPPEMENT_EN_FRANCAIS_L2.pdf

VERGNE, J. (1999) *Étude et modélisation de la syntaxe des langues à l'aide de l'ordinateur. Analyse syntaxique automatique non combinatoire. Synthèse et Résultats.* Habilitation à Diriger des Recherches, 29 septembre 1999, Caen.

ÅGREN, M. (2005) « Le marquage morphologique du nombre dans la phrase nominale. Une étude sur l'acquisition du français L2 écrit » Ms. Institut d'études romanes de Lund. Université de Lund.

ILIMP : Outil pour repérer les occurrences du pronom impersonnel *il*

Laurence Danlos

LATTICE, Université Paris 7, Institut Universitaire de France
Laurence.Danlos@linguist.jussieu.fr

Mots-clés : Pronom impersonnel (explétif), Pronom anaphorique, Lexique-grammaire, Automates, Résolution d'anaphores, Analyse syntaxique modulaire

Keywords: Expletive pronouns, Anaphoric pronouns, Lexicon-Grammar, Automata, Anaphora resolution, Modular syntactic analysis

Résumé Nous présentons un outil, ILIMP, qui prend en entrée un texte brut (sans annotation linguistique) rédigé en français et qui fournit en sortie le texte d'entrée où chaque occurrence du pronom *il* est décorée de la balise [ANAphorique] ou [IMPersonnel]. Cet outil a donc comme fonctionnalité de distinguer les occurrences anaphoriques du pronom *il*, pour lesquelles un système de résolution des anaphores doit chercher un antécédent, des occurrences où *il* est un pronom impersonnel (explétif) pour lequel la recherche d'antécédent ne fait pas sens. ILIMP donne un taux de précision de 97,5%. Nous présentons une analyse détaillée des erreurs et nous décrivons brièvement d'autres applications potentielles de la méthode utilisée dans ILIMP, ainsi que l'utilisation et le positionnement d'ILIMP dans un système d'analyse syntaxique modulaire.

Abstract We present a tool, ILIMP, which takes as input a French raw text and which produces as output the input text in which every occurrence of the word *il* is tagged either with the tag [ANA] for anaphoric or [IMP] for expletive. This tool is therefore designed to distinguish the anaphoric occurrences of *il*, for which an anaphora resolution system has to look for an antecedent, from the expletive occurrences of this pronoun, for which it does not make sense to look for an antecedent. The precision rate for ILIMP is 97,5%. The few errors are analyzed in detail. Other tasks using the method for ILIMP are described briefly, as well as the use of ILIMP in a modular syntactic analysis system.

1 Introduction

En TAL, la résolution d'anaphores est un sujet de recherche fort étudié car c'est une question cruciale pour des applications comme la Recherche d'Informations ou le Résumé Automatique. Parmi les anaphores, les pronoms sont largement traités car fréquents et facilement identifiables. Parmi les pronoms, on peut distinguer un élément (*il* en français, *it* en anglais) avec un emploi impersonnel (explétif) *(il pleut, it rains)* qui se distingue de l'emploi anaphorique *(il est cher, it is expensive)*. Un système de résolution des anaphores doit être capable de repérer les occurrences des pronoms impersonnels avant de s'attaquer aux pronoms anaphoriques et aux autres anaphores. De ce fait, il existe un certain nombre de travaux sur les emplois impersonnels du pronom anglais *it*, citons (Lapin, Leass, 1994), (Kennedy, Bogurev, 1996) et (Evans 2001). Mais à notre connaissance, il n'existe pas de travaux similaires sur le pronom français *il*. Ce travail présente un outil, ILIMP, qui est conçu pour reconnaître toutes les occurrences du pronom impersonnel *il* dans les textes français : ILIMP décore chaque occurrence de *il* de la balise [ANAphorique] ou [IMPersonnel]. Cet outil est à base de règles (comme c'est le cas du système de Lapin et Leass) ; il travaille sur des textes bruts sans annotation linguistique (contrairement au système de Lapin et Leass qui repose sur une analyse syntaxique).

Si ILIMP est un outil s'imposant en amont d'un système de résolution d'anaphores, c'est aussi un outil qui peut s'intégrer dans la chaîne de traitements d'un analyseur syntaxique modulaire. En effet, d'une part, les balises [IMP] et [ANA] sur le pronom *il* peuvent être vues comme un raffinement des étiquettes morpho-syntaxiques traditionnellement utilisées dans les taggeurs : l'étiquette "pronom" serait remplacée par deux étiquettes, "pronom anaphorique" et "pronom impersonnel". Or on sait que plus le jeu d'étiquettes morpho-syntaxiques d'un taggeur est riche, plus un analyseur venant en aval de ce taggeur a des chances d'aboutir à l'analyse syntaxique correcte (Nasr, 2004). D'autre part, nous verrons que des outils dérivés d'ILIMP peuvent être utilisés pour d'autres annotations linguistiques.

La Section 2 présente notre méthode qui est basée, pour l'aspect linguistique, sur le lexique-grammaire du LADL, et pour l'aspect informatique sur UNITEX. La Section 3 décrit la réalisation d'ILIMP, les difficultés que nous avons rencontrées et les choix effectués pour les surmonter. Finalement, la Section 4 donne une évaluation d'ILIMP et discute de son positionnement dans une analyse syntaxique modulaire.

2 Méthode

2.1 Lexique-grammaire

Comme la plupart des phénomènes linguistiques, les constructions impersonnelles reposent sur des conditions tant lexicales que syntaxiques. Par exemple, l'adjectif *violet* ne peut jamais être la tête lexicale d'une phrase impersonnelle, (1a), l'adjectif *probable* ancre une phrase impersonnelle lorsqu'il est suivi d'un complément phrastique, (1b), l'adjectif *difficile* ancre une phrase impersonnelle (resp. personnelle) lorsqu'il est suivi d'une infinitive introduite par la préposition *de* (resp. *à*), (1c) et (1d).

(1)a Il est violet
b Il est probable que Fred viendra
c Il est difficile de résoudre ce problème
d Il est difficile à résoudre (ce problème)

De ce fait, le lexique-grammaire du français développé par Maurice Gross et son équipe (Gross1994, Leclère 2003) est une ressource linguistique appropriée pour ILIMP puisqu'il décrit l'ensemble des têtes lexicales des phrases simples du français avec leurs arguments syntaxiques et les alternances possibles. Nous avons donc extrait (manuellement) du lexique-grammaire tous les items lexicaux qui peuvent ancrer une phrase impersonnelle en enregistrant leur complémentation syntaxique. Présentons un bref aperçu des constructions impersonnelles du français que nous avons recensées. On peut distinguer les constructions intrinsèquement impersonnelles, qui ne peuvent avoir comme sujet que *il,* des constructions avec un "sujet profond extraposé". Parmi les premières, on trouve 45 verbes météorologiques de la table 31i de (BGL, 1976a) *(Il pleut, Il fait beau)*, 21 verbes de la table 17 de (Gross, 1975) *(Il faut du pain /que Fred vienne)* et 38 expressions figées de (Gross, 1993) (Il é*tait une fois, quoi qu'il en soit)*. Pour les constructions impersonnelles à sujet profond extraposé, on peut distinguer celles à sujet phrastique de celles à sujet nominal. Parmi les premières, on trouve 682 adjectifs dispersés dans les tables de (Picabia, 1978) et (Meunier, 1981) *(Il est probable que Fred viendra)*, 88 expressions "être Prép X" des tables Z5P et Z5D de (Danlos 1980) *(Il est de règle de porter un chapeau)*, 21 verbes de la table 5 de (Gross 1975) *(Il plaît à Vanne que Fred vienne)*, et enfin 140 verbes de la table 6 et 92 verbes de la table 9 de (Gross 1975) construits au passif ou au se-moyen *(Il a été dit/se raconte que Fred viendra)*. Les constructions impersonnelles à sujet extraposé nominal ont pour tête lexicale des verbes qui sont dispersés dans les tables élaborées par (BGL, 1976b). On peut distinguer d'un côté des verbes comme *manquer* ou *rester* dont l'emploi en construction impersonnelle est tout à fait courant *(Il manque /reste du pain)*, et de l'autre côté des verbes "inacusatifs" (*Il est venu trois personnes)* ou des verbes construits au passif *(Il a été mangé trois gâteaux)*, dont l'emploi dans une construction impersonnelle relève d'un niveau de langue châtié.

2.2 UNITEX

UNITEX[1] est un outil qui permet d'écrire des patrons linguistiques (expressions régulières ou automates) qui sont localisés dans le texte d'entrée, avec un éventuel ajout d'annotations lorsque les automates sont en fait des transducteurs. Un texte brut donné en entrée à UNITEX est d'abord pré-traité : le texte est segmenté en phrases et les phrases segmentées en tokens. Chaque token est étiqueté avec *toutes* les parties du discours et traits flexionnels enregistrés dans le "full-form" dictionnaire DELAF (Courtois, 2004). Il n'y a pas de désambiguïsation, autrement dit le pré-traitement d'UNITEX n'est pas équivalent à un étiquetage morpho-syntaxique.

Pour réaliser ILIMP, l'idée de base consiste à repérer les constructions impersonnelles grâce à leur tête lexicale et à leur complémentation. Il s'agit donc d'écrire (manuellement) un ensemble de transducteurs comme celui présenté en (2) sous une forme linéaire simplifiée. La balise [IMP] est l'ajout d'information amenée par l'aspect transducteur de (2). Les éléments entre chevrons de (2) se glosent de la façon suivante : <être.V:3s> correspond à toutes les formes du verbe *être* conjugué à la troisième personne du singulier, <Adj1:ms> correspond aux adjectifs masculins singuliers de la classe Adj1 qui regroupe des adjectifs se comportant comme *difficile*, <V:W> correspond aux verbes à l'infinitif.

(2) Il [IMP] <être.V:3s> <Adj1:ms> de <V:W>

---

[1] UNITEX est un logiciel sous licence GPL, dont l'ancêtre est INTEX (Silberstein, 1994). La documentation et le téléchargement de UNITEX se trouvent sur le site http://ladl.univ-mlv.fr.

La balise [IMP] - abréviation de impersonnelle- vient décorer les occurrences de *il* qui apparaissent dans les phrases correspondant au patron de (2). Cette balise vient donc décorer *il* dans (1c). La balise [ANA] - abréviation de anaphorique- est la balise par défaut : elle vient décorer les occurrences de *il* qui n'ont pas été balisées par [IMP]. Cette balise vient décorer *il* dans (1d). Néanmoins, la situation est un peu plus complexe, car il existe une troisième balise [AMB] - abréviation de ambigu- qui sera expliquée dans la section 3.2. Après cet exposé des principes théoriques sous-tendant ILIMP, passons à sa réalisation pratique.

3 Réalisation de ILIMP

3.1 Contexte gauche de la tête lexicale

Dans l'exemple (1c), le contexte gauche de la tête lexicale de la phrase - la séquence de tokens à gauche de *difficile* - se réduit à *Il est*. Mais on trouve fréquemment dans les corpus des phrases comme (3a) ou (3b) dans lesquelles le contexte gauche de la tête lexicale est plus complexe. Dans (3a), ce contexte gauche inclut (de droite à gauche) l'adverbe *très* qui modifie l'adjectif, le verbe *paraître,* à la forme infinitive, qui est "un verbe support" (Danlos, 1992) pour les adjectifs, le pronom *lui* et finalement le verbe modal *peut* précédé par *il*. Dans (3b), le contexte gauche inclut le verbe support *s'avérer* qui est conjugué à un temps composé (*s'est avéré*) et nié (*ne s'est pas avéré*).

(3)a Il peut lui paraître très difficile de résoudre ce problème
b Il ne s'est pas avéré difficile de résoudre ce problème

De ce fait, pour chaque type de tête lexicale (e.g. adjectif, verbe) qui ancre une construction impersonnelle, on doit déterminer tous les éléments qui peuvent figurer dans son contexte gauche et intégrer ces éléments dans les patrons linguistiques. Cette tâche ne se heurte pas à de réelles difficultés, disons que c'est un travail minutieux et coûteux en temps[2]. Par contre, on se heurte à de réelles ambiguïtés avec le contexte droit de la tête lexicale, comme nous allons le montrer. Dans la suite de cet article, les contextes gauches des têtes lexicales sont présentés de façon simplifiée - comme en (2) - pour faciliter la lecture.

3.2 Contexte droit de la tête lexicale

3.2.1 Ambiguïtés syntaxiques

Les ambiguïtés syntaxiques sont légion dans le contexte droit car, comme il est bien connu, une séquence de parties du discours peut recevoir plusieurs analyses syntaxiques. A titre d'illustration, considérons le patron en (4a), dans lequel le symbole Ω correspond à une séquence non-vide de tokens. Ce patron correspond à deux analyses syntaxiques : (4b) dans lequel *il* est impersonnel et l'infinitive sous-catégorisée par *difficile,* et (4c) dans lequel *il*

[2] Ce travail peut être réutilisé dans un outil qui repère la tête lexicale d'une phrase simple, outil qui peut s'intégrer dans la chaîne de traitements d'un analyseur syntaxique modulaire. Ce travail revient à mettre sous forme d'automates la notion d'"amas verbal" de (Gerdes, Kahane, 2004).

est anaphorique et l'infinitive fait partie d'un *GN*. Ces deux analyses sont illustrées respectivement dans les phrases (4b) et (4e) - ces phrases diffèrent seulement par l'adverbe *ici/juste*.

(4)a Il est difficile pour Ω <de V:W>
b Il [IMP] est difficile pour $(\Omega)_{GN}$ <de V:W>
c Il [ANA] est difficile pour $(\Omega$ de < V:W>$)_{GN}$
d Il est difficile pour (les étudiants qui viennent ici)$_{GN}$ de résoudre ce problème
e Il est difficile pour (les étudiants qui viennent juste de résoudre ce problème)$_{GN}$

Pour traiter ces ambiguïtés syntaxiques, une solution consiste à déclarer explicitement qu'un patron comme (4a) est ambigu, ce qui revient à décorer l'occurrence de *il* dans (4a) par la balise [AMB] qui doit être interprétée de la façon suivante : "ILIMP ne peut pas déterminer si *il* est anaphorique ou impersonnel". Cependant cette étiquette n'est d'aucune utilité pour les traitements ultérieurs d'un système de résolution d'anaphores ou d'une chaîne de traitements syntaxiques : elle doit donc être utilisée avec modération. Une autre solution consiste à faire appel à des heuristiques basées sur des fréquences. Par exemple, l'heuristique suivante : les phrases qui suivent le patron de (4a) sont plus fréquemment analysées comme (4b) que comme (4c), par conséquent, *il* dans les phrases qui suivent (4a) peut recevoir la balise [IMP], même si cette balise est fausse dans quelques cas. Nous avons adopté cette dernière solution. ILIMP repose donc sur tout un ensemble d'heuristiques que nous avons établies soit à partir de notre connaissance/intuition linguistique soit à partir d'études quantitatives sur les corpus. L'évaluation de ILIMP révèlera si nos heuristiques sont judicieuses (Section 4).

3.2.2 Ambiguïtés lexicales

Dans un très petit nombre de cas (une dizaine), un item lexical peut ancrer une construction impersonnelle ou personnelle avec le même cadre de sous-catégorisation. C'est le cas pour l'adjectif *certain* construit avec un complément phrastique, comme illustré dans la phrase en (5a). Comme les deux lectures de (5a) semblent également fréquentes, *il* dans le patron (5b) reçoit la balise [AMB].

(5) a Il est certain Fred que viendra (*Jean/Cela est certain que Fred viendra)*
b Il [AMB] est certain que P[3]

3.2.3 Autres difficultés

Le dernier type de difficultés s'observe avec des constructions impersonnelles à sujet nominal extraposé qui ne diffèrent en surface que de façon très subtile par rapport à des constructions personnelles. Voir la paire en (6) qui ne diffère que par *du/de* mais où (6a) est impersonnelle et (6b) personnelle. Voir aussi la paire en (6') qui ne diffère que par les noms *valise/ priorité* mais où (6'a) est impersonnelle et (6'b) personnelle. Nous avons tenté d'établir des heuristiques pour distinguer ces cas, sans toutefois nous lancer, par exemple, dans l'utilisation (périlleuse) de traits sémantiques, comme ± concret.

(6)a Il manque du poivre (dans cette maison)

---

[3] *P* symbolise un patron destiné à représenter une phrase. Il est composé d'une séquence non vide de tokens incluant un verbe fini.

b Il manque de poivre, ce rôti
(6')a Il reste la valise du chef (dans la voiture)
b Il reste la priorité du chef (le chômage)

Pour conclure cette section sur la réalisation de ILIMP, disons qu'il est fait fréquemment appel à des heuristiques afin d'éviter une utilisation abusive de l'étiquette [AMB]. Ces heuristiques peuvent mener à des erreurs qui vont être examinées dans la section suivante.

4 Evaluation de ILIMP

Notre corpus de travail a été *Le Monde*. Plus précisément, un corpus de 3.782.613 tokens extraits du corpus *Le Monde 1994*. UNITEX segmente ce corpus en 71.293 phrases. Il contient 13.611 occurrences du token *il* sur 20.549 occurrences de pronoms personnels sujet de la troisième personne *(il, elle, ils, elles)*. Le pronom *il* est donc le pronom sujet de la troisième personne le plus fréquent, avec un taux de 66 %. De ce corpus, 8544 phrases qui incluent au moins une occurrence de *il* ont été extraites et elles totalisent près de 10.000 occurrences de *il* (une phase complexe enchâssant diverses propositions peut inclure plusieurs occurrences de *il*). Ces phrases ont été données en entrée à ILIMP et les résultats - les balises [IMP], [ANA], et [AMB] - ont été évalués manuellement par des amis, collègues et étudiants[4]. Ces évaluateurs devaient se fier uniquement à leur intuition de locuteur : ils ne devaient pas mettre en œuvre leur éventuelle compétence de linguiste pour essayer de détecter des ambiguïtés virtuelles d'occurrences de *il*. Dans ces conditions, l'attribution d'une balise [IMP] ou [ANA] est immédiate dans quasiment tous les cas, la balise [AMB] ne concernant qu'un nombre négligeable de cas (rappelons, section 3.2.2, qu'elle n'est lexicalement justifiée que pour une dizaine de têtes lexicales). Il ressort de cette évaluation un taux de précision de 97,5%. Nous allons examiner en détail les erreurs, en laissant de côté la balise [AMB].

4.1 Erreurs provenant d'ambiguïtés morphologiques

Les erreurs d'ILIMP provenant d'ambiguïtés morphologiques sont (évidemment) comptabilisées comme les autres erreurs provenant de la réalisation d'ILIMP. Ces dernières seront examinées dans les sections suivantes. Pour l'instant, examinons les erreurs dues au fait que le pré-processing d'UNITEX n'effectue pas de désambiguïsation : ce n'est pas un taggeur (Section 2.2). Considérons le patron en (7a) : rappelons (note 3) que P représente une séquence non vide de tokens qui inclut un verbe fini ; <V6:K> couvre les verbes de la table 6 au participe passé, e.g. *choisi*. Le patron en (7a) est destiné à couvrir les phrases impersonnelles comme (7b). Néanmoins, il couvre aussi (7c), où le pronom *il* est donc balisé à tort [IMP]. Cette erreur est due au fait que le dictionnaire DELAF comprend à juste titre deux entrées pour le mot *mètres* - forme finie du verbe *métrer* et pluriel du nom mètre - et qu'UNITEX ne fait aucune distinction entre ces deux entrées. La séquence *l'acier ou le béton pour soutenir une toiture de 170 mètres* correspond donc au patron P (elle contient un verbe fini).

(7)a Il [IMP] <avoir.V:3s> été <V6:K> (ADV) que P

---

[4] A savoir Isabelle Faugeras, Annie Meunier, Christian Leclère, Laurence Delort, Ane Dybro-Johansen, François Lareau, Alexis Nasr, Céline Raynal, Jacques Steinlin, François Toussenel et Mélodie Soufflard, que nous remercions chaleureusement.

b Il a été choisi que les séances se feraient le matin vers 9h
c Il a été choisi plutôt que l'acier ou le béton pour soutenir une toiture de 170 mètres

Tout taggeur devrait attribuer au mot *mètres* de (7c) une étiquette nominale. Si ILIMP travaillait non pas sur du texte brut mais sur la sortie d'un taggeur, l'erreur de balisage de *il* dans (7c) serait donc évitée. Cette stratégie est envisageable[5], mais ILIMP serait alors tributaire des erreurs d'un taggeur. D'une manière plus générale, en admettant qu'un système d'analyse syntaxique repose sur une approche modulaire séquentielle où collabore "en pipe-line" tout un ensemble de modules - taggeur, reconnaisseur d'entités nommées, ILIMP, chunker, etc.- la question se pose de savoir dans quel ordre enchaîner ces modules. Mais laissons cette question ouverte et revenons aux erreurs d'ILIMP travaillant sur un texte brut.

4.2 *il* balisé à tort [IMP] au lieu de [ANA] : 0,3%

Très peu d'erreurs : 33. Ce faible taux d'erreur est surprenant vu le recours fréquent à des heuristiques "brutales". A titre d'illustration, nous avons mis la balise [IMP] dans le patron *il y <avoir.V:3s>* avec ses variantes de contexte gauche. Cette heuristique ne donne lieu qu'à deux erreurs, citées en (8), sur environ 1500 phrases qui suivent ce patron où *il* est correctement balisé [IMP].

(8)a Il s'était réfugié en Angleterre. Il y avait très tôt connu les pianos de ….
b Il revient de Rimini. Il y a donné la réplique à Madeleine.

4.3 *il* balisé à tort [ANA] au lieu de [IMP] : 2%

Plus d'erreurs. Les erreurs de ce type viennent de ce que [ANA] est la balise par défaut : elles viennent donc de lacunes dans l'ensemble des patrons composant ILIMP. Parmi celles-ci, on peut d'abord distinguer celles dues à de la paresse/lassitude/manque de temps. Par exemple, nous avons autorisé des guillemets à certains endroits mais pas partout, de ce fait *il* est balisé à tort [ANA] dans (9a). De même, nous avons écrit certains automates pour traiter les cas avec inversion du sujet, mais nous n'avons pas pris le temps de les écrire tous, d'où l'erreur en (9b). Et nous avons fait l'impasse totale sur toute coordination, d'où (9c).

(9)a Il [ANA] était " même souhaitable " que celui-ci soit issu " de l'opposition ".
b Est-il [ANA] inconcevable que ...
c Il [ANA] est donc indispensable et légitime de les aider …

Un second type d'erreurs provient de lacunes lexicales dans nos patrons. Dans l'état actuel d'ILIMP, il manque principalement des adjectifs ancrant une construction impersonnelle, car les adjectifs n'ont pas été étudiés de façon aussi systématique que les verbes dans le lexique-grammaire. La liste des 682 adjectifs ancrant une construction impersonnelle – à sujet phrastique extraposable - peut donc encore être complétée. Un relecteur anonyme de cet article pose la question de savoir si on peut effectivement recenser les adjectifs ayant un usage dans une construction impersonnelle. Il/Elle note qu'un adjectif comme *myope*, qui *a priori* n'ancre pas de construction impersonnelle, pourrait en fait permettre une phrase impersonnelle comme (10).

---

[5] Elle demande qu'UNITEX soit adapté pour prendre en compte les résultats d'un taggeur, ce qui a été fait par Patrick Watrin pour Tree Tagger.

(10) Il semble tout à fait myope, voire aveugle, de penser que la situation ne peut se détériorer

Si cette phrase est joliment construite et compréhensible, elle nous semble quand même inacceptable (même en tentant de l'améliorer en ajoutant *de sa part* sur le modèle de *Il est idiot de sa part d'avoir refusé cette offre*). Nous sommes plusieurs linguistes à juger (10) aussi inacceptable que **Cette action/idée est myope*[6]. Est-ce à dire qu'il est exclu de trouver une phrase comme (10) en corpus ? Totalement exclu, peut-être pas, car elle présente un joli effet de style. Mais il semble qu'on peut affirmer qu'elle est hautement improbable. De ce fait, *myope* et *aveugle* peuvent être exclus de la liste des adjectifs ancrant une construction impersonnelle. De manière plus générale, nous pensons que la liste des adjectifs (ou verbes) ancrant une construction impersonnelle n'est pas ouverte aux glissements de sens : c'est une liste fermée, donc recensable.

Un troisième type d'erreurs provient de lacunes syntaxiques. En particulier, nous avons considéré obligatoire un sujet phrastique extraposé, alors qu'il existe des cas où il n'est pas réalisé, par exemple dans des expressions en *comme*, comme en (11). Nous avons ajouté certaines de ces expressions, mais nous n'avons pas mené d'étude linguistique pour connaître l'étendue du phénomène, nous avons donc des lacunes syntaxiques.

(11) Comme il / a été annoncé / conviendrait / arrive souvent / est bien connu

Un quatrième et dernier type d'erreurs concerne les constructions impersonnelles à sujet profond nominal. D'une part, comme nous l'avons expliqué en 3.2.3, ces constructions sont délicates à repérer pour des verbes courants comme *manquer* ou *rester*. D'autre part, nous n'avons écrit aucun patron pour repérer les formes impersonnelles avec un verbe au passif ou au se-moyen relevant d'un niveau de langue châtié, voir section 2.1, d'où des erreurs comme en (12).

(12) Il [ANA] s'était formé un cercle d'inimitié autour de cet individu abject …

Les trois premiers types d'erreurs sont évitables à faible coût, mais il n'en est pas de même pour le quatrième type.

4.4 Autres erreurs : 0,2%

Les autres erreurs proviennent de ce que le mot *il* n'est pas employé comme pronom sujet, mais par exemple comme élément d'un nom propre étranger, (13a)[7]. Il y a aussi des fautes de frappe/d'orthographe, (13b) et (13c). Ces erreurs sont imparables en partant d'un texte brut.

(13)a Cela a commencé dans la seconde moitié du 18ème, quand, à Milan, se publie cette revue illuministe appelée Il [ANA] Caffè
b Il [ANA] y vingt-cinq ans
c Puis il [ANA] ont franchi les obstacles dans les bois

---

[6] On peut avancer l'hypothèse d'une corrélation entre la possibilité pour un adjectif d'ancrer une construction impersonnelle et celle d'avoir un sujet abstrait comme *action* ou *idée*, voir la paire *Il est abracadabrantesque de penser que Fred va voter non* et *Cette idée est abracadabrantesque*.

[7] Si ILIMP prenait en entrée un texte où les entités nommées sont reconnues, l'erreur en (13a) serait évitée puisque la séquence *Il Caffè* serait reconnue comme une entité nommée. Dans le présent article, le mot *il* appartient souvent à la méta-langue et n'est alors pas employé comme pronom sujet.

4.5 Evaluation sur des corpus de genre différent

Nous avons aussi réalisé une évaluation d'ILIMP sur des textes littéraires du XIXème siècle concernant 1858 occurrences de *il*. Le taux de bons résultats baisse par rapport au genre journalistique : il passe de 97,5% de bons résultats à 96,8%. Cette baisse est due d'une part à des tournures impersonnelles qui ne sont plus usitées, voir (14), d'autre part à un nombre élevé de phrases impersonnelles avec inversion du sujet comme en (9b), cas que nous n'avons pas pris le temps de traiter systématiquement.

(14) Mais peut-être était-il un peu matin pour organiser un concert, …

Le pourcentage de *il* impersonnel dans ces textes littéraires augmente par rapport au corpus *Le Monde* : il passe de 42% à 49,8%. D'une manière générale, nous nous attendons à des différences importantes de pourcentage de *il* impersonnel selon les corpus[8], mais nous ne nous attendons pas à des différences significatives sur le taux de bons résultats d'ILIMP (surtout si nous prenons le temps de corriger les trois premiers types d'erreurs signalés dans la section 4.2) : nous pensons en effet que les têtes lexicales des constructions impersonnelles forment une liste *fermée* (voir section 4.2) et *stable* quel que soit le corpus.

5 Conclusion et recherche future

La méthode employée dans ILIMP pour repérer les occurrences de *il* impersonnelles ou anaphoriques, qui donnent des résultats satisfaisants, peut être utilisée pour d'autres langues et pour d'autres tâches. Nous avons déjà signalé (note 2) qu'on peut dériver d'ILIMP un outil repérant la tête lexicale d'une phrase simple. On peut aussi envisager d'utiliser ILIMP pour enrichir le calcul des fonctions syntaxiques en identifiant les "sujets profonds extraposés". Par ailleurs, la méthode peut être utilisée pour désambiguïser des mots aussi fréquents et ambigus que *il*. Ainsi, (Jacques, 2005) utilise une méthode similaire à la nôtre dans la première étape du traitement qu'elle propose pour désambiguïser le mot *que* : sans même utiliser la richesse du lexique-grammaire du LADL, elle augmente le taux de précision sur l'étiquetage de ce mot de 14% par rapport à Tree Tagger (passage de 75% de bons résultats à 89%).

On peut s'interroger sur la pertinence de ces "petits" outils qui se cantonnent sur un seul mot ou sur une fonctionnalité bien particulière. Ils sont certes bien modestes par rapport à un système qui produirait, pour n'importe quelle phrase donnée en entrée, une (et une seule) analyse syntaxique, complète, et ce, avec un taux de précision avoisinant les 98%. Mais force est de constater qu'un tel système n'existe pas encore. Nous avancerons donc pour la défense des petits outils le célèbre dicton : *Les petits ruisseaux font les grandes rivières*. Il reste à canaliser ces ruisseaux, i.e. à organiser un vaste effort de recherche pour déterminer comment ordonnancer les "petits" outils dans une chaîne de traitement aboutissant à une analyse syntaxique robuste, complète et correcte.

---

[8] Le quotidien *Le Monde* contient un certain nombre de longs articles racontant la vie et l'œuvre de personnes célèbres. Ces articles, quand ils concernent une personne célèbre de sexe masculin, enchaînent de nombreuses occurrences de *il* anaphorique référant à cet homme. On peut s'attendre à ce que le pourcentage de *il* impersonnel augmente dans les journaux traitant seulement d'actualité ou d'économie.

Remerciements

Je remercie Eric Laporte qui m'a fourni toute la logistique nécessaire pour réaliser ce travail à l'Université Marne-la-Vallée, et Olivier Blanc pour l'assistance technique vitale qu'il m'a apportée. Je remercie aussi Sylvain Kahane et Alexis Nasr pour leurs commentaires fructueux sur cet article.

Références

BGL : BOONS J-P., GUILLET A., LECLERE C. (1976a), *La structure des phrases simples en français: constructions intransitives*. Genève: Droz, 378 p. BGL

BGL : BOONS J-P., GUILLET A., LECLERE C. (1976b), La structure des phrases simples en français: classes de constructions transitives. Rapport de Recherches du LADL n° 6, Paris: Université Paris 7.

COURTOIS B. (2004), Dictionnaires électroniques DELAF anglais et français, *Syntax, Lexis and Lexicon-Grammar. Papers in honour of Maurice Gross* , Lingvisticæ Investigationes Supplementa 24, Amsterdam/Philadelphia : Benjamins, pp. 113–133.

DANLOS L. (1980), *Représentation d'informations linguistiques: les constructions N être Prép X.* Thèse de troisième cycle, Paris: Université Paris 7.

DANLOS L. (1992), Support Verb Constructions: linguistic properties, representation, translation, *Journal of French Linguistic Studies*, Vol. 2, n°1, Cambridge University Press, Cambridge.

EVANS R. (2001), Applying Machine Learning toward an Automatic Classification of *it*, *Literary and Linguistic Computing*, Vol. 16, n°1, pp. 45-57.

GERDES K., KAHANE S. (2004), L'amas verbal au cœur d'une modélisation topologique du français, *Journées de la syntaxe : ordre des mots dans la phrase française, positions et topologie,* Bordeaux, 8.

GROSS M. (1975), *Méthode en syntaxe*, Paris, Hermann.

GROSS M. (1993), Les phrases figées en français, *L'information grammaticale 59*, Paris, p. 36-41.

GROSS M. (1994), Constructing Lexicon-Grammars, *Computational Approaches to the Lexicon*, Oxford, Oxford University Press, p. 213-263.

JACQUES M.P. (2005), *Que* : la valse des étiquettes, *Actes de TALN 05*, Dourdan.

KENNEDY C., BOGURAEV B. (1996), Anaphora for Eveyone; Pronominal Anaphora Resolution without a Parser, in *COLING'96*, Copenhagen.

LAPIN S., LEASS H.J. (1994), An algorithm for pronominal anaphora resolution, *Computational Linguistics*, 20(4), p. 535-561.

LECLERE . C. (2003), The lexicon-grammar of French verbs: a syntactic database, In *Proceedings of the First International Conference on Linguistic Informatics,* Kawaguchi Y. et alii (eds.), UBLI, Tokyo University of Foreign Studies.

MEUNIER A. (1981), *Nominalisations d'adjectifs par verbes supports*. Thèse de troisième cycle, LADL, Université Paris 7.

NASR A. (2004), *Analyse syntaxique probabiliste pour grammaires de dépendances extraites automatiquement,* Habilitation à diriger des recherches, Université Paris 7

PICABIA L. (1978), *Les constructions adjectivales en français*, Genève: Droz.

SILBERZTEIN M. (1994), INTEX: a corpus processing system, in *COLING'94*, Kyoto, Japon, vol. 1, pp. 579-583.

Que : la valse des étiquettes

Marie-Paule Jacques

ERSS – Université Toulouse II le Mirail
5, allées Antonio Machado 31058 Toulouse Cedex 9
mpjacques@univ-tlse2.fr

Mots-clés : Analyse syntaxique automatique, étiquetage morphosyntaxique
Keywords: Automatic parsing, tagging

Résumé Nous présentons ici une stratégie d'étiquetage et d'analyse syntaxique de *que*. Cette forme est en effet susceptible d'appartenir à trois catégories différentes et d'avoir de multiples emplois pour chacune de ces catégories. Notre objectif est aussi bien d'en assurer un étiquetage correct que d'annoter les relations de dépendance que *que* entretient avec les autres mots de la phrase. Les deux étapes de l'analyse mobilisent des ressources différentes.

Abstract In this paper I present a method for tagging and parsing the grammatical word *que*. This word is particularly difficult to tag because it may belong to three different categories and may give rise to many constructions for each category. My aim is to assign the correct tag and to annotate dependency relations between *que* and the other words of the sentence.

1 Introduction

Le mot *que* est de ces formes redoutables pour le TAL français. Non seulement peut-il recevoir plusieurs étiquettes différentes selon ses emplois (voir section 2.1), mais encore entre-t-il dans quantité de constructions syntaxiques différentes, qui correspondent à autant de valeurs sémantiques et d'instaurations de dépendances différentes entre les éléments de la phrase. Or, il est bien souvent nécessaire de résoudre les questions posées par la présence d'un *que* dans une phrase pour proposer une analyse syntaxique satisfaisante de celle-ci.

Notre étude poursuit deux objectifs : i. assigner une étiquette morphosyntaxique correcte à *que* ; ii. annoter les relations syntaxiques qu'il entretient avec d'autres éléments de la phrase. Notre propos n'est pas de discuter les avantages de telle ou telle procédure d'étiquetage morphosyntaxique, par règles ou probabiliste (pour un tour d'horizon, Abney, 1996 ; Garside et al., 1997 ; Habert et al., 1997), mais de tester une méthode d'analyse qui **établit l'étiquetage définitif par l'analyse syntaxique.** Alors que bien souvent ces deux tâches se succèdent – étiquetage **puis** analyse syntaxique appuyée sur les résultats du tagger –, nous procédons à un étiquetage définitif de *que* en fonction de l'identification de constructions syntaxiques

caractéristiques de telle ou telle catégorie. Pour tracer un panorama des difficultés d'analyse, un tour d'horizon des diverses valeurs et constructions possibles de *que* s'impose.

2 Polycatégorie et polyfonctionnalité syntaxique

Dans le Trésor de la Langue Française informatisé (TLFi), la description de *que* couvre 14 pages format A4, c'est dire la variété de ses emplois. Selon les cas, il peut s'agir d'un adverbe, d'un pronom ou d'une conjonction de subordination.

2.1 Les diverses catégories de *que*

- ***que* adverbe**

Que se voit classer comme adverbe dans les phrases exclamatives du type :

> ***Que** cela nous semble alors loin !*

En fonction des auteurs, dictionnaires et autres ouvrages de grammaire, il y a désaccord sur la valeur de *que* dans une construction telle que :

> *Les geysers **n'**entrent en activité **que** la nuit et au petit matin.*

Dans cette structure de négation exceptive *ne V que*, *que* est considéré soit comme adverbe, soit comme une conjonction de subordination. Nous avons choisi la première solution, *que* adverbe, qui présente l'avantage d'une cohérence d'ensemble des constructions impliquant le premier morphème de négation *ne* : tous les éléments qui participent à la négation quels qu'ils soient sont étiquetés adverbe[1]. Ainsi, *que* entre dans un paradigme de formes pouvant co-occurrer avec *ne* pour former tous types de négations.

- ***que* pronom relatif**

Comme tous les pronoms relatifs, *que* est susceptible d'introduire une subordonnée dans laquelle il a une fonction syntaxique, généralement celle d'objet direct :

> *Le volcan est connu pour les risques **qu'**il constitue pour les populations avoisinantes.*

Dans certaines constructions clivées, *que* est ce qu'on appelle un relatif sans antécédent :

> *C'est dans les Andes **que** l'on trouve les plus hauts volcans du monde.*

Les clivées dans lesquelles c'est l'objet du verbe de la subordonnée qui est extrait – par exemple *c'est le chocolat que j'aime* – se ramènent au cas de figure évoqué précédemment.

---

[1] Un autre motif de ce choix tient au fait que, dans le corpus CRATER (cf. 3.1) qui nous sert d'étalon pour la mesure des performances de notre module d'analyse, *que* est étiqueté adverbe pour la négation exceptive.

- ***que* conjonction de subordination**

C'est en tant que conjonction de subordination que *que* manifeste la plus grande variété d'emplois. La place nous manque pour être absolument exhaustive, nous indiquons ici les constructions les plus typiques des textes sur lesquels nous avons basé notre inventaire[2].

Dans une part massive de ses emplois, *que* introduit une complétive, c'est-à-dire soit l'objet direct d'un verbe, soit le complément d'un nom ou encore le complément d'un adjectif :

> *Tout le monde est d' accord pour **penser que** le gaz toxique est venu du fond du lac.*
>
> *Les raisons invoquées proviennent **du fait que** la médecine n'est pas une science exacte.*
>
> *Il est **rare qu'**un séisme provoque directement une activité volcanique.*

Tout en continuant à être catégorisé comme conjonction par l'ensemble des grammaires et dictionnaires, *que* entre dans quantité de structures que, par commodité, nous regroupons sous l'appellation de corrélatives. Dans celles-ci, *que* n'introduit pas nécessairement une **phrase** subordonnée, ce qui a pu donner lieu à une analyse en terme d'ellipse (Riegel et al., 1994)[3]. Nous pouvons remarquer que, dans ces emplois, *que* fonctionne nécessairement avec un autre élément, un élément « déclencheur » situé avant lui dans la phrase, pour assurer une mise en relation de deux constituants : celui qui est contigu à (ou qui contient) l'élément déclencheur, celui qui est introduit par *que*. Ces structures corrélatives se caractérisent donc par la présence d'un couple de marqueurs tantôt contigus, tantôt discontinus, comme par exemple *plus/moins... que*, *d'autant plus/d'autant moins... que*, *aussi... que*, *tel... que*, *autre/même... que*, *tant/autant/aussi bien... que*, etc. :

> *Aucun phénomène n'a donné naissance à **autant** de mythes, de symboles, de légendes, de rites ou de superstitions **que** le 'feu de la terre'.*

2.2 Difficultés pour l'analyse

Les difficultés essentielles à surmonter viennent de ce que des constructions analogues sur le plan de la forme ne le sont cependant pas sur la valeur de *que*. Par exemple, nous n'insisterons pas sur la difficulté, bien connue, de distinguer de façon automatique la valeur de pronom relatif de celle de conjonction de subordination :

> *On appelle compétence du courant la **possibilité qu'**il a de transporter des matériaux.*
>
> *J. et M. semblent exclure la **possibilité que** des opportunités demeurent non perçues.*

De la même manière, la valeur d'adverbe ne se laisse pas simplement cerner par la présence de *ne* devant le verbe :

---

[2] Nous laissons de côté diverses collocations qui peuvent sans ambiguïté être étiquetées 'conjonction de subordination', comme *à condition que*, *afin que*, *de sorte que*, *parce que*, etc.

[3] Nous n'avons pas ici la place de discuter du bien-fondé de cette analyse.

> *Cependant on **n'**est plus persuadé **que** cette action rasante soit entièrement responsable des formes de champignons constatées dans les déserts*

Contrairement à ce que l'on pourrait penser, ce n'est pas seulement le morphème *plus* qui fait que l'on n'a pas de négation exceptive, comme on le voit dans la phrase suivante :

> *Le volcan **n'**émet **plus**, par de nombreux points du cratère, **que** de la vapeur d'eau surchauffée*

Autre exemple, la présence d'un marqueur lexical caractéristique d'une corrélative n'est pas un gage d'identification certaine de la valeur de *que*, comme le montrent les deux extraits suivants, dans lesquels, au sein d'une même construction *plus/moins difficile à Vinf que...*, *que* assume une fonction d'opérateur de comparaison aussi bien que d'objet direct du verbe :

> *Une solution française semble dans ce cas **plus difficile à envisager qu'**une reprise par un groupe étranger.*

> *Il n'en reste pas **moins difficile à expliquer que** les crêtes et sillons prélittoraux ne soient pas toujours parallèles à la côte.*

Face à la diversité des constructions et des éléments à prendre en compte pour les identifier, nous avons testé une stratégie qui intervient à deux moments de l'analyse syntaxique et mobilise des informations différentes dans chacun de ces deux moments. Nous indiquons maintenant le cadre de notre expérimentation, la teneur de la méthode et les résultats obtenus.

3 L'analyse de *que* : cadre, méthode et résultat

3.1 Cadre de l'expérimentation

L'expérimentation que nous décrivons prend place dans une analyse syntaxique de type modulaire, celle mise en œuvre par Syntex (voir ici même Bourigault, Frérot, 2005). A partir d'un texte étiqueté par TreeTagger (Schmid, 1994), divers modules se succèdent, chacun effectuant une partie de l'analyse syntaxique, ce qui autorise les modules les plus tardifs à utiliser les relations posées par les modules précédents pour leur propre analyse.

Syntex fournit une annotation de relations de dépendance entre mots. Par exemple, d'un nom pourront dépendre un déterminant, un ou plusieurs adjectifs, une ou plusieurs prépositions… ; d'une préposition pourront dépendre un nom ou un verbe ; d'un verbe pourront dépendre un nom et/ou un pronom en fonction sujet, idem en fonction objet, une ou plusieurs prépositions, etc. Chaque mot de la phrase est donc susceptible d'entrer dans deux types de dépendances – être régi par un autre mot, être recteur d'un autre mot –, mais qui ne sont pas toutes deux obligatoires : par exemple, le verbe d'une principale n'est pas régi.

Pour ce qui concerne *que*, TreeTagger lui attribue l'une des trois étiquettes 'Pronom relatif', 'Adverbe' ou 'Conjonction de subordination'. Notre analyse a pour objet : i. de revenir sur cet étiquetage pour éventuellement le rectifier, ii. de relier *que* aux éléments convenables dans les deux directions, c'est-à-dire vers un recteur et vers un régi (par exemple, le verbe de la subordonnée), et ce par la relation idoine.

Afin de déterminer les indices sur lesquels appuyer l'analyse et d'avoir un aperçu réaliste de la diversité des constructions, nous avons constitué un corpus d'étude en rassemblant, de façon opportuniste, la plus grande variété de textes dont nous pouvions disposer sous format électronique : articles du journal Le Monde, articles scientifiques du domaine de l'ingénierie des connaissances, recettes de cuisine, documents professionnels, textes littéraires, textes spécialisés de domaines aussi divers que la volcanologie, la géomorphologie, la médecine, le vol libre. Il s'agissait pour nous de ne pas nous cantonner à un seul genre, sans toutefois prétendre couvrir tous les genres possibles. Nous pensons tout de même avoir de cette manière recueilli une bonne gamme des possibilités d'emploi de *que*.

L'évaluation n'a pas été menée sur ce corpus, mais sur un corpus de test extrait de la partie française du corpus CRATER[4]. Celui-ci offre l'intérêt d'un étiquetage morphosyntaxique – relativement – fiable parce que vérifié par des annotateurs humains. Il constitue donc un banc d'essai idéal pour mesurer l'efficacité de notre analyse sur l'étiquetage. Nous avons extrait de ce corpus toutes les phrases contenant la forme *que*, puis nous avons annoté manuellement les 1100 premières pour les relations de dépendance.

3.2 Méthode d'analyse

L'analyse de *que* se fait en deux étapes principales intervenant à des moments différents dans le processus global. Son principe essentiel est d'exploiter les acquis des modules précédents pour repérer certaines constructions syntaxiques modélisées à partir du travail sur corpus et décider, en fonction de celles-ci, de la catégorie définitive de *que*. En résumé, **ce n'est que lorsqu'une construction syntaxique est positivement identifiée que *que* est réétiqueté.**

Dans un premier temps, sont repérées certaines constructions qui ne requièrent pour être identifiées que peu d'informations de structure. Puis, en toute fin d'analyse, un module spécifique repère les constructions faisant intervenir des relations à plus grande distance.

3.2.1 Première étape : des constructions locales

Rappelons que Syntex produit une analyse à partir d'un texte préalablement étiqueté. Cela implique que la qualité de l'analyse syntaxique dépend crucialement de la qualité de l'étiquetage. Par exemple, si un *que* adverbe est faussement étiqueté conjonction de subordination, le module de recherche des objets directs est mis en échec : une conjonction de subordination constitue une barrière après laquelle on ne peut avoir un objet direct. Pour éviter ces obstacles liés à des erreurs d'étiquetage, il est apparu nécessaire d'avoir le plus tôt possible dans l'analyse un processus de vérification et de correction de *que*. Mais, « très tôt dans l'analyse » signifie que l'on ne peut s'appuyer que sur très peu d'informations de structure, puisque celle-ci ne sera découverte que par les modules suivants et précisément qu'à condition qu'un mauvais étiquetage ne rende pas la tâche impossible.

Pour l'essentiel, dans cette première étape, sont donc analysés et réétiquetés des *que* dont le contexte proche fournit une information exploitable : des *que* adverbes placés immédiatement

---

[4] Université Lancaster. *UCREL Projects* [en ligne]. http://www.comp.lancs.ac.uk/ucrel/projects.html#crater (page consultée le 4 février 2005)

après le verbe, des *que* objets situés eux aussi à proximité du verbe, des *que* introduisant un complément de nom (*le fait que, l'idée que, l'hypothèse que, …*), des *que* pris dans une structure comparative très locale du type *plus/moins/aussi/autant/si Adj que*. L'objectif lors de cette phase est de découvrir le plus possible de *que* adverbe ou conjonction de subordination en opérant, grâce au contexte, une « déduction affirmative » (Vergne, Giguet, 1998) sur la catégorie de *que*. Dans le même temps, et parce que cette déduction est liée au repérage d'un type de construction syntaxique, sont annotées les relations de *que* avec son recteur. Pour ce qui est des relations avec les régis, seules sont annotées celles des complétives car c'est le seul cas dans lequel on est sûr que l'élément régi soit le verbe conjugué d'une proposition subordonnée. En effet, dans une structure corrélative, *que* n'introduit pas nécessairement une proposition (*davantage* ***que je ne croyais*** *vs. davantage* ***que l'autre jour***).

Les informations mobilisées mêlent des listes lexicales – une vingtaine de noms qui prennent un complément en *que*, une dizaine d'adverbes susceptibles d'entrer dans une structure corrélative mettant en jeu un adjectif ou un participe passé, moins d'une dizaine d'adverbes de négation (*pas, point, guère*, etc.), deux cents verbes qui prennent un objet direct en *que* (*penser que, croire que*, etc.) –, l'exploitation des quelques relations qui ont été posées – rattachement des adverbes autour des verbes et des adjectifs et rattachement des auxiliaires et des modaux aux participes passés pour les premiers, à un verbe à l'infinitif pour les seconds – et enfin, la prise en compte de la catégorie des mots avant et après *que*. Rappelons que la démarche consiste essentiellement à repérer des éléments positifs d'analyse dans le contexte proche (nous donnerons en 3.2.3 un exemple de règle d'analyse). L. Danlos (2005) adopte pour la désambiguïsation de *il* (pronom impersonnel *vs.* pronom anaphorique) une démarche très similaire – modélisation de patrons linguistiques qui constituent les 'marqueurs' d'une valeur donnée – et elle obtient avec cette approche 97,5% de bons résultats, ce qui est plus que satisfaisant.

3.2.2 *Deuxième étape : la structure globale*

Au moment de cette seconde étape, tous les modules ont fait leur travail et ont placé diverses relations qui vont être exploitées pour l'analyse de *que*. Le principe est de remonter vers la gauche de *que* à la recherche d'indices de telle ou telle structure. En fonction de la nature des constituants rencontrés, certains éléments sont recherchés. Par exemple, si dans ce parcours vers la gauche un nom est rencontré, on vérifie si ce nom régit un adjectif tel que *même* ou *autre*, si oui, on estime être face à une structure corrélative et l'analyse s'arrête là, si non, on se déplace sur le recteur du nom, on teste à nouveau une série de contraintes et ainsi de suite. Pour chaque nouveau constituant, une série de contraintes est évaluée et dès qu'une contrainte est satisfaite, l'analyse s'arrête. De cette façon, dans les deux exemples suivants, extraits de CRATER, on analyse successivement divers syntagmes prépositionnels jusqu'à arriver aux éléments qui permettent d'identifier deux structures corrélatives différentes.

> *L'indication de fonction peut ou non contenir le* ***même*** *numéro d'identification de fonction* ***que*** *celui qui se trouvait dans la demande d'activation de fonction d'origine.*
>
> *Les procédures s'appliquent* ***aussi bien*** *à une interface à débit de base* ***qu'***à *une interface à débit primaire.*

Les informations mobilisées sont donc les mêmes que pour la première étape, plus – et essentiellement – toutes les informations de structure disponibles. Ce sont celles-là qui sont essentielles car, le principe étant de progresser vers la gauche constituant par constituant, **en**

s'appuyant sur les relations que les modules précédents ont annotées, l'analyse s'interrompt dès qu'un constituant de la chaîne est « orphelin », c'est-à-dire n'est rattaché à aucun autre élément.

Cette seconde étape rajoute quelques règles d'analyse, notamment pour l'analyse des syntagmes nominaux et prépositionnels, pour lesquels on ne disposait d'aucune relation de dépendance lors de la première étape, mais aussi elle permet d'appliquer des règles déjà définies à des éléments de la phrase placés non immédiatement dans le contexte de *que*. Pour l'illustrer, nous prendrons comme exemple la règle qui s'applique à l'analyse d'un verbe conjugué à gauche de *que*.

3.2.3 Un exemple de règle d'analyse

Il s'agit ici d'une règle utilisée aux deux étapes de l'analyse pour repérer certaines constructions verbales. Elle permet de décider dans un certain nombre de cas d'étiqueter *que* comme conjonction de subordination (CSub) ou comme adverbe (Adv), en repérant les éléments d'une corrélative, d'une relation d'objet direct ou d'une négation exceptive.

Lorsque la forme *que* est immédiatement précédée d'un verbe conjugué ou précédée d'un adverbe lui-même précédé d'un verbe conjugué, alors on lance l'analyse du verbe, qui consiste à tester les conditions énumérées ci-dessous et, en cas de test positif, à attribuer à *que* l'étiquette mentionnée après la flèche.

- présence des adverbes : *plus, moins, davantage, autant, mieux, tellement* → CSub ;
- présence de *ne* immédiatement précédé par *rien, nul, personne* (par exemple *Nul ne comprendrait que*... qui n'est pas une négation exceptive) → CSub ;
- présence conjointe de *ne* et de *pas, point* ou un autre *que* adverbe → CSub ;
- présence seule de *ne* : négation exceptive → Adv ;
- appartenance du verbe à la liste de ceux qui prennent *que* comme objet direct → CSub + recherche du verbe de la subordonnée introduite par *que*.

Il est possible aussi qu'on ne se prononce pas, c'est-à-dire qu'on laisse l'étiquette attribuée par TreeTagger en l'état, si on n'a recueilli aucun indice pour faire mieux que lui.

Dans la seconde étape, la même règle est utilisée pour analyser des verbes placés plus loin de *que*, par exemple :

> *Un « champ électrique » implique une activité qui* ***ne*** *peut être exprimée d'une façon univoque* ***qu'****en fonction du temps et de deux ou trois dimensions.*

En suivant les relations de dépendance placées par l'analyse syntaxique, le module se déplace de l'adjectif *univoque* à son recteur *façon*, de celui-ci à la préposition *de*, de cette dernière au participe passé *exprimée*, puis au verbe *être* et enfin au verbe *pouvoir*, auquel s'applique la règle exposée ci-dessus, qui permet de reconnaître un *que* adverbe.

Il s'agit ici d'une partie de la règle qui couvre la plus grande proportion de *que* : dans le corpus CRATER, les *que* objets directs ou adverbes partie prenante d'une négation représentent 52% des relations vers un recteur. Dans la première étape, la règle permet d'identifier 76% des *que* objets et 61% des adverbes, un pourcentage qui est diversement amélioré dans la seconde étape

puisqu'il passe à 78% pour les objets et 82% pour les adverbes. Ces chiffres ne sont qu'une composante des résultats, intéressons-nous maintenant à ceux-ci.

3.3 Résultats

Dans la mesure où la tâche concerne aussi bien l'étiquetage que l'annotation de relations, nous avons évalué notre méthode sur ces deux points.

3.3.1 L'étiquetage

Pour ce qui concerne l'étiquetage, la seule mesure que nous présentons concerne la précision. En effet, le rappel est de 100% : absolument tous les mots reçoivent une et une seule étiquette. La mesure porte sur 1183 *que*, étiquetés initialement par TreeTagger. Le tableau ci-dessous récapitule les résultats :

Etape	Précision	
étiquetage initial	75%	
étape 1	89%	+14%
étape finale	92%	+3%

Tableau 1 : Mesure de la précision de l'étiquetage morphosyntaxique

On voit que l'essentiel du gain est obtenu après la première étape, ce qui tend à montrer que pour la détermination de la catégorie de *que*, des informations très locales suffisent. Il n'en est pas de même pour l'annotation des relations.

3.3.2 Les relations

Les mesures proposées ici sont restreintes aux cas où l'étiquetage de *que* est correct. Le nombre de relations évaluées n'est donc pas le même à chacune des étapes : 1789 pour la première, 1843 pour la seconde, toutes relations confondues. Il faut noter que pour les *que* adverbes, l'analyse n'inscrit qu'une relation, vers le recteur, alors que pour les *que* conjonction ou pronom relatif, deux relations doivent être annotées, l'une vers le recteur, l'autre vers le régi pour les conjonctions, vers un antécédent nominal pour le pronom. Ceci pose, dans le cas des relatifs sans antécédent, tels que ceux qui participent à une structure clivée (voir la section 2.1), de véritables problèmes de représentation : soit on note une relation d'antécédence fictive et arbitraire vers l'un quelconque des éléments du constituant clivé, soit aucune relation d'antécédence n'est mentionnée (c'est la position retenue actuellement).

Le rappel et la précision de l'annotation des relations sont indiqués dans le tableau 2.

Etape	Rappel	Précision
étape 1	60%	97%
étape finale	93% +33%	94% -3%

Tableau 2 : Mesures de rappel et de précision de l'annotation des relations

A titre de comparaison, la précision de Syntex pour le rattachement des prépositions varie de 78 à 87 % selon les corpus (Bourigault, Frérot, 2005).

Il apparaît ici clairement que l'annotation des relations bénéficie considérablement des informations sur les relations existantes au sein de la phrase, notamment parce qu'il devient alors possible de placer des relations à distance. Mais cette annotation de relations se trouve améliorée aussi pour la simple raison que la recherche des régis des conjonctions de subordination dans les corrélatives n'est pas du tout effectuée dans la première étape, et ce parce que l'on a besoin de disposer des dépendances pour, par exemple, ne pas prendre un pronom pour régi et arriver jusqu'au verbe dont il est sujet, en d'autres termes, ne pas analyser *tel que nous l'avons défini* comme *tel que nous*.

A l'heure actuelle, les performances sont limitées par deux types d'obstacles. D'une part, l'ambiguïté de certaines structures : comme nous l'avons déjà mentionné, la présence des marqueurs recherchés ne garantit pas d'être effectivement face à la structure supposée, ce qui est parfaitement illustré par les deux exemples suivants, formellement analogues. Dans le premier, la présence de *tels* conduit l'analyseur à manquer la relation avec le verbe *dire*.

> *On dit couramment de **tels** réseaux mésochrones **qu'**ils sont synchronisés.*
>
> *La nature sociale comporte de **tels** hasards **que** l'imagination des inventeurs est à tout moment dépassée.*

D'autre part, les limites du fichier d'entrée constituent des écueils actuellement infranchissables : si par exemple un verbe n'est pas étiqueté verbe mais nom, si une préposition n'a pas été rattachée ou encore si un adjectif ne régit pas l'adverbe qui manifesterait une comparaison, le processus d'annotation des relations est bloqué. Ce sont donc deux directions vers lesquelles porter les efforts.

Avant de conclure, soulignons que ces résultats sont partiels dans la mesure où ils se limitent à l'analyse de la forme *que*. En fait, il faudrait pouvoir évaluer l'impact d'un meilleur traitement de *que* sur l'analyse globale, ce que nous n'avons pas été en mesure de faire.

4 Conclusion

Nous avons présenté une stratégie qui effectue dans le même mouvement l'analyse et l'étiquetage de *que*, en prenant appui sur les acquis de l'analyse syntaxique au fur et à mesure de son déroulement. La confrontation à un corpus de référence montre que l'étiquetage est grandement amélioré avec peu d'informations de structure, qui sont en revanche indispensables pour l'annotation des relations de dépendance.

Parmi les pistes à continuer à explorer, nous voudrions mentionner la question de la représentation des relations. La représentation de certaines relations ne nous semble pas poser problème : les complétives, les objets directs, les adverbes sont reliés sans difficulté au nom, adjectif, verbe, etc., concerné. Mais les structures corrélatives, par leur diversité, ne se laissent pas facilement appréhender. A l'heure actuelle, nos choix d'annotation impliquent de représenter de la même façon des structures similaires en surface. Considérons par exemple une structure *aussi Adj que…* Il nous importe de reconnaître une corrélative, en identifiant la relation entre l'adjectif et l'adverbe *aussi*, quel que soit l'autre élément de la comparaison. Mais selon la nature de cet élément, **l'interprétation** de la structure n'est pas la même :

Jean est aussi gentil que beau	Jean gentil + beau
Jean est aussi gentil que son frère	Jean gentil + son frère gentil

Dans le premier cas, *que* régit un adjectif, l'interprétation doit être que les deux adjectifs se rapportent au même SN, dans le second cas, *que* régit un SN, on doit alors interpréter que les deux SN « partagent » l'adjectif. Il nous semble qu'il n'est absolument pas trivial de décider comment distinguer ces deux constructions, soit elles sont distinguées au moment de l'annotation, par des relations différenciées, soit on estime que l'annotation doit rester au plus près des structures de surface et que c'est dans l'interprétation qu'elles seront différenciées. Cela mérite réflexion et approfondissement, au-delà du cas particulier de *que*.

Remerciements

Je remercie vivement Cécile Fabre, Didier Bourigault ainsi que les relecteurs de TALN pour toutes leurs remarques et conseils.

Références

ABNEY S. (1996), Part-of-Speech Tagging and Partial Parsing, In K. Church, S. Young et G. Bloothooft (Eds.), *Corpus-Based Methods in Language and Speech*, pp. 118-136, Dordrecht, Kluwer Academic Publishers.

BOURIGAULT D., FREROT C. (2005), Acquisition et évaluation sur corpus de propriétés de sous-catégorisation syntaxique, Actes de *TALN'2005*.

DANLOS L. (2005), ILIMP : Outil pour repérer les occurrences du pronom impersonnel *il*, Actes de *TALN'2005*.

GARSIDE R., LEECH G., MCENERY T. (1997), *Corpus Annotation: Linguistic Information from Computer*, Londres, Longman.

HABERT B., NAZARENKO A., SALEM A. (1997), *Les linguistiques de corpus*, Paris, Armand Colin.

RIEGEL M., PELLAT J.-C., RIOUL R. (1994), *Grammaire méthodique du français*, Paris, P.U.F.

SCHMID H. (1994), Probabilistic Part-of-Speech Tagging Using Decision Trees, Actes de *NEMLAP*.

VERGNE J., GIGUET E. (1998), Regards Théoriques sur le "Tagging", Actes de *TALN'98*, 22-31.

Un système multi-agent pour la détection et la correction des erreurs cachées en langue arabe

(1) Chiraz Ben Othmane Zribi, (2) Fériel Ben Fraj, (3) Mohamed Ben Ahmed

Laboratoire RIADI – Université La Manouba
ENSI, La Manouba, Tunisie
(1) Chiraz.benothmane@riadi.rnu.tn, (2)Feriel.BenFraj@riadi.rnu.tn,
(3)Mohamed.BenAhmed@riadi.rnu.tn

Mots-clés : Erreurs orthographiques cachées, Détection, Correction, Système multi-agent, Analyse linguistique, Langue arabe

Keywords: Hidden spelling errors, Detection, Correction, Multi-Agent System, Linguistic analysis, Arabic language

Résumé

Cet article s'intéresse au problème des erreurs orthographiques produisant des mots lexicalement corrects dans des textes en langue arabe. Après la description de l'influence des spécificités de la langue arabe sur l'augmentation du risque de commettre ces fautes cachées, nous proposons une classification hiérarchique de ces erreurs en deux grandes catégories ; à savoir syntaxique et sémantique. Nous présentons, également, l'architecture multi-agent que nous avons adoptée pour la détection et la correction des erreurs cachées en textes arabes. Nous examinons alors, les comportements sociaux des agents au sein de leurs organisations respectives et de leur environnement. Nous exposons vers la fin la mise en place et l'évaluation du système réalisé.

Abstract

In this paper, we address the problem of detecting and correcting hidden spelling errors in Arabic texts. Hidden spelling errors are morphologically valid words and therefore they cannot be detected or corrected by conventional spell checking programs. In the work presented here, we investigate this kind of errors as they relate to the Arabic language. We start by proposing a classification of these errors in two main categories: syntactic and semantic, then we present our multi-agent system for hidden spelling errors detection and correction. The multi-agent architecture is justified by the need for collaboration, parallelism and competition, in addition to the need for information exchange between the different analysis phases. Finally, we describe the testing framework used to evaluate the system implemented.

1 Introduction

Le problème des erreurs cachées présente une incommodité pour les scripteurs lors de la saisie de leurs textes. Ces fautes ne sont autres que des erreurs orthographiques qui produisent pour autant des mots lexicalement corrects. La présence d'un vocable au sein d'un contexte syntaxique ou sémantique qui n'est pas le sien peut rendre insensée toute la phrase. L'exemple suivant illustre ce phénomène :

Exemple : Le jardinier utilise le *gâteau* pour bêcher la terre
Le mot *"gâteau"* est introduit dans un contexte qui ne lui est pas approprié. Cette faute de frappe peut être corrigée en la changeant par le vocable *"râteau"*.

Les statistiques réalisées par Eastman et Oakman (1991) (Verberne, 2002), affirment que les erreurs cachées comptent 25% parmi toutes les erreurs orthographiques commises et contenues dans leur corpus de référence. Mitton (1987) leur attribue une valeur plus grande à savoir : 40% parmi toutes les erreurs orthographiques étudiées. Ces deux valeurs assez importantes ont rendu l'étude de ce genre d'erreurs une nécessité en soi. En effet, plusieurs recherches ont été entreprises dans le but de remédier à ce problème. Nous pouvons, alors, citer comme exemples : les recherches de Golding (Golding, Dan, 1996) qui a étudié ce genre d'erreurs en langue anglaise. Il a ainsi proposé de multiples méthodes comme la méthode de Bayes (Golding, 1995), la méthode des trigrammes des parties du discours (Golding, Schabes, 1996) et la méthode à base de réseaux neuronaux dite Winnow (Golding, Dan, 1999). Le chinois a été aussi traité par les deux chercheurs Jianhua et Xiaolong (Xiaolong, Jianhua, 2001). Le suédois a fait l'objet d'une recherche pareille avec Bigert et Knutsson (Bigert, Knutsson, 2002).

En ce qui concerne la langue arabe, aucun travail n'a été réalisé sur les erreurs cachées malgré l'importance de l'entreprise d'une telle recherche. La langue arabe présente, en effet, des spécificités qui rendent le risque de commettre une erreur cachée plus important que pour les autres langues. Nous nous sommes donc proposés de nous intéresser à ce problème en construisant un système permettant à la fois de détecter et de corriger ce type d'erreurs pouvant survenir dans des textes arabes. A cause de la complexité de ce travail, nous avons été amenés à émettre certaines hypothèses pour restreindre les champs de nos investigations. Nous avons considéré alors l'arabe non voyellé avec une seule erreur cachée par phrase. L'erreur consiste en une seule faute d'édition à savoir ; l'ajout d'un caractère, l'omission d'un caractère, la substitution d'un caractère par un autre ou l'interversion de deux caractères adjacents. Des statistiques ont en effet montré que l'une (seulement) de ces opérations est à l'origine d'une erreur orthographique dans 90% des cas (Ben Hamadou, 1993).

Dans ce qui suit, nous décrivons dans la première section les spécificités de la langue arabe qui ont participé à augmenter le risque de commettre des erreurs cachées lors de la saisie des textes. Dans la seconde section, nous présentons la classification que nous avons adoptée pour ces erreurs. Sur la base de ladite classification et de nos besoins, nous avons opté pour une architecture multi-agent dont la conception est décrite dans la troisième section de l'article. La quatrième section est consacrée, quant à elle, à la description des résultats de l'évaluation du système mis en place.

2 Quelles difficultés pour la langue arabe ?

Pour la langue arabe, le problème des erreurs cachées s'aggrave et se complique encore plus que pour d'autres langues (notamment indo-européennes). En effet, on se trouve confronté dans cette langue, à des contraintes d'écriture et à diverses ambiguïtés : telles que l'agglutination des enclinomènes aux formes simples, l'ambiguïté grammaticale, et la proximité lexicale des formes textuelles.

2.1 Le phénomène d'agglutination

L'agglutination est l'ajout de préfixes appelés *'proclitiques'* et de suffixes dits *'enclitiques'* aux formes simples pour obtenir ce qu'on a appelle des "formes agglutinantes" ou encore des "hyperformes". Les proclitiques et les enclitiques forment l'ensemble des enclinomènes[1] de la langue arabe. Une erreur cachée peut être la conséquence d'une opération d'ajout ou d'omission d'un enclinomène à un radical. Cette opération peut donner naissance à une forme textuelle licite, mais malencontreusement, injectée dans un contexte qui ne lui est pas approprié.

2.2 L'ambiguïté grammaticale

Les mots en arabe présentent une ambiguïté grammaticale. Les statistiques réalisées par (Debili et al., 2002) sur un texte en langue arabe confirment cette ambiguïté. L'auteur a constaté l'importance du taux d'ambiguïté grammaticale pour les textes voyellés qui est égale à 5,63 en moyenne. Ce taux est augmenté par l'absence des voyelles pour atteindre en moyenne 8,71 catégories grammaticales par forme textuelle. L'ambiguïté grammaticale est l'une des causes de l'abondance des erreurs cachées en langue arabe car ce genre de fautes peut être dû à une confusion dans l'interprétation grammaticale des formes textuelles.

2.3 La proximité lexicale

La caractéristique des mots arabes la plus remarquable pour notre problématique est celle de la proximité lexicale. En effet, les mots en langue arabe sont très proches graphiquement les uns des autres, nous pouvons l'affirmer en nous basant sur l'étude réalisée par (Ben Othmane Zribi, 1998) à ce propos. Cette expérience consistait à appliquer les quatre opérations d'édition, précédemment citées, sur tous les mots d'un dictionnaire de la langue. Parmi les formes, automatiquement construites, on a procédé par dénombrer les graphies correctes, comptant ainsi ce qu'on appelle *"le nombre de mots approchants ou lexicalement voisins"*. Ces comptages nous ont donné une idée claire sur la similarité dite aussi le degré de ressemblance entre les vocables d'une langue. Ainsi, avons-nous constaté que les mots en langue arabe sont beaucoup plus proches les uns des autres avec un nombre moyen de formes voisines de 26,5 : valeur importante comparée à celle calculée pour la langue française égale à 3,5 et celle relative à l'anglais égale à 3. De ce fait, la ressemblance typographique des mots

---

[1] Ce sont des particules affixes représentant les pronoms personnels, les conjonctions, les prépositions, etc.

en arabe augmente le risque de tomber sur des erreurs cachées, comme elle augmente la taille de la liste des candidats à la correction lors de la phase de correction.

3 Typologie des erreurs cachées

Pour détecter les erreurs cachées, une simple analyse morphologique s'avère insuffisante puisque ces erreurs engendrent des formes morphologiquement correctes mais erronées sur le plan syntaxique ou sémantique (voir même pragmatique). Ainsi, une phrase contenant une erreur syntaxique est lexicalement correcte mais la structuration de ses mots est incorrecte. On parle alors d'un dérèglement syntaxique. Par contre, une phrase contenant une erreur sémantique est dénuée de sens puisque l'erreur est un mot venant s'intercaler dans un contexte sémantique qui n'est pas le sien.

3.1 Les anomalies syntaxiques

Les dérèglements grammaticaux peuvent être de différents types. Ainsi, la classe des anomalies grammaticale peut être subdivisée en des sous-classes d'erreurs syntaxiques qui se présentent comme suit :

- Les erreurs d'accord : Ce sont des dérèglements syntaxiques qui sont dus au non-respect des contraintes d'accord qui gèrent la langue. Ils causent des incompatibilités au niveau des informations morpho-syntaxiques relatives aux mots d'une même phrase. Exemple : (الكبيرة) "رفع المسافر الحقيبة الكبير", ("Le voyageur a soulevé le grand valise », la forme correcte est : la grande).

- Les erreurs liées à la transitivité : La transitivité est un lien syntaxique (et sémantique) entre un verbe et un complément d'objet lui succédant en général. L'absence de ce complément ou son apparition là où il ne faut pas rend incorrecte la phrase. Exemple : (بالحمّى) "مرض الرضيع الحمى" ("le nourrisson est tombé malade fièvre", la forme correcte est : à cause de la fièvre).

- Les erreurs d'agrammaticalité : Ce genre de fautes concerne l'agencement des catégories grammaticales au sein de la phrase. Exemple: (مكتبه) "جلس المدير في نكتبه" ("le directeur s'est assis dans nous l'écrivons", la forme correcte est : son bureau).

3.2 Les anomalies syntaxiques

Tout comme pour le niveau syntaxique, les anomalies sémantiques peuvent être scindées en des sous classes d'erreurs dont nous pouvons citer :

- Les incompatibilités sémantiques : Une incompatibilité sémantique consiste en l'injection d'un mot dans un contexte sémantique qui n'est pas le sien. Exemple : (سمكة) "وجد الصياد سكة كبيرة" ("Le pécheur a trouvé une grande voie", la forme correcte est : poisson ; en langue arabe le mot poisson est au féminin)

- Les incomplétudes sémantiques : L'omission ou l'insertion là où il ne faut pas d'une conjonction de coordination ou toute autre particule rendent parfois la phrase dénuée de sens (Aloulou, 1996). Exemple : "ضربت الولد بكى (فبكى) " ("j'ai frappé le garçon il a pleuré", la forme correcte est : ensuite il a pleuré).

4 La solution proposée

La complexité de notre problème ainsi constatée, ainsi que la hiérarchie présentée dans la classification des erreurs cachées dénotent la nécessité d'une interférence entre les différentes phases d'analyse pour le traitement de ce genre d'erreurs. En effet, la détection et la correction des erreurs syntaxiques nécessitent la contribution des connaissances sémantiques. De même que, le traitement des erreurs cachées de type sémantique nécessite des retours en arrière syntaxiques pour une meilleure détection et correction.

En conséquence, nous avons opté pour une architecture multi-agent où les différents agents travaillent en : collaboration, compétition, coordination et parallélisme afin d'atteindre l'objectif global du système. Chacun d'eux apporte sa contribution à la solution finale. Tous s'organisent dans une société commune où ils peuvent discuter et coopérer.

5 L'architecture du système

Pour qu'un système de vérification des textes en langue naturelle soit efficace, il doit avoir à sa disposition un ensemble d'informations linguistiques concernant ces textes. C'est pour cette raison que nous nous sommes proposés d'effectuer une analyse morpho-syntaxique des textes à traiter. Ces textes analysés serviront comme entrées pour notre système. La figure1 illustre l'architecture générale de notre système.

5.1 Le groupe syntaxique d'agents

Ce groupe est formé de quatre agents à savoir ; l'agent Accord, l'agent Transitivité, l'agent Compatibilité syntaxique et l'agent superviseur des trois premiers. Le superviseur syntaxique reçoit alors le texte à vérifier et l'envoie phrase par phrase à ses collègues de la même société d'agents.

- **L'agent Accord** vérifie la validité des contraintes d'accord en appliquant un ensemble de règles d'accord (on compte environ 800 règles).
- **L'agent Transitivité** essaie de détecter les anomalies pouvant exister entre les verbes et leurs compléments d'objet.
- **L'agent Compatibilité syntaxique** vérifie la structuration des Catégories Grammaticales des Hyperformes (HyperCGs) dans la phrase en considérant des séquences ternaires d'HyperCGs. Il utilise à cet effet, une matrice à trois dimensions qui indique la validité de ces séquences ternaires.

Les travaux de ces trois agents sont contrôlés par le superviseur. En effet, une fois que l'un d'eux détecte une anomalie, il informe ses collègues pour qu'ils arrêtent leurs traitements respectifs et annonce la nouvelle au superviseur pour que ce dernier déclenche le processus de correction.

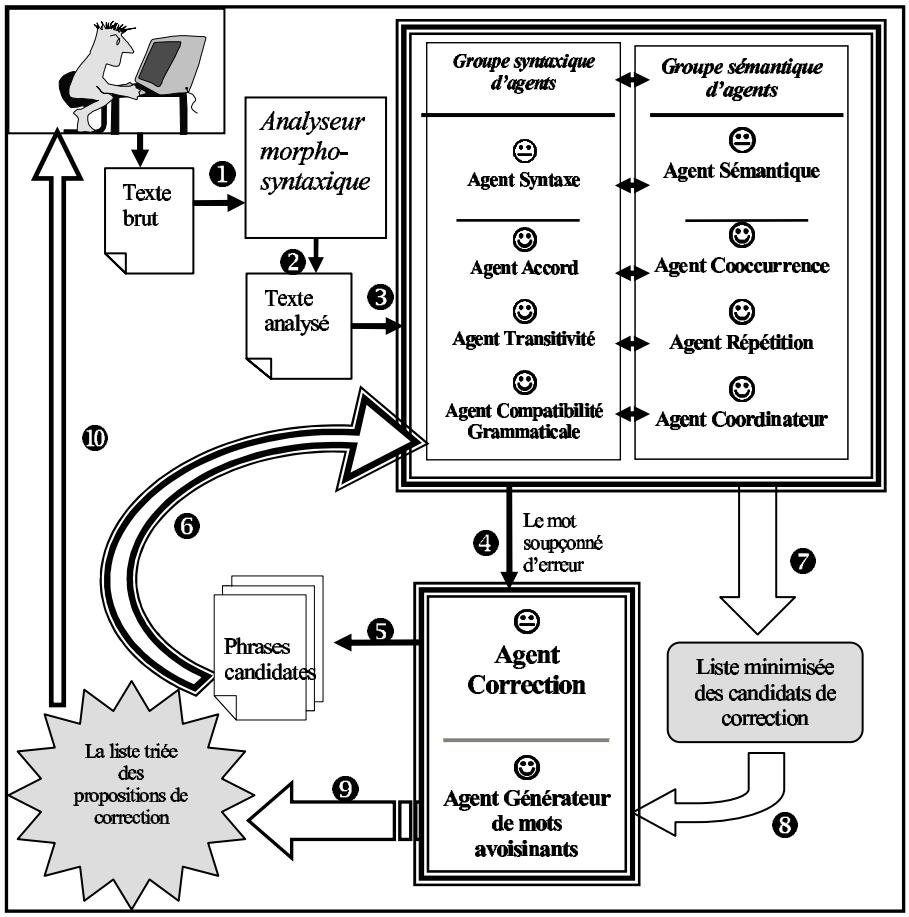


Figure 1 : Architecture du système Multi-Agent de détection/correction des erreurs cachées

5.2 Le groupe sémantique d'agents

Ce groupe est aussi constitué de quatre agents. Le premier est le superviseur qui envoie le texte découpé en phrases aux autres agents du même groupe. Les trois autres sont : l'agent Cooccurrence, l'agent Répétition et l'agent Coordinateur.

Notons :

$Ph = \{m_1, ..., m_i, ..., m_n\}$: la phrase soumise à la vérification sémantique.

$C = \{c_{-k}, ..., c_{-1}, c_1, ..., c_k\}$: l'ensemble des mots entourant le mot analysé (en considérant une fenêtre de taille k).

$L = \{l_1, \ldots, l_i, \ldots, l_n\}$: l'ensemble des lemmes des mots de la phrase.

- **L'agent Cooccurrence** vérifie pour chaque mot de la phrase s'il possède des affinités sémantiques avec son contexte. Il procède alors de deux façons, ne pouvant être qualifiées de différentes mais plutôt de complémentaires. Cet agent va, tout d'abord, chercher s'il existe des collocations[2] entre le mot cible d'analyse et les mots l'avoisinant dans le même contexte. Les collocations, si elles sont trouvées, vont

[2] Une collocation est une association habituelle de deux ou plusieurs termes (collocats) au sein d'un discours.

permettre de conforter chaque mot dans le contexte où il a été mis. Ainsi, pour chaque vocable est attribuée une valeur dite information mutuelle calculée à l'aide de la formule suivante :

$$I(m_i) = \max_{k \leqslant j \leqslant k} \text{Log} \frac{p(m_i, c_j)}{p(m_i) \times p(c_j)}$$

Avec $p(m_i)$ la probabilité d'observer m_i, $p(c_j)$ la probabilité d'observer c_j et $p(m_i,c_j)$ la probabilité de les observer ensemble.

Affirmer que le mot m_i est en collocation avec son contexte revient à évaluer la valeur $I(m_i)$ de la façon suivante :

- si $I(m_i)$ est positive, alors m_i est en collocation avec son contexte.
- si $I(m_i)$ s'approche de la valeur nulle alors m_i n'a pas de rapport avec son contexte.
- si $I(m_i)$ est négative alors m_i a des distributions complémentaires avec son contexte.

Outre les collocations, l'agent Cooccurrence va chercher s'il existe des cooccurrences ordinaires entre chaque mot cible d'analyse et son entourage. Pour ce faire, nous avons choisi d'utiliser la formule des probabilités conditionnelles de Bayes suivante :

$$p(m_i|C) = \frac{p(C|m_i) \times p(m_i)}{p(C)}$$

Plus la valeur $p(m_i/C)$ est élevée, plus le mot analysé m_i a d'affinité sémantique au sein de son contexte C.

- **L'agent Répétition**, pour sa part, cherche si le lemme de la forme textuelle à analyser se répète au sein du texte lui même. Il se base sur le principe qui dit que *'les mots ou plus précisément les lemmes des mots d'un texte ont tendance à se répéter dans le texte lui-même'*. En effet, d'après des comptages réalisés (Ben Othmane Zribi, Ben Ahmed, 2003) sur un corpus textuel en langue arabe appartenant à un domaine bien particulier, nous savons qu'une forme textuelle peut apparaître en moyenne 5,6 fois alors qu'un lemme peut apparaître en moyenne 6,3 fois et ce dans le même texte. Nous calculons alors pour chaque lemme sa fréquence d'occurrence au sein de tout le texte, avec la formule suivante :

$$p(l_i) = \frac{\text{nombre d'occurrences de } l_i}{\text{nombre total de lemmes}}$$

Nous considérons alors que plus la valeur de la probabilité $p(l_i)$ est élevée plus il y a affinité sémantique entre le mot m_i à vérifier dont le lemme est l_i et le texte où il a été mis.

- **L'agent Coordinateur** englobe les résultats trouvés par les deux agents Cooccurrence et Répétition dans la formule :

$$F(m_i) = \alpha * I(m_i) + \beta * p(m_i|C) + \lambda * p(l_i)$$

Avec $F(m_i)$ la fréquence totale d'apparition du mot m_i au sein du texte, α, β et λ sont trois coefficients liés aux trois probabilités contextuelles calculées, les valeurs associées à ces coefficients ne peuvent être prédites, mais il faut qu'elles soient obtenues à travers des tests et des comparaisons de pertinence. Toutefois, nous estimons que les valeur de α et β sont plus importantes que celle de λ vu que l'importance du contexte voisin du mot analysé est, sans aucun doute, plus grande que celle du contexte lointain de ce même mot.

Pour chaque mot, nous calculons la valeur de $F(m_i)$ qui, comparée aux valeurs des mots voisins et à une valeur seuil, va confirmer ou infirmer la validité de l'existence de ce mot entouré de ses voisins.

Le résultat final de la vérification sémantique est aussitôt envoyé au superviseur pour qu'il déclenche le processus de correction.

5.3 Collaboration des deux groupes d'agents : syntaxique et sémantique

Avant le déclanchement du processus de correction, les deux groupes d'agents syntaxique et sémantique entrent en communication par le biais de leurs superviseurs respectifs. Etant donné que ces deux groupes d'agents sont lancés en parallèle à la recherche d'une erreur cachée dans une phrase, le premier qui trouve doit informer l'autre et demander une confirmation sur l'erreur. Si consensus, il y a, le processus de détection est arrêté et l'erreur est envoyée vers l'agent Correction. Il s'agit dans ce cas là, d'une erreur cachée provoquant un dérèglement à la fois syntaxique et sémantique. Dans le cas contraire, le deuxième groupe d'agents continue de balayer toute la phrase à la recherche d'une autre erreur. Deux cas se présentent : si ce dernier n'a pas trouvé alors il informe le premier groupe, qui, lui a déjà trouvé une erreur, pour qu'il déclenche la correction. Si par contre, il a trouvé une autre erreur, il l'envoie à son tour au premier groupe demandant une confirmation. Si les deux groupes se mettent d'accord sur cette nouvelle erreur, alors c'est cette dernière erreur qui est envoyée pour être corrigée à la place de la première erreur. Enfin, s'il les deux groupes ne se sont pas mis d'accord sur aucune erreur chacun envoie de son côté sa propre présumée erreur à l'agent Correction.

5.4 L'agent Correction

Finalement, l'agent Correction vient corriger les fautes détectées par les deux vérificateurs : syntaxique et sémantique. Il procède alors par la génération de toutes les formes proches de la forme erronée, à une différence d'édition près pour former ainsi une liste contenant les candidats à la correction. Ladite liste est assez longue. Elle contient en moyenne plus de 27 formes candidates et peut atteindre 185 formes : valeur pouvant être augmentée par l'agglutination des enclinomènes (Ben Othmane Zribi, 1998). Pour réduire ce grand nombre de propositions, l'agent Correction substitue la forme erronée par chacune des formes proposées et forme ainsi un ensemble de phrases candidates. Ces dernières seront réinjectées, au fur et à mesure de leur production, dans les deux vérificateurs (autrement dit la partie détection du système). Celles qui contiennent toujours des anomalies seront éliminées et c'est le même sort que subissent les propositions qui leur sont respectives. La liste des propositions restantes est par la suite triée par ordre de pertinence et présentée à l'utilisateur.

6 Expérimentation et résultats

Notre objectif étant la réalisation d'un système capable de détecter et de corriger les erreurs cachées, nous avons alors implémenté à ce stade de notre travail une partie du système, déjà conçu, à savoir le groupe syntaxique d'agents et intégré l'agent Correction de (Ben Othmane Zribi, 1998).

De plus, pour évaluer le système ainsi réalisé, nous avons besoin d'un corpus textuel contenant suffisamment d'erreurs cachées. Chacune de ces dernières doit être identifiée avec sa forme corrective. Toutefois, faute de corpus contenant ce genre d'erreurs sous sa forme naturelle, nous avons choisi de créer notre propre corpus manuellement. Nous avons, ainsi, généré parmi les formes qui existent dans le corpus de test une liste d'erreurs cachées tout en respectant les hypothèses restrictives de nos champs d'investigation. Ce corpus, qui constituera l'entrée de notre système, contient environ 750 formes textuelles non voyellées, dans lesquelles nous avons introduit 100 erreurs cachées du type syntaxique.

6.1 Résultats de l'évaluation de la détection

L'expérimentation de notre système de détection des erreurs cachées a donné des résultats que nous jugeons satisfaisants avec un pourcentage de précision égal à **80%** et un pourcentage de rappel égal à 77%. La présence de bruit 20% (1- Précision) et de silence 23% (1 – Rappel) s'expliquer principalement par les causes citées ci-dessous :

- La largeur de la portée de vérification dans les phrases manipulées. En effet, malgré la phase de segmentation, le nombre de mots constituant certaines phrases reste important ce qui contrarie les principes de vérification de certains agents qui travaillent à base de phrases courtes.

- La compétition entre agents, qui a été l'une des principales hypothèses de notre choix d'une architecture multi-agent et qui s'est manifestée par le fait que le premier agent détecteur de faute arrête ses collègues et ce, sans savoir si la faute détectée est ou n'est pas une fausse alarme.

- La non exhaustivité de nos règles linguistiques et l'absence de certaines informations linguistiques. Les notions d'"animé" ou d'"inanimé", pourraient par exemple aider à effectuer une meilleure vérification syntaxique ou sémantique.

6.2 Résultats de l'évaluation de la correction

Cette phase a été testée à deux niveaux ; d'abord après l'obtention de toutes les propositions de correction, ensuite après la minimisation de la liste de ces propositions. Les résultats obtenus sont illustrés dans le tableau ci-après.

	Couverture	**Précision**	**Ambiguïté**	**Proposition**	**Position**
Initialement	100%	100%	100%	82,5	8,7
Après minimisation	93,3%	86,6%	86,6%	18,4	2,8

Figure 2 Evaluation du correcteur des erreurs cachées

Nous remarquons que notre méthode de minimisation de la liste des propositions a permis de diminuer, considérablement, le nombre moyen des propositions de 77% (de 82,5 à **18,4** propositions en moyenne). Cette diminution, bien qu'elle ait réduit l'ambiguïté de notre correcteur de 13,4%, ne s'est pas passée sans dégâts. Elle s'est faite aux dépens de la couverture (diminution de 6,4%) et de la précision (diminution de 13,4%).

7 Conclusions et perspectives

La partie du système ainsi implémentée a donné des résultats assez satisfaisants. Les choix que nous avons adoptés nous ont permis d'atteindre nos objectifs initialement dressés. Cependant, nous estimons que les résultats obtenus peuvent être encore améliorés d'abord par l'amendement des règles linguistiques utilisées et ensuite par la prise en considération des informations sémantiques. Aussi, l'implémentation du groupe sémantique d'agents figure parmi les perspectives proches de nos travaux de recherche.

Références

ALOULOU C. (1996), Utilisation de l'approche multi-critère pour orienter un processus de correction des erreurs d'accord dans des phrases de la langue arabe non voyellée, Mémoire de DEA, Institut Supérieur de Gestion, Université de Tunis III.

BEN HAMADOU A. (1993), Vérification et correction automatique par analyse affixale des textes écrits en langue naturelle : le cas de l'arabe non voyellé, Thèse d'état en informatique, Faculté des Sciences de Tunis.

BEN OTHMANE ZRIBI C. (1998), De la synthèse lexicographique à la détection et à la correction des graphies fautives arabes, Thèse de doctorat, Université de Paris XI, Orsay.

BEN OTHMANE ZRIBI C. et BEN AHMED M. (2003), Le contexte au service de la correction des graphies fautives arabes, *TALN'03*, Nantes.

BIGERT J., KNUTSSON O. (2002), Robust Error Detection : A Hybrid Approach Combining Unsupervised Error Detection and Linguistic Knowledge, In *Proceedings of Robust Methods in Analysis of Natural Language Data (ROMAND'02)*, Frascati, Italie.

DEBILI F., ACHOUR H., SOUISSI E. (2002), La langue arabe et l'ordinateur: De l'étiquetage grammatical à la voyellation automatique, *Correspondances N°71*, Lyon.

GOLDING A. R. (1995), A bayesian hybrid method for context- sensitive spelling correction, In *Proceedings of the Third Workshop on Very Large Corpora*, Cambridge, Massachusetts, USA, pages 39-53.

GOLDING A. R., SCHABES Y. (1996), Combining trigram-based and feature-based methods for context-Sensitive Spelling Correction, *ACL'96*, San Fransisco.

GOLDING A. R., DAN R., (1999), A winnow-based approach to context-sensitive spelling correction, *Machine Learning*, 34(1-3), 107-130.

VERBERNE S. (2002), Context sensitive spell checking based on word trigram probabilities, Mémoire de Mastère, Université de Nijmegen.

XIAOLONG W., JIANHUA L. (2001), Combine trigram and automatic weight distribution in Chinese spelling error correction, *Journal of computer Science and Technology, Volume 17 Issue 6*, Province, China.

Structure des représentations logiques, polarisation et sous-spécification

Sylvain Kahane

Modyco, Université Paris 10
Lattice, Université Paris 7
sk@ccr.jussieu.fr

Résumé – Abstract

Cet article s'intéresse à la structure des représentations logiques des énoncés en langue naturelle. Par représentation logique, nous entendons une représentation sémantique incluant un traitement de la portée des quantificateurs. Nous montrerons qu'une telle représentation combine fondamentalement deux structures sous-jacentes, une structure « prédicative » et une structure hiérarchique logique, et que la distinction des deux permet, par exemple, un traitement élégant de la sous-spécification. Nous proposerons une grammaire polarisée pour manipuler directement la structure des représentations logiques (sans passer par un langage linéaire avec variables), ainsi qu'une grammaire pour l'interface sémantique-syntaxe.

This paper aims at the structure of logic representations in natural languages. By logic representation we mean a semantic representation including a quantifier scope processing. We show that such a representation basically combines two underlying substructures, a "predicative" structure and a logic hierarchic structure, and that the identification of the two allows for an elegant processing of underspecification. We will propose a polarized grammar that directly handles the structure of logic representations (without using a linear language with variables), as well as a grammar for the semantics-syntax interface.

Mots Clés –Keywords

Logique du premier ordre, calcul des prédicats, représentation sémantique, relation prédicat-argument, quantificateur, grammaire d'unification polarisée, grammaire de dépendance, dag, interface syntaxe-sémantique.

First order logic, predicate calculus, semantic representation, predicate-argument relation, quantifier, polarized unification grammar, dependency grammar, dag, syntax-semantics interface.

1 Introduction

Les objectifs de cet article sont multiples. D'un point de vue mathématique, il s'agit de mieux comprendre la structure des formules logiques du premier ordre (= formules du calcul des

prédicats), notamment lorsque celles-ci sont utilisées comme représentations sémantiques d'énoncés en langue naturelle. Il s'agit en particulier de mieux comprendre la nature des quantificateurs et de leur portée. D'un point de vue linguistique, il s'agit de combiner deux modes de représentations utilisés en sémantique des langues naturelles : des représentations logiques, utilisées en sémantique non lexicale et issues des travaux de Frege et d'autres logiciens, et des représentations issues de la sémantique lexicale, qui traitent tous les signifiés lexicaux comme des prédicats, y compris les quantificateurs et la négation (cf. Dymetman & Coperman 1996 pour une problématique analogue).

La section 2 introduira différentes « écritures » pour une représentation logique et mettra en évidence les deux structures sous-jacentes, tandis que la section 3 proposera une grammaire permettant de manipuler directement de telles structures, de réaliser une interface sémantique-syntaxe et de contrôler la sous-spécification.

2 Représentation logique

2.1 Variations sur la représentation logique

Cette section s'articulera autour des différents modes de représentation du sens d'un énoncé comportant plusieurs quantificateurs. Nous travaillerons avec l'énoncé suivant :

(1) *Tout homme aime une femme.*

On lui associe généralement la représentation logique suivante :

(2) $\forall x$ [homme'(x) → $\exists y$ [femme'(y) ∧ aimer'(x,y)]]

En s'autorisant la confusion entre un prédicat et son extension, on peut utiliser la représentation suivante, plus proche de la langue naturelle :

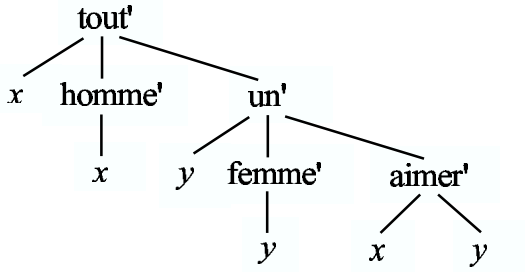
(3) $\forall x \in$homme' [$\exists y \in$femme' aimer'(x,y)]

Dans une telle représentation, un quantificateur est lié à trois objets : une *variable*, une *restriction* (sur cette variable) et la *portée*. Pour le quantificateur $\forall$ en (3), il s'agit respectivement de la variable x, de la restriction homme'(x) et de la sous-formule $\exists y \in$femme' aimer'(x,y). De nombreuses approches en sémantique (Woods 1975, Copestake *et al.* 1999) rendent cette structure explicite en associant à (1) la représentation suivante :

(4) tout'(x, homme'(x), un'(y, femme'(y), aimer'(x,y)))

On pourra ainsi représenter les autres quantificateurs (*certains, quelques, la plupart de,* etc.) de la même façon que *tout* et *un*, c'est-à-dire par un prédicat à trois arguments. La formule en (4) possède une structure arborescente que nous donnons en (5) :

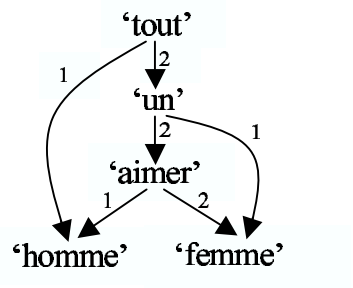
(5)



La représentation des quantificateurs comme des prédicats à trois arguments est conséquente à l'utilisation d'une variable liée, qui n'a pas de contribution sémantique, mais uniquement un rôle structural en liant différentes positions de la formule. La variable n'est utile qu'en raison

de l'écriture linéaire[1] de la formule en (4). On peut en effet supprimer la variable en (5) en associant à (4) non pas un arbre, mais un dag[2], c'est-à-dire une structure réentrante où les différentes occurrences de la variable correspondent à un seul nœud. Mais ce nœud fait alors double emploi avec le nœud de homme' et on peut considérer que la variable *x* n'est en fait qu'une réification de cet élément sémantique. Ceci est assez cohérent avec le fait que le sémantème 'homme', le signifié du nom *homme*, représente en fait une entité (un objet du monde), qui même s'il est indéterminé, n'est pas en soi un prédicat unaire (le prédicat unaire homme' que nous avions introduit correspond à l'expression *être un homme* plutôt qu'au nom *homme* proprement dit). Nous obtenons ainsi une nouvelle représentation pour (1)[3] :

(6)



Bien que directement liée aux représentations logiques les plus classiques, cette représentation n'a jamais été proposée à notre connaissance. Polguère (1992) propose une représentation très similaire, en ajoutant à une structure prédicative (voir caractérisation plus bas) un argument de portée pour les quantificateurs ; mais contrairement à ici, la portée est explicitement représentée comme une portion de la structure prédicative, alors que nous l'encodons, comme nous le verrons plus loin, par une relation hiérarchique.

Dans cette nouvelle représentation, les quantificateurs ('tout' et 'un') sont explicitement des opérateurs à deux arguments (cf. Barwise & Cooper 1981). En fait, même dans la représentation (4), en raison du caractère lié de la variable, le quantificateur est aussi un opérateur à deux arguments, dont on peut donner le lambda-terme λPλQ.[tout'(*x*,P(*x*),Q(*x*))], ou même λPλQ.[tout'(λ*x*[P(*x*),Q(*x*)])], en explicitant le caractère lié de la variable *x*.

Nous adopterons dorénavant la représentation proposée en (6), que nous appellerons la *représentation sémantique* de l'énoncé (1). Nous allons maintenant étudier plus en détail la structure de cette représentation. Elle combine en fait deux sous-structures : une structure prédicative et une structure hiérarchique liée à la portée.

Notons auparavant que notre représentation sémantique n'est pas moins riche que la représentation logique traditionnelle (en (2), (3) ou (4)), puisqu'on peut revenir à la représentation logique en réifiant (à la façon de Davidson 1967 pour les évènements) les sémantèmes sur lesquels pointent les quantificateurs, c'est-à-dire en leur associant une variable, en leur ajoutant cette variable comme argument et en donnant cette variable comme argument aux prédicats pointant sur eux (voir section 3.4 pour un exemple).

---

[1] On utilise aussi la variable pour désigner les référents de discours. Néanmoins, cet usage devrait être nettement séparé de l'autre, car dans ce deuxième cas la variable renvoie à un objet du monde et ne devrait pas figurer comme argument d'un sémantème, si l'on distingue sens et dénotation.

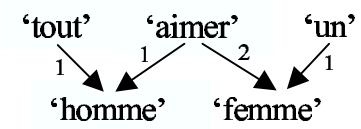
[2] On appelle *dag* un graphe orienté acyclique (de l'anglais *directed acyclic graph*).

[3] La structure en (5) est un arbre ordonné, c'est-à-dire que les fils de chaque nœud sont ordonnés, indiquant ainsi pour chaque prédicat quel est son premier (deuxième …) argument. En (6), nous adoptons une autre convention : les différents arcs du graphe sont étiquetés pour indiquer les différentes places argumentales. Par ailleurs, comme nous le verrons, ce graphe est partiellement hiérarchisé : les arcs droits correspondent à la structure arborescente sous-jacente (la structure hiérarchique logique, section 2.2).

2.2 Représentation logique et structure prédicative

La *structure* que nous appelons *prédicative* (qu'on aurait pu aussi appeler structure argumentale) est un graphe de relations prédicat-argument, où les prédicats représentent des sémantèmes, c'est-à-dire les signifiés des éléments lexicaux de l'énoncé. Une telle structure est utilisée comme représentation sémantique par la théorie Sens-Texte (Mel'cuk 1988, Polguère 1992, Kahane 2001). Une relation prédicative correspond à une relation actancielle ou modificative et doit être « validée » par une relation de dépendance syntaxique[4]. En tant que prédicat sémantique, les quantificateurs n'ont qu'un argument : le nom qu'ils déterminent. La portée est d'une nature différente : cela n'aurait pas de sens de considérer qu'il existe en (1) une relation prédicative entre *tout* et *un*, car il n'y a aucune relation syntaxique entre eux. La structure prédicative de (1) est donc :

(7)

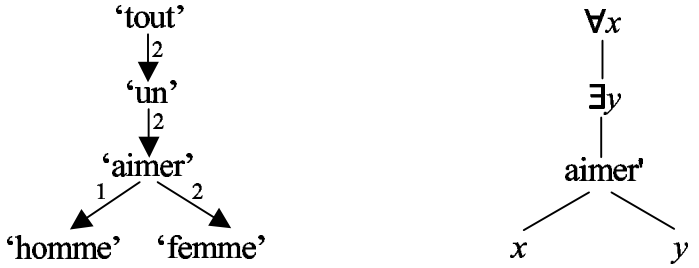


Par définition, la structure prédicative est très proche de la structure syntaxique, lorsque celle-ci est représentée par un arbre de dépendance syntaxique. Nous donnons en (8) l'arbre syntaxique de (1) ainsi que le graphe hiérarchisé obtenu en superposant au graphe en (7) la hiérarchie de l'arbre syntaxique.

(8)

On voit que la structure syntaxique est en quelque sorte une hiérarchisation de la structure prédicative. Or la représentation sémantique possède également une hiérarchie sous-jacente, mais celle-ci est différente et rend compte des relations de portée[5]. Nous donnons en (9) la *structure hiérarchique logique* de (1), ainsi que l'arbre de décomposition de la formule logique en (3) :

(9)



Comme on le voit en (9), c'est la structure hiérarchique que privilégie la représentation logique traditionnelle, au détriment de la structure prédicative. On peut noter des distorsions

---

[4] Il existe en fait un certain nombre de distorsions possibles entre la structure prédicative et la structure syntaxique (Mel'cuk 1988, Kahane 2001). C'est le cas par exemple des phénomènes de montée, comme dans *Pierre semble dormir*, où la relation prédicative entre 'dormir' et 'Pierre' est validée par la relation syntaxique sujet entre *semble* et *Pierre*.

[5] Nous laissons de côté une question fort intéressante, mais qui dépasse le cadre de cet article. Il s'agit des liens entre la structure informationnelle et la structure logique. La théorie Sens-Texte n'a par exemple jamais introduit de structure hiérarchique explicite, considérant que celle-ci n'est qu'une conséquence de la structure communicative et notamment de la partition thème-rhème (Polguère 1992, Mel'cuk 2003).

notables entre la hiérarchie logique et la hiérarchie syntaxique, sources de réelles difficultés pour l'interface sémantique-syntaxe, lorsque la structure sémantique repose sur la hiérarchie logique comme en (2)-(4). La structure logique est d'ailleurs plus directement liée à l'ordre linéaire qu'à la structure syntaxique, à notre avis. Ainsi, en (10)a et b, dans l'interprétation préférée[6], 'tout' est dans la portée de 'un', à l'inverse de (1), alors que l'ordre des quantificateurs dans la phrase est également l'inverse de celui de (1) :

(10) **a.** *Une femme est aimée de tout homme.*
b. *Il y a une femme que tout homme aime.*

La représentation sémantique que nous avons adoptée peut être étendue, par exemple pour traiter des modifieurs. Dans *un homme heureux*, 'heureux' sera traité comme un prédicat unaire ayant comme argument 'homme' et se trouvant dans la portée de 'homme' (la relation prédicat-argument et la portée auront des sens opposés). Ce prédicat peut lui-même être modifié, par exemple, par 'très'. La portion de la représentation se trouvant dans la portée d'un sémantème représentant une entité restreint l'extension de ce sémantème et contribue à la restriction du quantificateur portant sur cette entité[7].

Nous allons maintenant proposer une grammaire permettant de générer nos représentations sémantiques, avant de l'étendre pour en faire une interface sémantique-syntaxe.

3 Une grammaire pour les représentations logiques

Nous allons proposer une grammaire « lexicalisée » pour les représentations logiques, c'est-à-dire une grammaire qui construit la représentation en combinant des morceaux de structure associés aux différents sémantèmes. Nous nous tournons vers un formalisme capable de manipuler des graphes, les grammaires d'unification polarisées (Kahane 2004). Nous rappelerons brièvement ce formalisme avant de présenter notre grammaire sémantique.

3.1 Grammaires d'unification polarisées

Les grammaires d'unification polarisées sont des grammaires permettant de générer des ensembles de structures finies. Une structure repose sur des *objets*. Par exemple, pour un graphe (orienté), les objets sont les nœuds et les arcs. Chaque arc est lié à deux nœuds par les

---

[6] Comme on le sait, il est possible qu'un quantificateur en deuxième position ait une portée large. Par exemple (10)a peut éventuellement recevoir la même interprétation que (1), mais une telle interprétation est plus difficile d'accès et doit être conditionnée par le contexte ou les connaissances du monde.

[7] Nous pouvons traiter d'autres types de restrictions, comme une relative ou une participiale. Ainsi, pour la phrase *Tout homme aimant une femme est heureux*, les relations de portées seront :

'tout' —2→ 'heureux' —1→ 'homme' ——→ 'un' —2→ 'aimer' —2→ 'femme'.

La relation de portée entre 'homme' et la participiale est « validée » par la relation prédicative 'aimer' —1→ 'homme', qui est dans le sens inverse (comme pour un adjectif).

Les syntagmes nominaux complexes, comme dans *Tout locuteur de deux langues est heureux*, posent un problème intéressant. Ici les relations de portées sont :

'tout' —2→ 'deux' —2→ 'heureux' —1→ 'locuteur' —1→ 'langue'.

Il est également intéressant de comparer des exemples comme (i) *Les soldats fatigués se sont arrêtés* et (ii) *Les soldats, fatigués, se sont arrêtés*. La structure prédicative est la même ('s'arrêter' —1→ 'soldat' ←1— 'fatigué'), mais, en (i) seulement, 'fatigué' restreint la portée du quantificateur et doit être dans la dépendance logique de 'soldat'.

fonctions source et cible. Ce sont ces fonctions qui fournissent la structure proprement dite. Une *structure polarisée* est une structure dont les objets sont polarisés, c'est-à-dire étiquetés par une valeur appartenant à un ensemble fini P de polarités. L'ensemble P est muni d'une opération commutative et associative notée « . », appelée *produit*. Un sous-ensemble N de P contient les polarités dites *neutres*. Une structure polarisée est dite *neutre* si tous ses objets sont neutres.

Nous allons utiliser un système de polarités P = {●,○,●}, avec N = {●,●}, et un produit défini par le tableau suivant (où ⊥ représente l'impossibilité de se combiner). Nous appellerons nos polarités ● = noir = saturation, ○ = blanc = contexte (obligatoire) et ● = gris = neutre absolu.

·	●	○	●
●	●	○	●
○	○	○	●
●	●	●	⊥

Les structures peuvent être combinées par *unification*. L'unification de deux structures A et B donne une nouvelle structure A⊕B obtenue en « collant » ensemble ces structures par l'identification d'une partie des objets de la première structure avec ceux de la deuxième. Lorsque deux structures polarisées A et B sont unifiées, la polarité d'un objet de A⊕B obtenu par identification de deux objets de A et B est le produit de leurs polarités.

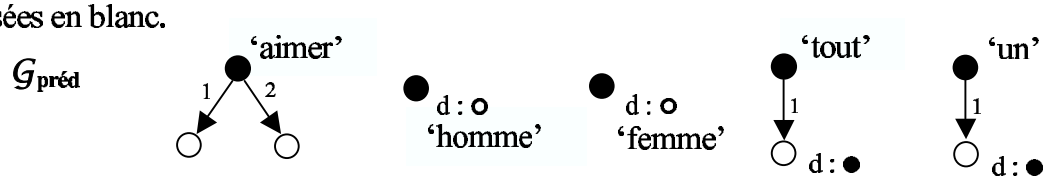
Une *grammaire d'unification polarisée* (GUP) est définie par une famille finie T de types d'objets (avec des fonctions attachées aux différents types d'objets), un système (P,.) de polarités, un sous-ensemble N de P de polarités neutres, et un ensemble fini de structures élémentaires polarisées, dont les objets sont décrits par T et dont une est éventuellement marquée comme la structure initiale (et appelée top dans la suite). Les structures *générées* par la grammaire sont les structures neutres obtenues par combinaison de l'éventuelle structure initiale et d'un nombre fini de structures élémentaires.

Rappelons que le formalisme est monotone (avec l'ordre ● < ○ < ● sur les polarités) et que les structures peuvent être combinées absolument dans n'importe quel ordre.

3.2 Une grammaire sémantique

Nous allons présenter notre grammaire en plusieurs étapes pour en faciliter la compréhension et pour en montrer le caractère élémentaire.

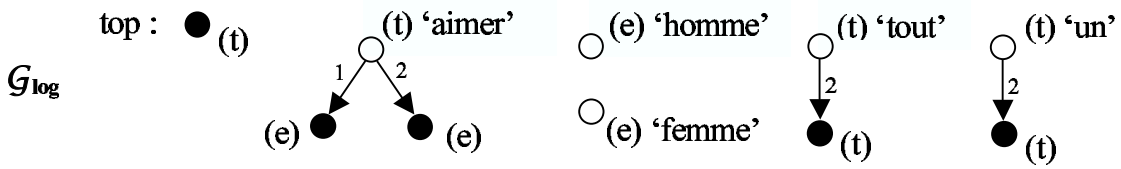
La structure prédicative est un graphe. Elle peut donc être générée par une grammaire aussi simple que $\mathcal{G}_{\text{préd}}$, où chaque prédicat ancre un morceau du graphe avec ses arguments : le nœud du prédicat est saturé et les positions des arguments doivent être remplies et sont polarisées en blanc.



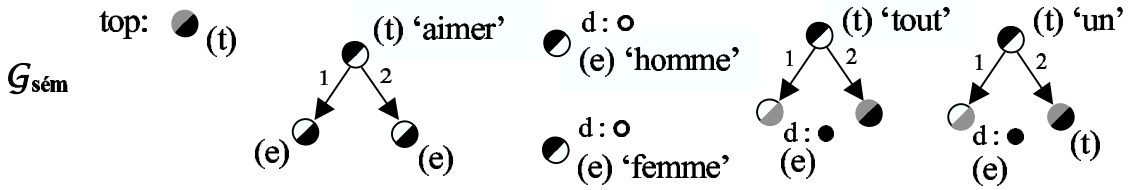
Nous avons ajouté une polarité supplémentaire d : ○ et d : ● sur certains nœuds. La polarité d : ○ indique que ces sémantèmes sont indéterminés. Un sémantème d'une entité déterminée, comme 'Paul', recevra une polarité d : ●. La polarisation multiple fonctionne ainsi : si un objet A de polarités (x,y) et un objet B de polarités (z,t) sont identifiés, l'objet résultant reçoit le couple de polarités $(x.z,y.t)$.

La structure logique est un arbre. La structure d'arbre est très facile à encoder avec une GUP (Kahane 2004) : il suffit d'assurer que chaque élément autre que la racine ait un unique

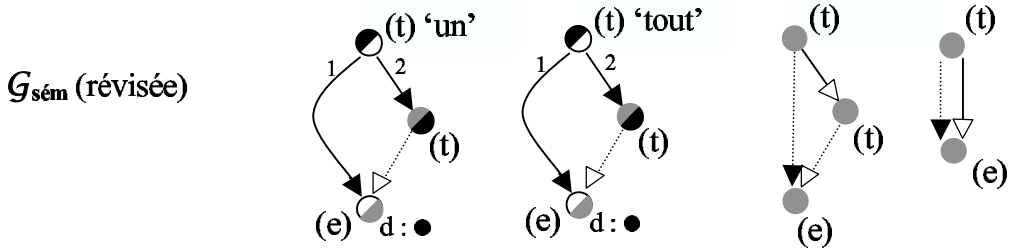
gouverneur (la connexité étant elle assurée par définition). Nous obtenons la grammaire $\mathcal{G}_{log}$, où chaque sémantème ancre la racine d'un sous-arbre. Nous typons, selon l'usage, les nœuds en deux classes : e pour les entités et t pour les propositions (*truth value*).



En superposant[8] $\mathcal{G}_{préd}$ et $\mathcal{G}_{log}$, nous obtenons la grammaire sémantique $\mathcal{G}_{sém} = \mathcal{G}_{préd} \times \mathcal{G}_{log}$. Du fait de la superposition, chaque nœud hérite d'un couple de polarité. Un nœud qui n'apparaît pas dans une des deux grammaires est considéré comme neutre absolu là où il n'apparaît pas. Nous notons (○,●) = ◒ et (●,○) = ◓. On remarquera que ◒ et ◓ se comportent comme des polarités opposés de type besoin-ressource.



Nous devons encore enrichir un peu $\mathcal{G}_{sém}$ pour assurer que la restriction d'un quantificateur est dans sa portée. Une GUP permet facilement de contrôler le chemin entre deux nœuds : nous ajoutons pour cela un lien de dominance dans la structure du quantificateur entre la portée et la restriction. Une grammaire dédiée permet de « réécrire » ce lien de dominance en une chaîne de dépendance. On peut facilement contrôler la nature de cette chaîne, puisqu'une GUP peut simuler n'importe quelle grammaire de réécriture (Kahane 2004). Nous donnons ci-dessous la grammaire $\mathcal{G}_{sém}$ révisée, avec les structures complètes pour les quantificateurs et, à droite, les deux règles nécessaires à la propagation de la dominance, où les liens de dominance sont représentés par des flèches pointillées.



3.3 Interface sémantique-syntaxe

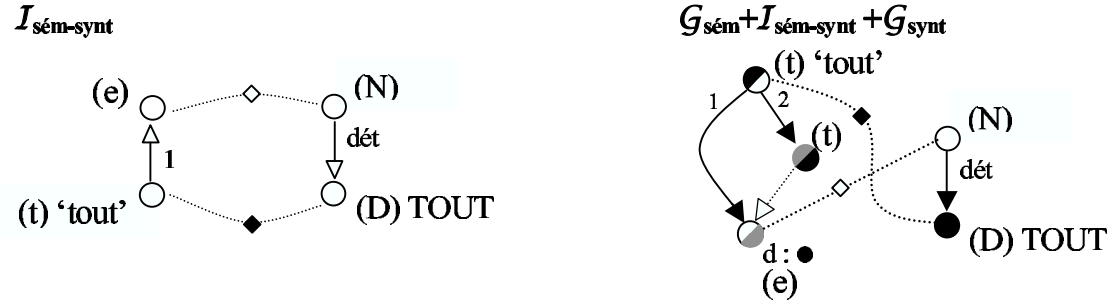
Nous appelons interface sémantique-syntaxe une grammaire capable de faire se correspondre une représentation sémantique et une représentation syntaxique. La difficulté de la tâche vient du fait qu'il y a entre les deux représentations des distorsions importantes au niveau de la hiérarchie entre les éléments[9]. Le problème a été étudié dans le cadre de nombreux

[8] Comme l'a fait remarquer un relecteur, la superposition des grammaires est un cas particulier de synchronisation (voir section 3.3). Celle-ci est possible quand les deux grammaires manipulent les mêmes objets.

[9] Le fait que notre représentation syntaxique soit un arbre de dépendance plutôt qu'un arbre syntagmatique ne change pas fondamentalement le problème. Nous préférons pour notre part traiter la question de l'ordre linéaire séparément et considérer une structure syntaxique sans ordre linéaire.

formalismes à commencer par la grammaire de Montague (1973). Celle-ci privilégie l'organisation logique, alors que des travaux plus récents tendent à favoriser l'organisation syntaxique, comme Copestake *et al.* 1999 avec HPSG ou Kallmeyer & Joshi 1999 avec TAG.

Pour notre part, nous résolvons le problème de manière triviale, en synchronisant notre grammaire sémantique avec une grammaire de dépendance syntaxique, sans privilégier aucun des deux niveaux de représentation. Seule la structure prédicative est prise en compte dans cette interface sémantique-syntaxe ($\mathcal{I}_{\text{sém-synt}}$). En combinaison avec $\mathcal{G}_{\text{sém}}$ et $\mathcal{G}_{\text{synt}}$, elle permet de construire parallèlement les représentations sémantiques et syntaxiques[10]. En GUP, la synchronisation de deux grammaires s'effectue simplement en alignant les deux grammaires et en synchronisant certains nœuds (cf. Shieber & Schabes 1990, Perrier 2004, Kahane & Lareau 2005). La synchronisation est assurée par des liens dit de synchronisation qui obligeront, lorsqu'on identifie deux nœuds dans une des structures, à identifier les nœuds qui sont synchronisés avec eux dans l'autre structure. L'obligation d'unifier les liens de synchronisation (et donc de synchroniser les éléments qu'ils relient) est encore une fois assurée par la polarisation (représentée dans des losanges).



Notre grammaire est équivalente au niveau de l'expressivité aux autres interfaces sémantique-syntaxe évoquées et s'apparente tout particulièrement à la grammaire proposée par Perrier 2004, qui est à notre connaissance le premier à avoir utilisé une grammaire polarisée pour traiter les représentations logiques et la sous-spécification. Notre grammaire peut être utilisée pour l'analyse comme pour la synthèse et sera compatible avec diverses procédures. Une procédure pour une GUP consiste à donner un ordre dans le traitement des structures (par exemple suivre l'ordre imposé par la hiérarchie logique ou syntaxique) et choisir un ordre dans la neutralisation des polarités. L'un de ces ordres mérite quelques commentaires.

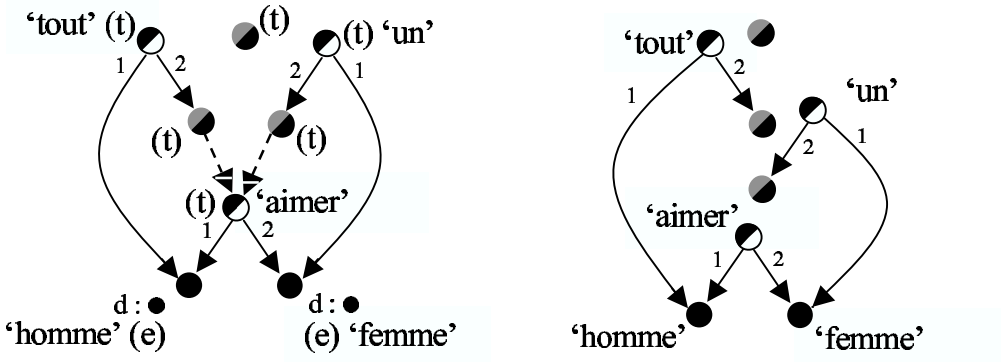
3.4 Sous-spécification et neutralisation partielle

Lorsqu'on utilise notre grammaire pour l'analyse, on remarque que la structure syntaxique ne permet pas de décider quelle doit être la portée respective des quantificateurs. Pour ne pas introduire une ambiguïté qu'on n'a pas les moyens de résoudre, on peut préférer construire une représentation sémantique où la hiérarchie logique reste sous-spécifiée. Cette question a été beaucoup étudiée dans la littérature récente (Reyle 1993, Bos 1995, Copestake et al. 1999, Kallmeyer & Joshi 1999, Perrier 2004). Dans notre cadre, la solution est élémentaire : il suffit de ne pas chercher à neutraliser les polarités qui contrôlent la hiérarchie logique (c'est-à-dire

[10] Notre grammaire syntaxique $\mathcal{G}_{\text{synt}}$ a été décrite dans de nombreux articles (notamment Kahane 2001 et Kahane & Lareau 2005). Le fait que les noms doivent avoir un unique déterminant est ici assuré par la polarité d de la grammaire sémantique. Il faut néanmoins noter que des éléments sémantiquement déterminés comme 'Paris' ou 'la France' peuvent être ou non syntaxiquement déterminés (*la France* vs *en France*) et qu'un traitement syntaxique de la détermination doit venir compléter le traitement purement sémantique.

de ne pas considérer $\mathcal{G}_{\text{préd}}$ lors de l'analyse). On obtient, pour (1) ou (10), la *représentation sémantique sous-spécifiée* (11) (en version simplifiée à droite, en mettant en vis-à-vis les nœuds qui fusionneront pour obtenir (6)), où comme on le voit les seules polarités non neutres sont les polarités blanches associées à la hiérarchie logique :

(11)



Autrement dit, l'interface syntaxe-sémantique est uniquement réalisée entre la structure syntaxique et la structure prédicative. C'est ce que préconise depuis les années 60 la théorie Sens-Texte, originellement conçue pour la traduction automatique, où la levée des ambiguïtés logiques est généralement inutile.

Notons que la représentation sémantique en (11) peut être traduite en une écriture linéaire en réifiant les différents nœuds : p pour la racine, p_1 et p_2 pour les quantificateurs 'tout' et 'un', e pour le prédicat 'aimer', x et y pour les entités 'homme' et 'femme' et h_1 et h_2 pour les portées des quantificateurs (h pour *hole*, d'après la Hole Semantics de Bos 1995) :

(12) top:p, p_1:'tout'(x,h_1), p_2:'un'(y,h_2), e:'aimer'(x,y), x:'homme', y:'femme'

ou encore avec une double réification des entités (en introduisant les variables a_1 et a_2) :

(13) top:p, p_1:tout'(x,a_1,h_1), p_2:un'(y,a_2,h_2), e:aimer'(x,y), a_1:homme'(x), a_2:femme'(y)

Les relations de dominance, ainsi que la nécessité de neutraliser les trous h_1 et h_2, seront données par des conditions telles que :

(14) $p \geq p_1 > h_1 \geq e$, $p \geq p_2 > h_2 \geq e$, $\{p,h_1,h_2\} = \{e,p_1,p_2\}$

Même si les deux représentations, (11) et (13)+(14) (qui est la représentation de Copestake *et al.* 1999) sont (quasiment) équivalentes, on peut voir un avantage à utiliser des structures polarisées et à ne pas utiliser d'écriture linéaire forçant l'introduction de multiples variables.

4 Conclusion

Notre contribution concerne la représentation sémantique des énoncés et tout particulièrement la structure logique, c'est-à-dire la représentation des phénomènes de portée. Nous n'avons pas réellement abordé de questions de logique, c'est-à-dire la façon dont nos représentations pouvaient servir directement à un calcul logique. De ce point de vue, nous nous sommes couverts par la possibilité de revenir aux formules usuelles de la logique du premier ordre, mais une description directe de l'inférence dans notre formalisme serait certainement intéressante. Notre objectif était d'avoir une représentation qui soit à la fois suffisante pour l'interprétation logique, mais qui permette aussi un traitement adéquat d'autres phénomènes sémantiques qui ne se manifestent pas en termes de portée.

Le véritable point de notre contribution est d'avoir rappelé que la structure sémantique des énoncés n'est pas linéaire et qu'il y a tout intérêt dans ces conditions à utiliser un langage qui

permette de manipuler directement des arbres, des graphes et des produits des deux. Nous avons ainsi montré qu'une représentation sémantique est la superposition d'une structure prédicative et d'une structure hiérarchique encodant les relations de portée. Lorsqu'on ne privilégie plus la relation de portée, comme le fait la logique classique, il devient facile de définir une représentation sous-spécifiée en « neutralisant » une des deux structures seulement.

Remerciements

Je remercie chaleureusement pour leurs commentaires Pascal Amsili, Laurence Danlos, Laura Kallmeyer, François Lareau, Guy Perrier, Alain Polguère, Igor Mel'cuk, Benoît Sagot et les trois relecteurs de TALN.

Références

BARWISE J. & COOPER R. (1981), Generalized quantifiers and natural language, *Linguistics and Philosophy*, *4,* 159-219.

BOS J. (1995), Predicate Logic Unplugged, *Tenth Amsterdam Colloquium.*

COPESTAKE A., FLICKINGER D. & SAG I. A. (1999), Minimal Recursion Semantics : An Introduction, draft, 26 p.

DAVIDSON D. (1967), The Logical Form of Action Sentences, in N. Rescher (ed.), *The Logic of Decision and Action*, University of Pittsburgh Press, 81-95. Reprinted in D. Davidson, *Essays on Actions and Events*, Oxford: Clarendon Press, 1990,105-122.

DYMETMAN M., COPPERMAN M. (1996), Extended dependency structures and their formal interpretation, *Proceedings CoLing*, Copenhague.

KAHANE S. (2001), Grammaires de dépendance formelles et théorie Sens-Texte, Tutoriel, *Actes TALN*, *vol. 2*, 17-76.

KAHANE S. (2004), Grammaires d'unification polarisées, *Actes TALN*, Fès, 233-242.

KAHANE S. & LAREAU F. (2005), Grammaire d'Unification Sens-Texte : polarisation et modularité, *Actes TALN*, Dourdan, 10 p.

KALLMEYER L. & JOSHI A. (1999), Factoring predicate argument and scope semantics: Underspecied semantics with LTAG, *Proceedings of the 12th Amsterdam Colloquium.*

MEL'CUK I. (1988), *Dependency Syntax: Theory and Practice*, SUNY Press.

MEL'CUK I. (2003), *Communicative Organisation of Natural Language*, Benjamins.

MONTAGUE R. (1973), The proper treatment of quantification in ordinary English, in J. Hintikka (éd.), *Approaches to Natural Language*, 221-242, Reidel.

PERRIER G. (2004), La sémantique dans les grammaires d'interaction, *Actes TALN,* Fès, Maroc, 351-360.

POLGUERE A. (1992), Remarques sur les réseaux sémantiques Sens-Texte, in A. Clas (éd.), *Le mot, les mots, les bons mots*, Presses de l'Université de Montréal.

REYLE U. (1993), Dealing with ambiguities by underspecification, *Journal of semantics, 10*, 123-179.

SHIEBER S. M. & SCHABES Y. (1990), Synchronous tree-adjoining grammars, *Proceedings CoLing*, *vol. 3*, 253-258, Helsinki, Finland.

WOODS W.A. (1975), What's in a link: Foundations for semantic networks, in D. Bobrow & A. Collins, *Representation and Understanding–Studies in Cognitive Science*, 55-82, Academic Press, Orlando.

Representational and architectural issues in a limited-domain medical speech translator

Manny Rayner (1), Pierrette Bouillon (1), Marianne Santaholma (1),
Yukie Nakao (2)
(1) University of Geneva, TIM/ISSCO
40, bvd du Pont-d'Arve,
CH-1211 Geneva 4, Switzerland
mrayner@riacs.edu, Pierrette.Bouillon@issco.unige.ch,
Marianne.Santaholma@eti.unige.ch
(2) National Institute for Communications Technology,
3-5 Hikaridai, Seika-cho, Soraku-gun,
Kyoto, Japan 619-0289
yukie-n@khn.nict.go.jp

Mots-clefs : reconnaissance de la parole, traduction de la parole, aide au diagnostic médical

Keywords: speech understanding, speech translation, computer-aided diagnosis

Résumé Cet article dresse un aperçu du système MedSLT, un système de traduction de la parole dans le domaine médical pour un vocabulaire limité. Il met l'accent sur le problème du choix du type de représentation pour les constructions temporelles et causales. Nous montrons que celles-ci ne peuvent pas être représentées par des structures plates, généralement utilisées pour ce type d'application, mais qu'elles nécessitent des stuctures plus riches, enchâssées, qui permettent d'obtenir une traduction plus adéquate. Nous expliquons comment produire ces représentations et écrire des règles de traduction économiques qui mettent en correspondance les représentations sources dans la représentation interlingue correspondante

Abstract We present an overview of MedSLT, a medium-vocabulary medical speech translation system, focussing on the representational issues that arise when translating temporal and causal concepts. Although flat key/value structures are strongly preferred as semantic representations in speech understanding systems, we argue that it is infeasible to handle the necessary range of concepts using only flat structures. By exploiting the specific nature of the task, we show that it is possible to implement a solution which only slightly extends the representational complexity of the semantic representation language, by permitting an optional single nested level representing a subordinate clause construct. We sketch our solutions to the key problems of producing minimally nested representations using phrase-spotting methods, and writing cleanly structured rule-sets that map temporal and phrasal representations into a canonical interlingual form.

1 Introduction

As a subject, automatic speech translation is now a little more than ten years old. First generation systems, like Verbmobil (Wahlster, 2000), Spoken Language Translator (Rayner *et al.*, 2000) and Janus III (Lavie *et al.*, 1997) were essentially proofs of concept. We are now progressing to the stage where people want to build systems that have some claim to be useful: prominent recent examples are NESPOLE! (Lavie *et al.*, 2001), Tongues (Black *et al.*, 2002) and Phraselator (Phraselator, 2004). For obvious reasons, one application area that stands out is medical translation; this paper will focus on representational issues in MedSLT, a medium-vocabulary medical speech translation system (Rayner & Bouillon, 2002; Rayner *et al.*, 2003a).

There are many different contexts in which medical translation could potentially be useful. In the scenario targeted by the MedSLT system, we envisage that a doctor wishes to perform a preliminary examination of a patient who does not speak the doctor's language. The system allows the doctor to pose normal examination questions in her own language, translating them into the patient's language. The task appears tractable, given current speech technology. Medical examination questions are fairly stereotypical. It is also feasible for the dialogue to be one-way, with the patient responding non-verbally, so the user (i.e. the doctor) can reasonably be assumed to have had time to acclimatize themselves to the system and learn its capabilities.

The first question we need to ask is what metrics we should use to evaluate the success or failure of the system; evaluation of machine translation is notoriously difficult, and must normally be carried out with reference to a specific task. In the context of the MedSLT task, the translation system is basically a diagnostic tool; thus, the critical question is whether the patient's responses will give the doctor misleading information. This has several implications. There is no particular requirement that translations be completely literal; often, a non-literal translation will be as good, or indeed better. It is however important for translations to be concrete and clear in meaning, even if this involves losing nuances in the source utterances — this contrasts sharply with many translation and interpretation tasks, where nuances of meaning can be vital. Above all, the system must be extremely reliable, since the consequences of a mistranslation can be serious.

Putting these requirements together, we arrive at the basic design. Given the uncertainty inherent in current speech understanding and machine translation technology, processing cannot be fully automatic. We always need to make sure that the system has understood correctly before it translates. Since translation will not in general be completely literal, the system needs to echo back to the source-language user an accurate paraphrase of the translation it proposes to ask. The user will then have the option of either approving the translation and proceeding, or else aborting.

As always, there is tension between precision and recall. If linguistic coverage is too restricted, the system becomes hard to use. However, robust coverage must be balanced against the potentially very serious consequences of a mistranslation in a safety-critical task. Precision, which in this context is going to mean precision on the utterances which the user approved for translation, is thus more important than recall.

Here, we will be particularly concerned with the representations used by the MedSLT system for source, target and interlingual levels of structure. Following on from the previous points, the key issues are the following:

- If we want to prioritise reliability, we prefer to have a tightly constrained set of representational primitives.

- If the system is going to have adequate coverage, it needs to be able to represent all the relevant concepts in the domain. In this domain, the problematic cases mostly involve temporal and causal relations.

- With regard to the abstract structure of the representation, the critical dimension is the opposition between nested and flat representions. Nested representations (parse trees, logical forms etc) are more fine-grained, and make it easier to support a wide range of constructs. Conversely, flat representations hide linguistic structure, but are inherently more robust. In particular, they are well-suited to speech understanding architectures based on phrase-spotting and other methods suitable for processing noisy input. This point is sufficiently important that many researchers working in spoken language understanding simply take for granted that all semantic representations will be flat lists of key/value pairs; (Young, 2002) provides a good overview of current trends here. Flat structures also greatly simplify the task of writing translation rules which define correspondences across widely differing language pairs.

In the rest of the paper, we describe the representational solution we have developed for MedSLT. Linguistic representations are flat enough that it is simple to write phrase-spotting patterns and translation rules, but sufficiently expressive that they can capture all the key concepts of the domain. In the following sections, we first present the system; we then explain our approach focussing on the temporal and causal constructions. The central idea is to reduce causal concepts to temporal ones, which greatly simplifies the range of concepts that needs to be represented.

2 The MedSLT system

This section provides a brief overview of the current MedSLT prototype. The system is built on top of the Nuance toolkit platform (Nuance, 2003), and offers speech-to-speech translation from English into French, Japanese and Finnish[1]. It supports three separate medical diagnosis subdomains (headaches, chest pain, and abdominal pain) well enough that the full range of routine examination questions for each subdomain is covered. The vocabulary for each subdomain is between 300 and 450 words.

Translation is one-way in the doctor to patient direction, which means that most communication is in the form of yes-no questions that can be answered non-verbally. The system has a limited notion of dialogue context, so that it is possible to ask elliptical follow-on questions. For example, if the preceding question was "Is the pain sharp?", then "dull?" will be interpreted as "Is the pain dull?". Supporting ellipsis compensates to some extent for the restriction to yes-no questions. Instead of asking a single WH-question ("Where is the pain?", the doctor can ask an initial yes-no question with a series of elliptical follow-ups ("Is the pain in the front of the head?"... "The back of the head?"... "The left side?"... "The right side?"[2].

[1] Versions with French, Japanese and Spanish input and Spanish output are in various stages of preparation.

[2] The system does in fact also support WH-questions, since several doctors said they would like the option of using them as introductions to yes-no questions: "Where is the pain?"... "Is it in the front of the head?"

There are two versions of the system, using different speech understanding components. At the start of the project, we felt that there was a case to be made for using grammar-based recognition methods. Initially, we had no training data for creating statistical language models; also, the system is designed for expert users, and an earlier study we had been involved in (Knight *et al.*, 2001) suggested that grammar-based recognition can be more suitable for this type of user. These arguments are obviously not particularly strong. We wanted to be able to compare grammar-based speech understanding with a more standard architecture based on statistical language modelling and robust parsing, and have the option of reverting to the standard architecture if that seemed appropriate. In particular, this implied that source-language semantic representations needed to be such that they could reasonably be produced using phrase-spotting techniques.

In the grammar-based version, speech recognition uses a set of CFG-based language models (one per subdomain), compiled, using the REGULUS 2 toolkit, from a single linguistically motivated unification grammar (Rayner *et al.*, 2003b; Regulus, 2005). This makes it possible to support efficient structure-sharing between many similar subdomains with overlapping vocabulary and structure. Each subdomain-specific grammar is defined by a small training corpus, typically containing 500 to 1000 examples. The same corpus material is also used to perform probabilistic tuning of the resulting CFG language model. The statistical/robust version uses a normal class N-gram language model built using the Nuance SayAnything© package, together with a set of phrase-spotting rules. (Rayner *et al.*, 2004) reports experiments in which we compare performance for the two different versions of the system.

Both versions of the system use the same translation engine. Translation is interlingual and rule-based. Target language generation is also performed using suitably compiled linguistically motivated unification grammars. Output speech is produced using either a commercial TTS engine or concatenated recorded wavfiles, depending on the language.

3 Translating temporal and causal constructions

Initial versions of the MedSLT system (Rayner *et al.*, 2003a) used a completely flat representation format and a transfer-based translation architecture. For example, the English query "does the pain radiate to the jaw?" was represented as

```
[[utterance_type,ynq],[symptom,pain],[state,radiate],
 [tense,present],[prep,to_loc],[body_part,jaw]]
```

The Japanese translation "ago made itami wa hirogarimasu ka" (jaw-to pain-TOPIC radiate-POLITE-PRES Q) is represented as

```
[[utterance_type,sent],[symptom,itami],[state,hirogaru],
 [tense,present],[prep,made],[body_part,ago]]
```

When the scheme works, as it does here, the advantages are apparent: although the source and target versions have fairly different syntactic structures, the elements of the flat representations are in one-to-one correspondence. Transfer can be effected in a straightforward compositional fashion, and the constrained nature of the domain ensures that only valid target language translations can be produced from the target representation.

Problems arise, however, for causal and temporal constructions. For structurally similar languages, the same kind of solution tends to work reasonably well. For example, the representation of the English query "is the pain aggravated by coughing?" is

```
[[utterance_type,ynq],[symptom,pain],[event,aggravate],
 [tense,present],[cause,coughing]]
```

This can be translated into French as "la douleur est-elle aggravée par la toux?", which is represented similarly as

```
[[utterance_type,ynq],[symptom,douleur],[event,aggraver],
 [tense,present],[cause,toux]]
```

For unrelated language-pairs, this kind of solution is much more problematic. Although a literal translation of "is the pain aggravated by coughing?" into Japanese is not completely impossible, natural translations will not use a verbal construct corresponding to "aggravate", or a nominal construct corresponding to "coughing". It is instead preferable to use a subordinating conjunction construction, for example "seki wo suru to itami wa hidoku narimasu ka" (cough-OBJ make when pain-TOPIC worse become Q)[3].

Examples like these create a dilemma. Flat key/value representations are very suitable for robust phrase-spotting architectures, but there is no good way to handle a construction like a subordinate clause using a flat representation; both syntactically and semantically, a subordinate clause is clearly a nested structure. Unfortunately, the nature of the medical diagnosis domain means that temporal and causal constructions are extremely common. We have already seen "aggravate"; other typical examples are "relieve" ("does massage relieve the headache?"), "cause" ("is the headache caused by stress?"), "precede" ("is the headache preceded by nausea?") and "associated with" ("is the headache associated with vomiting?"). In English, too, natural phrasing often requires use of a subordinating conjunction. Although it is possible to say "is the pain relieved by lying down?", many people would prefer "is the pain better when you lie down?"

When we realised how important these phenomena were, our first reaction was to conclude that flat feature/value representations were simply inappropriate to a domain as complex as medical diagnosis questions: perhaps it was necessary to use general nested representations instead. If this were true, it would greatly complicate implementation of both the speech understanding and translation components of the system.

Further analysis, however, convinced us that this view of the situation was too extreme, and that a sensible compromise solution existed between the opposing positions of flat feature/value lists and general nested structures. Most importantly, we can in the context of this task reduce all temporal and causal relationships to one of the following canonical schemas: (1) [WHEN] Clause1 WHEN Clause2; (2) [BEFORE] Clause1 BEFORE Clause2; (3) [AFTER] Clause1 AFTER Clause2.

Figure 1 shows examples of how different concepts can be paraphrased in this way. Our new strategy then became the following: move to an interlingual translation architecture, and use the canonical versions of the temporal and causal relations as the interlingual representation.

Note that we are in no way claiming that temporal and causal relationships can be conflated in general; in many other contexts, we would certainly have to distinguish them. What we

[3]In practice, "itami wa" (pain-TOPIC) would often be omitted, since the topic is clear from context.

is the headache **aggravated** by bright light? →
headache **is worse WHEN** you are exposed to bright light?

does massage **relieve** the headache? →
headache **is better WHEN** you receive massage?

does stress **give** you headaches? →
you **have** headache **WHEN** you are stressed?

is the headache **associated with** vomiting? →
you vomit **WHEN** you **have** a headache?

is the headache **accompanied** by nausea? →
you experience nausea **WHEN** you **have** a headache?

is the headache **preceded** by scintillations? →
you experience scintillations **BEFORE** you **have** a headache?

do you get headaches **after** a large meal? →
you have headache **AFTER** you eat a large meal?

Figure 1: Examples of reducing causal and temporal concepts to canonical form

are doing, rather, is exploiting the constraints of the medical diagnosis task to simplify the semantic representation language. In this very specific context, the justification for replacing causal questions with temporal ones is that the patient will not normally know what causes the symptoms, even if they believe they do — they only know about the temporal sequence of events. For this reason, the doctor will not receive misleading information if the patient answers the temporal question, irrespective of whether it was originally phrased as temporal or causal.

At the level of concrete representations, we conservatively extend the representation language by allowing one level of nesting in the key/value lists, so as to make it possible to represent the subordinate clause construction. Thus for example the representation of "do you have headaches when you drink coffee?" is

```
[[utterance_type,ynq],[pronoun,you],[state,have_symptom],
 [tense,present],[symptom,headache],[sc,when],
 [[clause,[[utterance_type,dcl],[pronoun,you],
  [action,drink],[tense,present],[cause,coffee]]]]
```

We have carefully chosen the above example so that the source and interlingua representations are in this case identical; in other words, "do you have headaches when you drink coffee?" is the canonical way to say this question. Following Figure 1, we design the rules which map source language representations to interlingua so that we get the same interlingual form for other phrasings of the same question. For example, "are your headaches caused by coffee?", with source representation

```
[[utterance_type,ynq],[symptom,headache],[substance,coffee],
 [event,cause],[tense,present]]
```

and "does coffee give you headaches?", with source representation

```
[[utterance_type,ynq],[symptom,headache],[substance,coffee],
 [event,give],[tense,present]]
```

will both yield the same interlingual form as "do you have headaches when you drink coffee?"

In order to realise the scheme we have just sketched out, we had to solve two main technical problems. First, we needed to be able to produce nested source language representations for utterances containing subordinate clauses. Second, we required a clean way to structure the rules which map source language representations into interlingual ones. We consider these two sets of issues separately.

3.1 Producing nested source language representations

Producing nested representations in the grammar-based version of the recogniser is straightforward: these can be built up in the usual way using compositional semantics. The challenge is to produce them in the version of the system which uses statistical recognition, where we are limited to robust surface processing on a noisy recognition string. This section briefly describes our implemented solution.

Processing consists of three phases. First, a set of rules is applied that attempts to detect start- and end-boundaries for subordinate clauses. A typical rule in this group[4] is

```
boundary([when],[not_word(do/does/have/has/can)],start).
```

This guesses the start of a subordinate clause after the word "when", and before a word that is not one of the words "do", "does", "have", "has" or "can".

Once the recognition string has been segmented into clauses, a second set of rules is applied, to guess key/value pairs. A typical rule in the second group is

```
pattern([lean/leaning,forward],[action,lean_forward]).
```

This guesses the key/value pair `[action,lean_forward]` if a sequence is found consisting of the word "lean" or "leaning" followed by the word "forward".

Finally, a set of post-processing rules is applied, which fills in default values for unset features in the representation of each clause. For example, `tense` is by default set to `present`, and `utterance_type` to `ynq` if a verb is present, and `phrase` otherwise.

3.2 Mapping temporal and causal concepts into canonical form

Translation rules in the MedSLT system are implemented using the Prolog-based formalism defined by the Regulus toolkit (Rayner *et al.*, 2005). Basically, this allows definition of rules mapping lists of key/value pairs to lists of key/value pairs. Lists can optionally contain up to

[4]The form of the rules has been simplified slightly for presentational purposes.

one level of nesting, using the `[clause, ...]` representation of subordinate clauses shown above. Rules may be conditional on the presence or absence of partially specified elements in the rest of the list; there is also support for use of macros. Macros may be non-deterministic, in which case the rule expands into multiple copies. All these features are illustrated in the following (artificial) example,

```
transfer_rule([[polarity,OnOff]], [[event, @onoff(OnOff)]])
    :- context([event,switch]).

macro(onoff(on), switch_on).
macro(onoff(off), switch_off).
```

This says that the lists `[[polarity,on]]` and `[[polarity,off]]` are respectively mapped to the lists `[[event,switch_on]]` and `[[event,switch_off]]` in a context which also contains the key/value pair `[event,switch]`.

We now describe how we realise within this framework the types of transformation informally sketched in Figure 1. Comparing the left- and right-hand sides of the examples, we can see that the changes involved are of two types. On the one hand, the portions marked in bold pick out transformations associated with causal relations. Thus, informally, we transform "aggravated by" into "is worse when". Alongside these, we have transformations which map nominal concepts into associated verbal counterparts; so, again informally, we map "bright light" into "be exposed to bright light". The problem we need to solve here is how to structure the rule-base so that these two groups of rules can be kept separate and thus orthogonal.

The solution we have implemented is to realise the nominal-to-verbal transformations as macros, and the causal-to-temporal transformations as rules using those macros. The following typical rule (slightly simplified) handles a transformation which could be informally described as "X causes (symptom)" to "you have (symptom) when X occurs":

```
transfer_rule(
    [Noun,[event,cause]],
    [[state,have_symptom],[sc,when],
     [clause,
      [[utterance_type,dcl],[pronoun,you],[tense,present],
        @causal_noun_to_vp(Noun)]]])
  :- context([symptom,_]).
```

This operates in a context where there is an element matching `[symptom,_]` in the environment. The left-hand side is a list consisting of `[event,cause]` together with a causal noun; the right-hand side is a list containing the key/value pairs `[state,have_symptom]`, `[sc,when]`, and a subordinate clause where the subject is "you" and the verb-phrase is the verbal counterpart of the noun on the left-hand side.

The non-deterministic macro `causal_noun_to_vp` contains one definition for each causal noun. Typical entries are

```
macro(causal_noun_to_vp([[substance,tea]]),
      [[action,drink],[substance,tea]]).
```

```
macro(causal_noun_to_vp([[substance,large_meal]]),
      [[action,eat],[substance,large_meal]]).
macro(causal_noun_to_vp([cause,massage]),
      [[state,experience],[cause,massage]]).
```

The first two entries are obvious: the nominal concepts "tea" and "large meal" map into the verbal concepts "drink tea" and "eat large meal". The third entry shows another common pattern. In many cases, the verb associated with the nominal concept is semantically neutral; we represent this using the key/value pair `[state,experience]`. The rules which map from the interlingua to the target language may give `[state,experience]` a more specific lexical realisation. Thus for example when moving from interlingua to English we map `[[state,experience], [cause,massage]]` to "receive massage"; if the target language is Japanese, it is mapped to the neutral "massaaji suru" (do massage).

Although this representational scheme is quite simple, we have been surprised to see what a wide range of complex translation mismatches it can handle. One particularly interesting case concerns WH-pronouns. These are represented similarly to other causal concepts, so for example "what relieves your headaches?" is represented as

```
[[utterance_type,whq],[spec,what],[event,relieve],
 [tense,present],[symptom,headache]]
```

We map this into interlingual form by simply adding another definition of `causal_noun_to_vp`,

```
macro(causal_noun_to_vp([spec,what]),
      [[state,experience],[spec,what]]).
```

For Japanese, `[[state,experience], [spec,what]]` can be mapped directly into the expression "nani wo suru" (do what); thus we translate "what relieves your headaches?" into the quite natural "nani wo suru to zutsu ga osamarimasu ka" (what-OBJ do when headache-SUBJ get-better-Q). A detailed evaluation of the performance of the system can be found in (Rayner *et al.*, 2004).

4 Summary and conclusions

We have presented an overview of MedSLT, a medium vocabulary medical speech translation system, focussing on the representational issues that arise when translating temporal and causal concepts. Although flat key/value structures are strongly preferred as semantic representations in speech understanding systems, we argue that it is infeasible to handle the necessary range of concepts using only flat structures.

By exploiting the specific nature of the task, we have shown that it is possible to implement a solution which only slightly extends the representational complexity of the semantic representation language, by permitting an optional single nested level representing a subordinate clause construct. We have sketched our solutions to the key problems of producing minimally nested representations using phrase-spotting methods, and writing cleanly structured rule-sets that map temporal and phrasal representations into a canonical interlingual form.

References

BLACK A., BROWN R., FREDERKING R., SINGH R., MOODY J. & STEINBRECHER E. (2002). TONGUES: Rapid development of a speech-to-speech translation system. In *Proceedings of HLT: Human Language Technology Conference*.

KNIGHT S., GORRELL G., RAYNER M., MILWARD D., KOELING R. & LEWIN I. (2001). Comparing grammar-based and robust approaches to speech understanding: a case study. In *Proceedings of Eurospeech 2001*, p. 1779–1782, Aalborg, Denmark.

LAVIE A., LANGLEY C., WAIBEL A., PIANESI F., LAZZARI G., COLETTI P., TADDEI L. & BALDUCCI F. (2001). Architecture and design considerations in NESPOLE!: a speech translation system for e-commerce applications. In *Proceedings of HLT: Human Language Technology Conference*, San Diego, California.

LAVIE A., WAIBEL A., LEVIN L., FINKE M., GATES D., GAVALDA M., ZEPPENFELD T. & ZHAN P. (1997). JANUS-III: Speech-to-speech translation in multiple languages. In *Proceedings of the 6th European Conference on Speech Communication and Technology (EuroSpeech99)*, p. 99–106.

NUANCE (2003). http://www.nuance.com. As of 25 February 2003.

PHRASELATOR (2004). http://www.phraselator.com. As of 8 Dec 2004.

RAYNER M. & BOUILLON P. (2002). A phrasebook style medical speech translator. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (demo track)*, Philadelphia, PA.

RAYNER M., BOUILLON P., HOCKEY B., CHATZICHRISAFIS N. & STARLANDER M. (2004). Comparing rule-based and statistical approaches to speech understanding in a limited domain speech translation system. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation; also ftp://issco-ftp.unige.ch/pub/publications/tmi_045.pdf*, Baltimore, MD.

RAYNER M., BOUILLON P., VAN DALSEM V., HOCKEY B., ISAHARA H. & KANZAKI K. (2003a). A limited-domain English to Japanese medical speech translator built using REGULUS 2. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (demo track)*, Sapporo, Japan.

M. RAYNER, D. CARTER, P. BOUILLON, V. DIGALAKIS & M. WIRÉN, Eds. (2000). *The Spoken Language Translator*. Cambridge University Press.

RAYNER M., HOCKEY B. & BOUILLON P. (2005). *Using Regulus*. http://cvs.sourceforge.net/viewcvs.py/regulus/Regulus/doc/RegulusDoc.htm. As of 30 January 2005.

RAYNER M., HOCKEY B. & DOWDING J. (2003b). An open source environment for compiling typed unification grammars into speech recognisers. In *Proceedings of the 10th EACL (demo track)*, Budapest, Hungary.

REGULUS (2005). http://sourceforge.net/projects/regulus/. As of 30 January 2005.

W. WAHLSTER, Ed. (2000). *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer.

YOUNG S. (2002). Talking to machines (statistically speaking). In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*, p. 9–16, Denver, CO.

Sur le statut référentiel des entités nommées

Thierry Poibeau

Laboratoire d'Informatique de Paris Nord – CNRS et Université Paris 13
99, av. J.-B. Clément – F-93430 Villetaneuse
thierry.poibeau@lipn.univ-paris13.fr

Mots-clés : Entités nommées, référence, sémantique lexicale
Keywords: Named entities, reference, lexical semantics

Résumé Nous montrons dans cet article qu'une même entité peut être désignée de multiples façons et que les noms désignant ces entités sont par nature polysémiques. L'analyse ne peut donc se limiter à une tentative de résolution de la référence mais doit mettre en évidence les possibilités de nommage s'appuyant essentiellement sur deux opérations de nature linguistique : la synecdoque et la métonymie. Nous présentons enfin une modélisation permettant de rendre explicite les différentes désignations en discours, en unifiant le mode de représentation des connaissances linguistiques et des connaissances sur le monde.

Abstract We show in this paper that, on the one hand, named entities can be designated using different denominations and that, on the second hand, names denoting named entities are polysemous. The analysis cannot be limited to reference resolution but should take into account naming strategies, which are mainly based on two linguistic operations: synecdoche and metonymy. Lastly, we present a model that explicitly represents the different denominations in discourse, unifying the way to represent linguistic knowledge and world knowledge.

1 Introduction

Les entités nommées désignent les noms de personnes, de lieux, d'organisations mais aussi les dates ou les unités monétaires. Alors que ces éléments ont longtemps été délaissés par les systèmes de traitement automatique des langues, le renouveau lié au travail sur corpus a révélé qu'il s'agissait en fait d'éléments majeurs pour l'analyse. Les conférences en extraction d'information (*Message Understanding Conferences*, cf. MUC-7, 1998) ont mis en avant plusieurs tâches génériques, au premier rang desquelles l'analyse des entités nommées. Il s'agit en fait d'une double tâche : d'une part une tâche de reconnaissance des séquences pertinentes, d'autre part une tâche de typage des séquences ainsi reconnues en fonction d'une ontologie pré-établie. Sur des textes de type journalistique, les systèmes obtiennent

généralement d'assez bons scores, avec un taux combiné de rappel et de précision supérieur à 0,90.

Les entités nommées sont particulièrement importantes pour l'accès au contenu du document car elles forment les briques élémentaires sur lesquelles repose l'analyse. Les entités sont généralement considérées comme directement référentielles : ce sont les désignateurs rigides de Kripke (1982) qui font référence aux objets du monde, organisés en ontologie. Nous montrons ici les limites de cette approche et nous contestons la vue traditionnelle qui rend compte d'une façon très simplifiée de la complexité de la langue. Nous proposons quelques pistes pour mieux prendre en compte le sémantisme complexe des entités.

Nous présentons dans un premier temps les systèmes classiques de repérage d'entités nommées. Ceux-ci reposent sur un ensemble de règles permettant de typer les séquences pertinentes d'après un ensemble de catégories prédéfinies. Nous montrons ensuite les limites de cette approche : la plupart des entités sont polysémiques et leur analyse dépend étroitement du contexte. Partant de ce constat, l'analyse mise en œuvre ne peut qu'être dynamique et refléter les effets de sens en contexte. Nous proposons enfin, dans une dernière partie, une modélisation à base de structures de traits permettant de mieux rendre compte des phénomènes en jeu.

2 Systèmes de repérage et de catégorisation des entités nommées

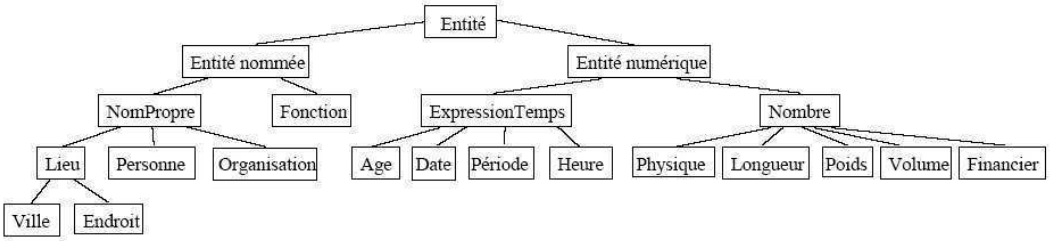
Les entités nommées sont traditionnellement typées suivant une hiérarchie pré-établie. Nous examinons ce type de hiérarchie et les problèmes posés par les textes.

2.1 Hiérarchies de types d'entités

Sous l'influence des conférences américaines d'évaluation MUC, les travaux en extraction d'entités nommées ont traditionnellement été effectués sur des textes journalistiques ou des dépêches d'agence. L'identification des entités nommées inclut trois types d'expressions :

- ENAMEX : les expressions de noms propres incluant les noms de personnes, de lieux et d'organisations.
- TIMEX : les expressions temporelles comme les dates et les heures.
- NUMEX : les expressions numériques telles que les expressions monétaires et les pourcentages.

Voici par exemple la hiérarchie de types définie par le LIMSI pour le système QALC (Ferret *et al.*, 2001) et calqué sur la hiérarchie définie pour MUC (Grishman et Sundheim, 1995).



Les hiérarchies comportent ainsi, pour les plus simples, une douzaine de types de base (feuilles de la hiérarchie) mais ont souvent besoin d'être étendues pour couvrir de nouveaux

besoins (de nouvelles tâches, de nouveaux domaines). Il n'est ainsi pas rare de faire face à des hiérarchies de plus de 200 éléments (Sekine, 2004).

2.2 Repérage et classification des entités nommées

De nombreux travaux ont porté sur l'identification des noms propres dans des textes journalistiques, notamment les *Message Understanding Conferences* (MUC7, 1998). La reconnaissance des entités nommées à partir de textes écrits est actuellement la tâche d'extraction d'information qui obtient les meilleures performances. Les performances sont mesurées en utilisant des mesures classiques comme P&R, fondée sur un taux combiné de précision et de rappel suivant la règle suivante : $P\&R = \frac{2 * PRECISION * RAPPEL}{PRECISION + RAPPEL}$. Les taux obtenus sont comparables à ceux des humains, de l'ordre de 0,90 P&R sur des dépêches journalistiques.

Deux grandes approches sont généralement suivies pour leur identification : une approche linguistique « de surface » et une approche probabiliste. L'approche linguistique est fondée sur la description syntaxique et lexicale des syntagmes recherchés. Des règles de grammaire utilisent des marqueurs lexicaux (ex. *Mr* pour *Mister* ou *Inc.* pour *Incorporated*), des dictionnaires de noms propres et des dictionnaires de la langue générale (essentiellement pour repérer les mots inconnus) sont utilisés pour repérer et typer les syntagmes intéressants (Aberdeen *et al.*, 1995 ; Grishman, 1995 ; Appelt et Martin, 1999). De son côté, l'approche probabiliste utilise un modèle de langage entraîné sur de larges corpus de textes pré-étiquetés. Cette approche est particulièrement robuste lorsque les textes sont bruités, c'est pourquoi la grande majorité des systèmes dédiés à l'oral adoptent une telle approche (Kubala, 1999).

3 Difficultés et limites de la catégorisation

Nous montrons dans cette partie les limites de l'approche traditionnelle. Nous montrons notamment la polysémie fréquente des entités nommées et les conséquences pour l'analyse automatique.

3.1 Polysémie des entités nommées

La catégorisation des entités repose en grande partie sur l'hypothèse que les entités sont référentielles et peuvent donc recevoir un type rendant compte de leur référent. Cette analyse est le plus souvent erronée comme on peut aisément le montrer, les entités nommées référant le plus souvent à plusieurs classes (Leroy, 2004). Les dates se confondent ainsi fréquemment avec des événements :

> *Le 11 septembre 2001 a représenté un tournant dans l'histoire américaine. (Elie Wiesel, site www.france-amerique.com)*

Il est parfois difficile de classer les noms d'organisations, qui peuvent être catégorisés comme institution, communauté d'individus ou encore bâtiment.

> *Le journal télévisé a eu lieu hier en direct de l'ONU.*
> *L'ONU était en grève hier.*
> *L'ONU a fêté ses 50 ans.*
> *L'ONU n'acceptera pas une attaque frontale de l'Irak (forum du Monde)*

On retrouve le même phénomène pour les noms de lieu :

L'Europe veut garder la tête du FMI. (Libération, 10 mars 2004)

Les noms de personnes sont encore plus labiles. Nous laisserons de côté les exemples où le nom de personne est en fait devenu la désignation d'une entreprise, d'un stade ou d'un lieu quelconque.

Une rencontre d'un niveau technique assez médiocre à l'Abbé Deschamps. (stade d'Auxerre, Journal L'équipe)

Même sans prendre en compte ces phénomènes de transfert de sens, les exemples de catégorisation changeante sont légion. Un nom de personne peut référer à une œuvre, à un objet ou à tout autre élément ayant un lien direct ou non avec la personne.

J'ai tout Chirac sur l'étagère
Pierre est garé en face. (cf. Cadiot et Visetti 2001, p. 167)

Il existe par ailleurs des phénomènes plus rares mais bien attestés où un nom propre, particulièrement un nom de la personne, ne renvoie pas au référent traditionnel. C'est par exemple le cas de l'antonomase, largement étudiée par S. Leroy (2003) : *mon oncle est un vrai Harpagon* ne réfère pas à un personnage réel s'appelant Harpagon.

On voit à travers ces exemples rapides que les entités nommées ne se comportent pas fondamentalement différemment des autres unités linguistiques. L'exemple le plus connu est certainement celui du *Prix Goncourt* introduit initialement par Kayser (1988). Celui-ci distingue plusieurs sens différents, suivant qu'il s'agisse du prix, de sa valeur monétaire, du livre qui a obtenu le prix, de l'institution ou encore de l'auteur.

3.2 Référentialité des entités nommées

Nous retrouvons sous ces thèmes déjà évoqués à de multiples reprises dans la littérature, le questionnement habituel sur la notion de polysémie. Comment traiter ces cas ? Leurs sens sont-ils mutuellement incompatibles ou peut-on définir une théorie unificatrice du sens rendant compte de ces multiples emplois ?

Les entités nommées constituent à cet égard une classe de syntagmes proche des termes. De nombreux travaux étrangers ne font d'ailleurs le plus souvent pas de différences entre ces éléments (Collier *et al.*, 2000), d'autant que certains domaines contribuent à rendre floues les distinctions qui semblent classiques par ailleurs. Dans une phrase issue d'un corpus de biologie comme *Abd-a interagit avec trx*, a-t-on affaire à des entités nommées ou à des termes ? *Abd-a* constitue-t-il un nom de gènes ou une classe ? Les réponses ne sont pas toujours claires. Le propre des entités nommées est malgré tout d'être référentielles. C'est essentiellement ce trait qui définit leur nature et qui les distingue des termes. Leurs propriétés référentielles ont été bien étudiées, depuis Frege (1892) et Kripke (notion de désignateur rigide, cf. Kripke, 1982), jusqu'aux travaux récents dans des cadres linguistiques divers (Charolles, 2002).

On remarque cependant la même variation au niveau des entités nommées qu'au niveau des termes. Il existe de multiples façons de désigner une personne ou un objet, il n'y a pas de nom unique et inévitable (cf. l'exemple de Frege autour de l'*étoile du soir* et *l'étoile du matin* qui désignent toutes deux Vénus …). On peut dès lors s'interroger sur la nature référentielle de l'entité nommée employée en discours. Il ne s'agit bien évidemment pas de nier les possibilités référentielles de l'entité nommée mais il est permis de se demander si l'emploi

d'une entité nommée en discours ne laisse pas de côté son statut référentiel : celui-ci serait en quelque sorte dans un état latent. Concrètement, cela signifie qu'en discours il n'y a pas de réelle différence entre une affirmation portant sur une entité, sur un terme ou sur un autre groupe nominal. Comme le rappelle Charolles (2002, p. 7-10), dans la littérature linguistique, les auteurs distinguent la dénotation de la référence. La dénotation « prend acte du fait que le signifié (des signes linguistiques) exprime un concept muni d'une certaine intension (i.e. un ensemble d'attributs caractéristiques en propre) et que cette intension délimite virtuellement son extension (c'est-à-dire une classe d'êtres satisfaisants ces attributs) ». L'acte de référence est quant à lui « un accord non entre deux pensées (…) mais entre deux pensées à propos de quelque chose et cela par le biais de la production, en contexte, d'une expression référentielle ». Nous reprenons à notre compte l'essentiel de ces définitions mais à l'inverse de Charolles, nous soutenons que l'on peut alors parler d'opération linguistique pour la dénotation, contrairement à la référence qui ressortit à une logique visant à apparier niveau linguistique et niveau extra-linguistique.

En discours, l'aspect dénotationnel est primordial, passant bien avant l'aspect référentiel. Cette distinction n'est pas seulement d'ordre philosophique : il s'agit d'une dimension essentielle de l'interprétation du texte dans la mesure où la dénotation, à l'inverse de la référence, peut supporter une sous-spécification par défaut[1]. Ainsi, dans l'exemple de biologie ci-dessus, la résolution du fait que *Abd-a* réfère à un gène particulier ou à une classe n'a pas lieu d'être dans la mesure où cela est inutile pour l'interprétation de la phrase. Le fait que l'*étoile du soir* et l'*étoile du matin* désignent la même planète n'est pas fondamental au niveau linguistique mais est en revanche essentiel pour assigner une valeur de vérité aux assertions émises sur ces objets[2]. On rejoint ici le point de vue de Searle (1983) pour qui « l'acte de référence, le processus de référenciation », inclut une dimension externe à la langue « à savoir que l'entité visé par cet acte, la chose (au sens le plus général du terme) à laquelle il renvoie, existe au-delà de ce que le locuteur en dit, en dehors de son esprit et de celui de celles ou ceux à qui il s'adresse ». (Charolles, 2002, p. 38)

3.3 Catégorisation à partir d'une ontologie

La plupart des auteurs qui se sont intéressés à la question du traitement automatique de la métonymie ou de la synecdoque proposent des approches fondées sur des ressources générales comme Wordnet (cf. l'atelier *The lexicon and figurative language* durant ACL 2003 [Wellington, 2003]). Il est en effet difficile d'acquérir directement à partir du corpus des informations comme une décomposition des objets en relation *partie-tout*. L'analogie peut jouer un rôle et permettre le repérage de relations de dépendance spécifiques (voir [Nehaniv, 1999] ou [Lepage, 2003]) mais ceci reste toutefois marginal dans la plupart des cas.

Les exemples et les remarques de la section précédente semblent toutefois aller dans le sens de la théorie défendue par Cadiot et Visetti (2001), qui contestent l'existence préalable d'une

---

[1] On retrouve ici la notion de profondeur variable de Kayser et Coulon [1981], que nous interprétons ainsi : l'analyse n'a pas à se préoccuper d'un certain nombre de traits sémantiques latents tant que ceux-ci ne sont pas activés de manière explicite. D'une manière plus globale, il n'est pas toujours nécessaire de chercher à désambiguïser finement le rôle de l'entité suivant la tâche ou l'application visée.

[2] Les cas où l'analyse de la référence est obligatoire sont somme toute relativement rares pour les applications de traitement automatique des langues. Le cadre privilégié est la commande de système, la communication avec des robots ou tout autre cadre où la parole entraîne une action très directe sur le monde.

ontologie, notamment pour les groupes nominaux déterminés et plus généralement, pour les unités dites référentielles. La notion de « facette » (Cruse, 2004) rend partiellement compte de l'aspect polysémique des entités envisagées mais plusieurs auteurs ont souligné que les facettes ne rendent pas compte des liens entre les différents sens envisagées. Cadiot et Visetti montrent bien qu'il s'agit essentiellement de phénomènes de synecdoque et de métonymie, qui ruinent toute tentative directement référentielle mais qui appellent plutôt une analyse dynamique par « profilage » de sens en fonction du contexte, pour reprendre une partie de la terminologie employée par les deux auteurs (sur ces questions, voir aussi Fass, 1988). Chibout (1999) donne une typologie claire de plusieurs figures de style impliquant des liens sémantiques complexes entre unités linguistiques mises en jeu. Une modélisation par un ensemble de traits activables en contexte semble particulièrement approprié.

Plusieurs auteurs ont proposé des modèles informatiques visant à résoudre des problèmes relativement similaires aux nôtres. La plupart se situent dans le cadre du dialogue homme-machine et s'en tiennent à identifier les problèmes de référence qui se posent lors de l'analyse (Salmon-Alt, 2001). Citons toutefois l'étude de Pospescu-Belis *et al.* (1999) qui présente un modèle de résolution des anaphores fondé sur le modèle des représentations mentales. L'auteur évoque la notion de point de vue et la variabilité de la référence dans un cadre conversationnel. Il faut d'ailleurs souligner le rôle de la tâche pour ce type d'analyse.

Il semble toutefois difficile de se passer de toute catégorisation préalable, de se fonder uniquement sur une approche constructiviste, voire herméneutique, à partir d'unité de sens activables (ou non) en contexte[3]. Les catégories prédéterminées souffrent des mêmes défauts que les facettes de Cruse : l'analyse ne dit rien des rapports entre date et événement, et, plus généralement, n'explique pas le continuum qui existe entre les différents aspects sémantiques des entités en contexte. Mais au moins une telle analyse ne préjuge pas de la sémantique de l'entité en contexte, elle laisse ouverte une liste de possibles. On laissera à la sémantique lexicale, voire infra lexicale, le soin de mettre au jour les liens entre traits activables au sein du mot.

4 Proposition de modélisation

Nous gardons de ce qui précède l'idée d'hypothèses construites en parallèle, pouvant être activées ou non en contexte. Nous nous inspirons des structures de traits à la DATR (Evans et Gazdar, 1996) et du lexique génératif (Pustejovsky, 1995) pour coder les différents traits activables pour les entités.

4.1 Modélisation sous forme de structures de traits

La hiérarchie de types définie dans le cadre des conférences MUC et présentée dans la première partie de cette étude a prouvé son efficacité. Il nous semble donc pertinent de nous appuyer sur celle-ci pour définir des types de base.

La plupart des propositions qui ont par la suite été faites pour raffiner cette hiérarchie ou d'autres catégorisations ayant une granularité comparable (Sekine, 2004) ont été confrontées

[3] Une proposition allant toutefois dans ce sens est celle de Sabah (1996) qui propose un modèle informatisé appelé « carnet d'esquisses ». On retrouve sous ce terme une vue dynamique de la sémantique en train de se construire de manière dynamique, en fonction du co-texte et du contexte.

aux problèmes de polysémie soulevés ci-dessus. Pour reprendre un exemple déjà entrevu : comment faire un choix parmi les différents sens du mot ONU, a-t-on affaire à l'institution, au bâtiment ou à l'ensemble des personnes qui composent l'organisme ? Il s'agit essentiellement d'un mécanisme de focalisation permettant de mettre en avant un aspect de l'entité en contexte. Nous proposons donc d'adjoindre un trait *saillance* aux entités afin de mettre en avant de manière explicite cette focalisation qui peut changer de manière dynamique en fonction du contexte. Reprenons quelques uns des exemples déjà traités :

L'ONU n'acceptera pas une telle décision.

```
Entity{
    Lexical_unit=ONU;
    Sem{
        Type=organization;
        Focalisation=diplomatic_org; }
}
```

L'ONU désigne dans ce cadre l'organisation politique. Cette information peut être repérée à partir d'informations sur le monde, sur le rôle de l'ONU sur la scène diplomatique et sur l'analyse du groupe verbal qui suit l'entité. On peut rapprocher cette information du rôle *télique* défini par Pustejovsky (1995) : le propre de l'ONU, c'est de prendre et de faire appliquer des décisions d'ordre diplomatique.

Le journal télévisé a eu lieu en direct de l'ONU.

```
Entity{
    Lexical_unit=ONU;
    Sem{
        Type=organization;
        Focalisation=location; }
}
```

Ici le verbe *avoir lieu* fait clairement allusion à un processus en train de se dérouler dans un lieu donné. La focalisation porte donc sur la localisation de l'événement.

L'ONU était en grève hier.

```
Entity{
    Lexical_unit=ONU;
    Sem{
        Type=organization;
        Focalisation=human_org; }
}
```

Le même phénomène opère ici. La notion de *grève* active la notion d'organisation composée d'individus (*human_org*). Le trait de focalisation est donc mis à jour. L'exemple qui suit est à cet égard moins clair :

L'ONU a fêté ses 50 ans.

```
Entity{
    Lexical_unit=ONU;
    Sem{
        Type=organization;
        Focalisation=none; }
}
```

L'auteur prête ici des traits humains à l'institution. Il ne s'agit plus d'une simple synecdoque mais d'un phénomène métaphorique, en tant que tel plus difficile à analyser automatiquement. D'un côté, le verbe *fêter* invite à voir une communauté d'individus dans le sujet. Mais c'est bien l'institution qui a 50 ans. Une focalisation sur la communauté d'individus serait quelque peu abusive dans ce cas. Peut-être faut-il s'en tenir à une certaine sous-spécification au niveau

de l'analyse automatique : l'institution et les individus qui la composent forment ici une unité syncrétique qui échappe partiellement à l'analyse.

4.2 Du formulaire d'entités à la résolution des anaphores nominales

L'analyse des entités suivant le modèle que nous proposons demande à avoir accès à des informations sur les entités. Ces informations peuvent être qualifiées de connaissances sur le monde ou encore de connaissances de sens commun. La façon d'encoder ces informations doit être générique mais il semble encore hors de portée de développer des systèmes à large couverture ayant une telle masse de connaissances sur le monde. Le plus grand projet connu allant dans ce sens est le projet CYC de Doug Lenat (Lenat et Guha, 1990) mais les articles de l'auteur ont depuis révélé que la tâche était sans doute hors d'atteinte et les réalisations ont jusqu'à maintenant été peu concluantes (Lenat, 2001).

L'analyse des entités demande une modélisation fine du domaine visé. Une tentative allant dans ce sens est celle des conférences d'extraction d'information à partir de textes MUC (Message Understanding Conferences). Plusieurs tâches génériques ont été définies dans le cadre de ces conférences, dont le remplissage de « formulaires d'entité ». Le formulaire permet de relier à une entité donnée un ensemble d'informations de natures diverses. En nous fondant sur les expériences menées dans le cadre de MUC, nous pouvons offrir par exemple une représentation enrichie de la notion d'*organisation*.

```
Entity{
    Lexical_unit=ONU;
    Sem{
        Type=organization;
        Focalisation=none;
    }
    EntityTemplate{
        IsLocatedIn = New_York;
        IsComposedOf = employees && diplomats;
        IsLeadedBy = Kofi_Annan;
        KindOf =diplomatic_org)
    }
}
```

Ce type de structures permet d'avoir accès aux différentes composantes de l'entité et permet d'expliquer certaines dénominations fondées sur la métonymie ou la mise en avant d'un des aspects de l'objet visé. Le modèle développé rend explicite les apparents changements de catégorie qui peuvent s'expliquer par une fonction d'accès à un des aspects de l'entité. Les outils d'extraction d'information (Poibeau, 2003) peuvent ici s'appliquer pour contribuer à produire ces ressources automatiquement. Ils demandent toutefois la mise au point de systèmes collaboratifs, dans la mesure où toute l'information ne peut pas être acquise directement à partir du corpus. Le peu de ressources disponibles pour le français, notamment au niveau sémantique, rend la tâche particulièrement difficile.

4.3 La question des anaphores nominales

Les mécanismes de mise en avant d'un aspect de l'entité en fonction du contexte restent à étudier. L'entité est constituée d'un ensemble de traits permettant d'avoir accès à ses différentes composantes. Il est ainsi possible d'expliquer certaines anaphores nominales qui résistent traditionnellement à l'analyse :

L'organisation de Kofi Annan…
Syn(L'organisation de Kofi Annan) = ONU

Justification: IsLeadedBy(ONU)=Kofi Annan

Ces mécanismes peuvent être modélisés au moyen de fonctions à la Melc'uk pour offrir une version unifiée du lexique, intégrant connaissances linguistiques et connaissances sur le monde. En ne cherchant pas à rendre explicite la référence de l'entité par rapport à un modèle du monde, notre modélisation se démarque nettement des travaux effectués dans le cadre de l'analyse du dialogue (Popescu-Belis, 1999 ; Salmon-Alt, 2001). Elle rend en revanche explicite les effets de sens au sein de l'entité, afin d'expliquer les phénomènes de synecdoque et de métonymie entrevus précédemment.

5 Conclusion

Nous avons essayé de montrer dans cet article qu'une même entité peut être désignée de multiples façons et que les noms désignant ces entités sont par nature polysémiques. L'analyse ne peut donc se limiter à une tentative de résolution d'anaphores mais doit mettre en évidence les possibilités de nommage s'appuyant essentiellement sur deux opérations de nature linguistique : la synecdoque et la métonymie. Nous avons enfin présenté une modélisation permettant de rendre explicite les différentes désignations en discours, en unifiant le mode de représentation des connaissances linguistiques et des connaissances sur le monde.

6 Références

ABERDEEN J., BURGER J., DAY D., HIRSCHMAN L, ROBINSON P., VILAIN M. (1995) « MITRE: Description of the Alembic System as Used for MUC-6 », In *Actes MUC-6*, Morgan Kaufmann Publishers, San Francisco.

APPELT, D., MARTIN D. (1999) « Named Entity Recognition in Speech: Approach and results using the TextPro System », *Proceedings of the DARPA Broadcast News Workshop*, Herndon, Virginia.

CADIOT P., VISETTI Y.-M. (2001) *Pour une théorie des formes sémantiques*, PUF, Paris.

CHAROLLES M. (2002) *La référence et les expressions référentielles en français*. Ophrys. Paris.

CHIBOUT K. (1999) *La polysémie lexicale : observations linguistiques, modélisation informatique, études ergonomique et psycho-linguistique*. Thèse de doctorat, Université 11.

COLLIER N., NOBATA C., TSUJII J. (2000) « Comparison between Tagged Corpora for the Named Entity Task », In *Actes 18th International Conference on Computational Linguistics* (COLING 2000), Saarbrücken, Allemagne. pp. 201-207.

CRUSE A. (2004) *Meaning in language, an introduction to semantics and pragmatics*, Oxford University Press, Oxford.

EVANS R. et GAZDAR G. (1996) « DATR : A language for lexical knowledge representation », *Computational Linguistics,* n°22(2). pp. 167-216.

FASS D. (1988). « Metonymy and metaphor: what's the difference? ». In *Actes 12th International Conference on Computational Linguistics* (COLING 1988), Budapest, Hungrie, pp. 177-181.

FERRET O., GRAU B., HURAULT-PLANTET M., ILLOUZ G., MONCEAUX L., ROBBA I., VILNAT A. (2001) « Finding An Answer Based on the Recognition of the Question Focus », *Actes de*

TREC 2001.

FREGE G. (1971) « Über Sinn und Bedeutung ». In *Zeitschrift für Philosophie und philosophische Kritik*, 100. Trad. fr. « Sens et dénotation ». In *Ecrits logiques et philosophiques*, Seuil, Paris. pp. 102-126.

GRISHMAN R. (1995) « Where's the Syntax? The NYU MUC-6 System », In Actes *MUC-6*, Morgan Kaufmann Publishers, San Francisco.

KAYSER D., COULON D. (1981) « Variable-Depth Natural Language Understanding ». 7th *International Joint Conference on Artificial Intelligence*. Vancouver. pp.64-66.

KAYSER D. (1988) « What kind of thing is a concept? », In *Computational Intelligence* n°4(2). pp. 158-165

KRIPKE S. (1982) *La logique des noms propres*, Éditions de Minuit, Paris.

KUBALA F., SCHWARTZ R., STONE R., WEISCHEDEL R. (1998) « Named Entity Extraction from Speech », *Proceedings of the DARPA Broadcast Transcription and Understanding Workshop*, Lansdowne, Virginie.

LENAT D., GUHA R. (1990) *Building large knowledge-based systems*, Addison-Wesley, Reading.

LENAT D. (2001) « Common sense and the mind of HAL ». In Stork D. (éd.), *From 2001 to 2001: HAL legacy*.

LEPAGE Y. (2003) *De l'analogie rendant compte de la commutation en linguistique*, Habilitation à Diriger des recherches, Université de Grenoble.

LEROY S. (2003) « Antonomase, métaphore et nom propre : identification ou catégorisation ? », In *Travaux linguistiques du CerLiCO*. Vol 16 (*1. Morphosyntaxe du lexique – 2 : catégorisation et mise en discours*), PUR, Rennes. pp. 161-178

LEROY S. (2004) *Le nom propre en français*, Ophrys, Paris.

MUC-7 (1998) *Proceedings of the Seventh Message Understanding Conference*.

NEHANIV C.L. éd. (1999) *Computation for Metaphors, Analogy, and Agents*, Lecture Notes in Artificial Intelligence n°1562, Springer-Verlag, Allemagne.

POIBEAU T. (2003) *Extraction automatique d'information*. Hermès. Paris.

POPESCU-BELIS A., ROBBA I., SABAH G. (1998) « Reference Resolution Beyond Coreference: a Conceptual Frame and its Application ». *Proceedings of Coling-ACL 1998*, Montreal, Canada. pp.1046-1052.

PUSTEJOVSKY J. (1995) *The generative lexicon*, MIT Press, Cambridge, 1995.

SABAH G. (1996) « Le carnet d'esquisses : une mémoire interprétative dynamique », *Actes Représentation des formes et Intelligence artificielle*, Rennes.

SALMON-ALT S. (2001) « Reference Resolution within the Framework of Cognitive Grammar », *International Colloquium on Cognitive Science*, San Sebastian.

SEARLE J. (1983) *L'intentionnalité. Essai de philosophie de l'esprit*, Éditions de Minuit, Paris.

SEKINE S. (2004) « Definition, dictionaries and tagger for Extended Named Entity Hierarchy », *Actes LREC 2004*, Lisbonne.

WELLINGTON A. éd. (2003) *The Lexicon and Figurative Language*. Atelier ACL 2003. Sapporo. Japon.

Production automatique du résumé de textes juridiques: évaluation de qualité et d'acceptabilité

Atefeh Farzindar et Guy Lapalme
RALI
Département d'informatique et de recherche opérationnelle
Université de Montréal
C.P. 6128, succursale Centre-ville
Montréal, Québec, Canada H3C 3J7
{farzinda, lapalme}@iro.umontreal.ca

Mots-clefs : résumé automatique, fiches de résumé, textes juridiques, évaluation d'un résumé.

Keywords: automatic text summarization, summary table, legals texts, evaluation of a summary.

Résumé- Abstract

Nous décrivons un projet de production de résumé automatique de textes pour le domaine juridique pour lequel nous avons utilisé un corpus des jugements de la cour fédérale du Canada. Nous présentons notre système de résumé LetSum ainsi que l'évaluation des résumés produits. L'évaluation de 120 résumés par 12 avocats montre que la qualité des résumés produits par LetSum est comparable avec celle des résumés écrits par des humains.

We describe an automatic text summarisation project for the legal domain for which we use a corpus of judgments of the federal court of Canada. We present our summarization system, called LetSum and the evaluation of produced summaries. The evaluation of 120 summaries by 12 lawyers shows that the quality of the summaries produced by LetSum is approximately at the same level as the summaries written by humans.

1 Introduction

La jurisprudence est une référence importante pour les juristes. Pour cette raison les juristes consultent quotidiennement des milliers de documents juridiques. De jour en jour, la masse d'information textuelle sous forme de jurisprudence accessible sur internet ou dans les bases de données des entreprises et des gouvernements ne cesse d'augmenter. Ce qui nécessite le développement des outils spécifiques afin de pouvoir accéder au contenu des textes. Le but d'un résumé d'un jugement est d'abord de livrer l'essence du texte clairement et avec concision pour permettre une consultation facile et rapide; il doit fournir suffisamment d'informations sur le jugement pour permettre au lecteur de décider si celui-ci peut être pertinent à sa recherche. Actuellement, des jugements sont résumés manuellement par les professionnels ce qui est très coûteux.

Notre approche au résumé automatique a l'avantage de fournir des moyens clairs de concevoir des documents juridiques en fonction de résumés courts pour différents types d'utilisateurs: des étudiants, des avocats et des juges.

Le domaine juridique est un domaine ayant un grand besoin de résumés mais avec des exigences spécifiques. Dans ce projet, nous nous sommes intéressés au traitement des décisions des cours judiciaires du Canada. Nous avons collaboré avec les avocats du Centre de Recherche en Droit Public (CRDP), chargés de créer la bibliothèque de droit virtuelle des décisions judiciaires canadiens CanLII [1].

Dans cet article, nous décrivons plutôt les aspects qualitatifs de l'évaluation d'un résumé que la méthodologie de la production de résumé automatique. À la section 2, nous rappelons notre approche de production automatique de résumé de jugements et son l'implantation, LetSum (*Legal Text Summarizer*). La section 3 présente les évaluations effectuées avec LetSum. L'évaluation de 120 résumés automatiques par 12 avocats montre que la qualité des résumés produits par LetSum est excellente. La comparaison des résumés de LetSum avec cinq systèmes de recherche ou commerciaux montre l'intérêt d'utiliser d'un système de résumé spécialisé pour le domaine juridique.

2 Résumé de textes juridiques

Notre méthode a été développée suite à une analyse manuelle de 75 jugements et de leurs résumés rédigés par les résumeurs professionnels. Nous avons déjà présenté la problématique (Farzindar, 2004) et notre méthode pour capturer la structuration thématique des documents et identifier les unités textuelles saillantes (Farzindar *et al.*, 2004). Nous identifions d'abord le plan d'organisation d'un jugement et ses différents thèmes discursifs qui regroupent les phrases autour d'un même sujet. Chaque phrase dans un thème donne des informations complémentaires sur le sujet. Pour les phrases reliées à un thème, nous pouvons en interpréter le sens d'après leur contexte afin d'en extraire les idées clés.

[1] Canadian Legal Information Institute http://www.canlii.org

La création du résumé par LetSum se fait en quatre étapes décrites en détail dans (Farzindar, 2005).

Segmentation thématique qui détermine l'organisation du document original et relie les segments du texte associés avec des sept thèmes suivants:

- DONNÉES DE LA DÉCISION: donne la référence complète de la décision et la relation entre les parties sur le plan juridique.
- INTRODUCTION: qui? a fait quoi? à qui?
- CONTEXTE: recompose l'histoire du litige et l'histoire judiciaire.
- SOUMISSION: présente le point de vue d'une partie sur le problème.
- QUESTIONS DE DROIT: identifie le problème juridique dont le tribunal est saisi.
- RAISONNEMENT JURIDIQUE: décrit l'analyse du juge, la détermination des faits et l'expression des motifs de la solution retenue.
- CONCLUSION: présente la décision finale de la cour.

Selon nos observations, quatre thèmes jouent les rôles principaux: INTRODUCTION, CONTEXTE, RAISONNEMENT JURIDIQUE et CONCLUSION. La présence de ces quatre thèmes dans le jugement et dans le résumé est obligatoire. Dans la structure du résumé, nous préservons ces quatre thèmes et nous extrayons les phrases qui leur appartiennent. Le thème QUESTIONS DE DROIT est optionnel dans le jugement.

Filtrage qui identifie les segments qui peuvent être supprimés dans les documents, sans perdre les informations pertinentes pour le résumé. Dans un jugement, les citations occupent un volume important du texte soit 30% du jugement, alors que leur contenu est moins important pour le résumé. Nous identifions les citations principalement pour les supprimer en ne conservant que leurs références juridiques. En plus, le thème SOUMISSION contenant des discours des avocats, identifié par le segmenteur thématique, sera éliminé dans cette étape.

Sélection des unités textuelles candidates pour le résumé qui construit une liste d'unités saillantes pour chaque niveau structural du résumé en calculant les poids pour chaque phrase dans le jugement. La sélection est basée sur des règles sémantiques et des mesures statistiques.

Production du résumé qui choisit les unités pour le résumé final et les combine afin de produire un résumé représentant au maximum 15% du jugement. Le critère de sélection des unités est basé sur l'importance du segment thématique contenant les unités candidates.

La présentation du résumé final est sous forme d'une fiche de résumé contenant des rubriques homogènes d'informations. Cette fiche présente les informations considérées importantes associées à des thèmes précis, ce qui en facilite la lecture et la navigation entre le résumé et le jugement source. Pour chaque phrase du résumé produit, l'utilisateur peut en déterminer le sujet en regardant le thème associé à son segment thématique. La figure 1 montre un exemple de sortie de LetSum comme une fiche de résumé. Cette fiche de résumé montre les thèmes identifiés dans le jugement qui étaient pertinents pour le résumé. La taille du résumé est de 10% de celle du jugement original (le document source a dix pages).

Dans les prochaines sections, nous présentons l'évaluation des résumés générés par LetSum.

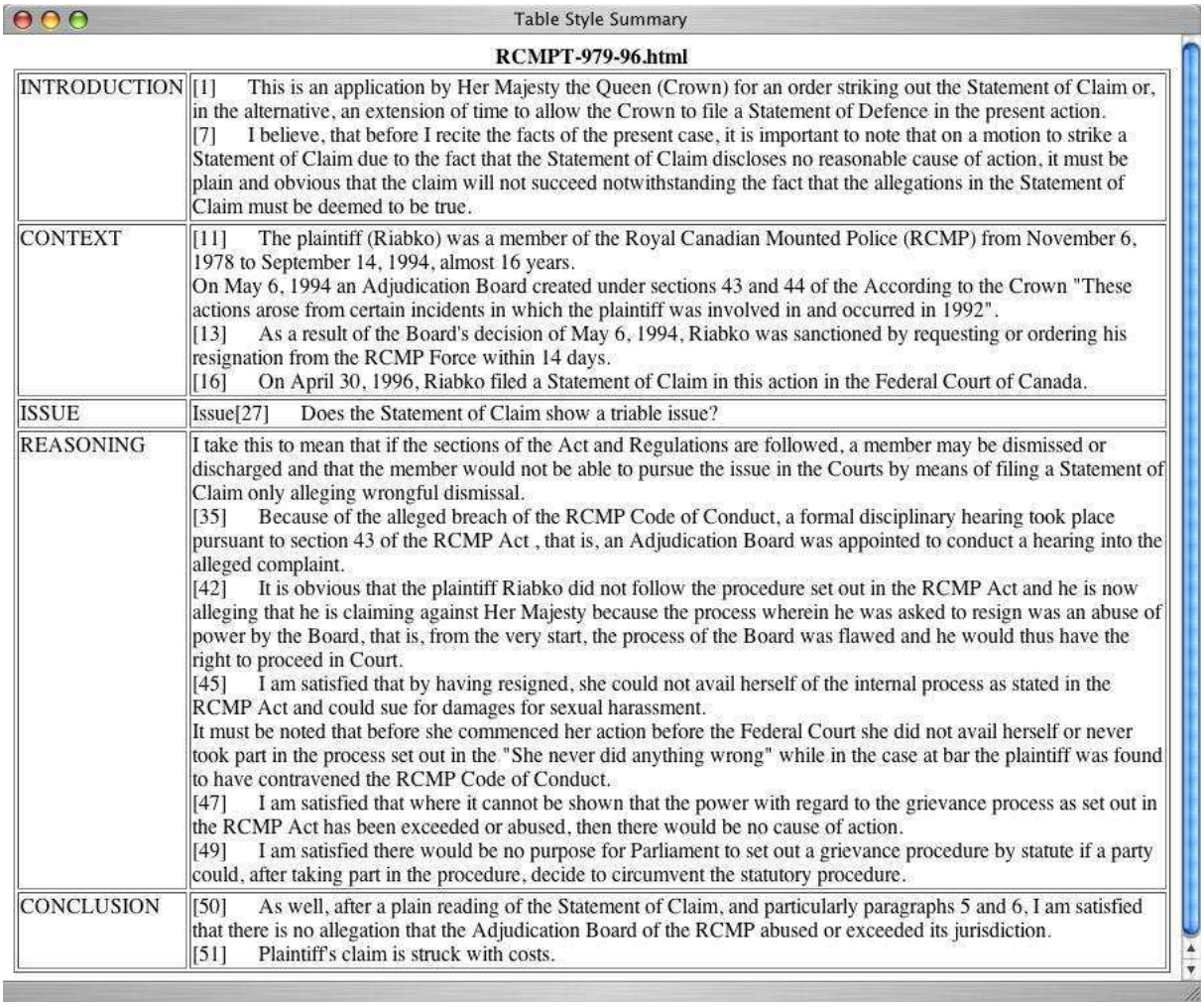


Table Style Summary

RCMPT-979-96.html

INTRODUCTION	[1] This is an application by Her Majesty the Queen (Crown) for an order striking out the Statement of Claim or, in the alternative, an extension of time to allow the Crown to file a Statement of Defence in the present action. [7] I believe, that before I recite the facts of the present case, it is important to note that on a motion to strike a Statement of Claim due to the fact that the Statement of Claim discloses no reasonable cause of action, it must be plain and obvious that the claim will not succeed notwithstanding the fact that the allegations in the Statement of Claim must be deemed to be true.
CONTEXT	[11] The plaintiff (Riabko) was a member of the Royal Canadian Mounted Police (RCMP) from November 6, 1978 to September 14, 1994, almost 16 years. On May 6, 1994 an Adjudication Board created under sections 43 and 44 of the According to the Crown "These actions arose from certain incidents in which the plaintiff was involved in and occurred in 1992". [13] As a result of the Board's decision of May 6, 1994, Riabko was sanctioned by requesting or ordering his resignation from the RCMP Force within 14 days. [16] On April 30, 1996, Riabko filed a Statement of Claim in this action in the Federal Court of Canada.
ISSUE	Issue[27] Does the Statement of Claim show a triable issue?
REASONING	I take this to mean that if the sections of the Act and Regulations are followed, a member may be dismissed or discharged and that the member would not be able to pursue the issue in the Courts by means of filing a Statement of Claim only alleging wrongful dismissal. [35] Because of the alleged breach of the RCMP Code of Conduct, a formal disciplinary hearing took place pursuant to section 43 of the RCMP Act , that is, an Adjudication Board was appointed to conduct a hearing into the alleged complaint. [42] It is obvious that the plaintiff Riabko did not follow the procedure set out in the RCMP Act and he is now alleging that he is claiming against Her Majesty because the process wherein he was asked to resign was an abuse of power by the Board, that is, from the very start, the process of the Board was flawed and he would thus have the right to proceed in Court. [45] I am satisfied that by having resigned, she could not avail herself of the internal process as stated in the RCMP Act and could sue for damages for sexual harassment. It must be noted that before she commenced her action before the Federal Court she did not avail herself or never took part in the process set out in the "She never did anything wrong" while in the case at bar the plaintiff was found to have contravened the RCMP Code of Conduct. [47] I am satisfied that where it cannot be shown that the power with regard to the grievance process as set out in the RCMP Act has been exceeded or abused, then there would be no cause of action. [49] I am satisfied there would be no purpose for Parliament to set out a grievance procedure by statute if a party could, after taking part in the procedure, decide to circumvent the statutory procedure.
CONCLUSION	[50] As well, after a plain reading of the Statement of Claim, and particularly paragraphs 5 and 6, I am satisfied that there is no allegation that the Adjudication Board of the RCMP abused or exceeded its jurisdiction. [51] Plaintiff's claim is struck with costs.

Figure 1: Fiche de résumé produit par LetSum, composé de 350 mots alors que le jugement source avait 4500 mots

3 Évaluation

La comparaison avec un résumé modèle comme référence pour des résumés automatiques est très naturelle, mais des résumés rédigés par des personnes différentes ne sont pas toujours convergents au niveau du contenu. La rédaction d'un résumé demande une analyse du texte pour en dégager les idées, les arguments, le style et les thèmes. Les rédacteurs humains dégagent les affirmations essentielles du document et les expriment dans leur propre style, ce qui donne lieu à plusieurs résumés pour le même document. Il est donc difficile de définir une métrique claire pour juger différents aspects d'un résumé comme la complétude, la thématique et la cohérence.

Plusieurs campagnes d'évaluation de systèmes de résumé comme SUMMAC[2] (Mani *et al.*, 1998) et DUC[3] (organisé par NIST) ont montré l'importance de définir des mesures pour l'évaluation d'un résumé. Spark Jones et Galliers (Spark-Jones & Galliers, 1995) ont proposé de diviser les évaluations en deux types: **intrinsèque** et **extrinsèque**. L'évaluation intrinsèque mesure les propriétés concernant la nature du sujet à évaluer et son objectif, alors que

[2]TIPSTER Text Summarization Evaluation Conference

[3]Document Understanding Conferences http://www-nlpir.nist.gov/projects/duc

l'évaluation extrinsèque mesure les aspects concernant les impacts et les effets de sa fonction. Nous avons évalué LetSum avec ces deux types d'évaluations.

Nos résumés du système ont été évalués en deux étapes: nous avons d'abord évalué les modules du système séparément, ensuite nous avons mesuré la qualité globale des résumés produits. Nous avons également comparé les résumés de LetSum avec des résumés produits par quatre autres systèmes et des résumés manuels.

Pour l'évaluation des modules de LetSum, nous avons utilisé une évaluation intrinsèque à trois niveaux: la qualité des divisions en thèmes par le segmenteur thématique, la détection correcte des citations par le module de filtrage, et le contenu des unités sélectionnées par le module de sélection et production.

Comme évaluation intrinsèque, nous avons utilisé ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) (Lin & Hovy, 2003). ROUGE est maintenant bien reconnue comme mesure d'évaluation des résumés et a été utilisée pour la première fois dans la compétition de DUC 2004 comme seule mesure de fiabilité pour certaines tâches. ROUGE est basé sur le calcul statistique de co-occurrence de n-grammes. Cette méthode dont les résultats sont bien corrélés avec les jugements humains permet d'optimiser les systèmes et d'accélérer leur évaluation. ROUGE comporte deux méthodes d'évaluation. ROUGE-N, dont le score est basé sur le nombre de n-grammes (normalement $1 \leq n \leq 4$) communs entre le résumé automatique est le résumé modèle. Par exemple, ROUGE-2 calcule le nombre de paires de mots successifs communs entre les résumés candidat et modèle. La deuxième est ROUGE-L, qui considère les phrases comme une suite des séquences des mots. Cette évaluation calcule la plus longue sous-séquence commune des mots afin d'estimer la similarité entre deux résumés.

Pour l'évaluation extrinsèque de LetSum, nous avons demandé à des utilisateurs juristes de juger le contenu des résumés et leur acceptabilité. Pour chaque résumé, le recouvrement du contenu sur les idées clés du document a été évalué par deux avocats.

3.1 Évaluation des modules de LetSum

Nous avons évalué les quatre modules de LetSum séparément. Les deux premiers modules, **segmentation thématique** et **filtrage** sont évalués séparément, alors que les deux autres modules de sélection des unités pertinentes et production ont été évalués dans le cadre de l'évaluation des résumés finals de LetSum. Nous avons comparé les sorties de module de segmentation thématique avec le corpus que nous avons annoté manuellement (avec validation d'un avocat du CanLII).

Pour l'évaluation du module de **segmentation thématique**, nous avons utilisé un corpus de test contenant 10 jugements de la cour fédérale. Ces jugements n'ont pas été utilisés pour entraîner le système, ni servi à la construction du dictionnaire des marqueurs. Pour l'évaluation de ce module les points considérés importants sont: détection des thèmes, degré de pertinence d'un thème pour un segment, couverture des segments thématiques, précision des frontières entre deux thèmes. Pour cette évaluation on peut calculer la précision et le rappel. La précision mesure la proportion des unités pertinentes parmi toutes les unités produites par le système. Le rappel mesure la proportion des unités pertinentes parmi tous les unités pertinentes. F-mesure considère les deux mesures ensemble. Nous avons obtenu une précision de 100% et un rappel de 95% soit F-mesure 99% . Sur 40 thèmes annotés dans le corpus, 38 thèmes ont été identifiés correctement.

Pour l'évaluation du module de **filtrage** de citations, nous avons utilisé 15 jugements de la cour fédérale qui n'ont pas servi à entraîner le module de filtrage. Pour cette évaluation, nous avons comparé les unités de citations identifiées par le **filtrage** avec les citations annotées manuellement dans les jugements. Le résultat d'évaluation du module de filtrage est de 98% pour la précision et 95% pour le rappel ce qui donne 96% pour la F-mesure. Sur 60 cas de citation, 57 unités ont été identifiées correctement. Certaines citations n'étaient pas identifiées correctement à cause la langue de rédaction des références. Dans les jugements canadiens, les juges citent parfois les références de droit tels quels peu importe qu'elles soient en anglais ou en français. Pour les citations en français, lorsqu'il y a des marqueurs d'énumération, le système les identifie mais en absence des marqueurs d'énumération, il ne peut pas les distinguer.

3.2 Évaluation de LetSum par ROUGE

Pour le module de **sélection** et **production**, il faut mesurer les topiques extraits des documents par le système. Il est possible d'aligner automatiquement les unités de deux textes pour comparer la similarité entre les résumés modèles et les résumés produits afin de calculer la fraction du résumé modèle exprimée dans le contenu du résumé produit par le système. Pour cette évaluation des résumés de LetSum, nous avons utilisé ROUGE en les comparant avec des résumés modèles écrits par des humains. Nous avons généré 50 résumés automatiques avec cinq systèmes: système de recherche *MEAD* (Radev *et al.*, 2003), un système de recherche et commercial français *Pertinence Mining* (Lehmam, 1995), un système commercial de *Microsoft Word* (option de résumé dans MS Word) et une méthode *StartEnd* que nous avons définie. Le *StartEnd* est un système basé sur les positions des segments dans le document et LetSum afin de comparer notre système avec d'autres systèmes de résumés. Pour la méthode *StartEnd*, nous avons mis au point cette approche suite à nos analyses du corpus des résumés manuels. Pour définir le *StartEnd* nous avons fait trois expérimentations.

D'après nos études, le début du jugement situé à la fin des DONNÉES DE LA DÉCISION (nom de la cour, lieu de l'audience, date, les références et etc.) est une partie importante qui comprend le début du thème INTRODUCTION. Nous avons défini un baseline qui prend 15% du début du texte. Ce baseline couvre des thèmes INTRODUCTION et CONTEXTE.

Un autre baseline prend 15% de la fin du jugement avant la signature du juge. Ce baseline couvre les unités des thèmes RAISONNEMENT JURIDIQUE et CONCLUSION. Nos expériences avec ROUGE ont montré que le score de premier baseline était plus élevé que le deuxième, ce qui signifie l'importance du commencement du document par rapport à sa fin.

Cette expérience, nous a conduit à définir une approche de résumé avec un taux de compression 15%, basé sur l'algorithme suivant: prendre 8% du début du jugement et en complétant la dernière phrase si cette dernière a été coupée et prendre 4% de la fin du jugement en ajoutant la première phrase complète. Cette dernière approche, que nous avons nommée *StartEnd* est donc assez appropriée pour les documents de style juridique même si son implémentation est assez simple.

Nous avons comparé avec ROUGE les résumés de LetSum, *StartEnd* et ceux de trois autres systèmes avec les résumés humains. Les résultats de l'évaluation sont montrés à la table 1. Un score plus élevé est meilleur et indique un système plus performant. LetSum est classé au premier rang avec les meilleures notes d'évaluation. D'après cette évaluation, le deuxième système est *StartEnd*, ce qui montre l'importance de l'étude sur les documents des domaines

Système	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4	ROUGE-L
LetSum	**0.57500**	**0.31381**	**0.20708**	**0.15036**	**0.45185**
StartEnd	0.47244	0.27569	0.19391	0.14472	0.34683
MEAD	0.45581	0.22314	0.14241	0.10064	0.32089
MsWord	0.44473	0.21295	0.13747	0.09727	0.29652
Per. Mining	0.32833	0.15127	0.09798	0.07151	0.22375

Table 1: Résultat d'évaluation intrinsèque avec ROUGE, LetSum a des meilleurs résultats

spécifiques. Le fait qu'une approche simple puisse dépasser des méthodes complexes de production de résumé met en évidence la différence entre des organisations des documents et elle montre aussi l'intérêt de développer un système spécifique pour un domaine. Il est plus en plus difficile de produire un résumé général pour tous types d'utilisateurs sans prise en compte du besoin des usagers et de la tâche demandée.

3.3 Évaluation extrinsèque de LetSum

L'objectif de cette évaluation est de mesurer l'utilité du résumé automatique par rapport à un résumé écrit par un humain et de comparer la qualité des résumés automatiques générés par différents systèmes. Ce test est basé sur un jugement humain. Cette évaluation est toutefois très coûteuse, parce qu'elle demande des ressources humaines et un temps considérable.

Pour cette évaluation, nous avons utilisé les résumés automatiques produits par cinq systèmes présentés à la section précédente et les résumés écrits par des humains. Il faut noter que dans cette évaluation, nous n'avons considéré que les textes du résumé. Nous n'avons pas généré le format tabulaire d'organisation du résumé comme celui qui est présenté à la figure 1. Nous voulions aussi normaliser l'apparence de la sortie de tous les systèmes pour ne pas influencer les juges. Ce choix pénalise toutefois LetSum car nous ne tenons pas compte de la structure thématique extraite par notre méthodologie. Les évaluateurs ne savaient pas quels résumés avaient été produits par ordinateur et lesquels avaient été écrits manuellement.

Nous avons fait évaluer 120 résumés par les juristes. Le corpus de test contient dix jugements choisis au hasard dans différentes collections de jugements de la Cour fédérale du Canada. Nous avons généré 50 résumés automatiques et nous avons collecté 10 résumés manuels écrits par les arrêtistes de la Cour fédérale. Pour chaque résumé, nous avons répété le test deux fois, ceci nous donne deux avis par résumé. Chacun des 12 avocats du CanLII a évalué 10 résumés sur une période d'une heure sur deux aspects: contenu et qualité.

Pour l'évaluation du **contenu**, nous avons défini sept points importants à retrouver dans un jugement. Si un lecteur peut déterminer les points en question en lisant le résumé, on en déduit que le résumé contient suffisamment d'informations pour couvrir les idées clés d'un jugement. Sept questions (présentées en haut de la table 2) ont été déterminées avec l'aide d'un avocat de CanLII. L'ensemble des réponses de ces questions montre le degré de couverture sur des idées clés du jugement source exprimées dans le résumé. La deuxième partie de l'évaluation portait sur la **qualité** d'un résumé selon trois critères:

Lisibilité : La facilité de distinction et de perception du contenu du résumé qui en facilite la compréhension. Ce critère donne une appréciation globale du résumé. On demande si le

Après avoir lu le résumé peut-on déterminer:
Q1. Qui sont les parties en litige?
Q2. Quel est le problème en litige?
Q3. Les questions de droit soulevées?
Q4. Comment le juge a appliqué le droit aux faits?
Q5. Les motifs couvrent-ils les questions de droit?
Q6. Le résumé contient-il les motifs déterminants pour arriver à la conclusion?
Q7. Le résultat final de la cour?

Résumé	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Moyenne
Humain	55,00	90,00	90,00	70,00	**80,00**	**85,00**	**95,00**	80,71
LetSum	50,00	90,00	80,00	**75,00**	75,00	**85,00**	85,00	77,14
StartEnd	**65,00**	**100,00**	**100,00**	70,00	**80,00**	70,00	90,00	**82,14**
MEAD	55,00	**100,00**	95,00	65,00	50,00	50,00	40,00	65,00
MsWord	30,00	80,00	85,00	60,00	45,00	45,00	60,00	57,86
Per. Mining	25,00	65,00	55,00	35,00	35,00	35,00	45,00	42,14
Moyenne	46,67	87,50	84,17	62,50	60,83	61,67	69,17	67,50

Table 2: Questions juridiques utilisées lors de l'évaluation du contenu du résumé par les juristes et les résultats d'évaluation extrinsèque, les pourcentages des réponses positives pour les sept questions juridiques

résumé est: clair, assez clair, peu clair ou incompréhensible.

Cohérence : La présence simultanée d'éléments qui correspondent au même contenu ou qui s'accordent entre eux, qui s'harmonisent. Ce critère contient le fil conducteur du texte pour en assurer la continuité et la progression de l'information. On demande si la cohérence du texte dans le résumé est: très bonne, bonne, médiocre ou très mauvaise.

Pertinence des phrases : Caractère de ce qui est plus ou moins approprié, qui s'inscrit dans la ligne de l'objectif poursuivi. La pertinence des phrases mesure si les phrases du résumé contiennent un lien clair et direct avec le sujet dont il est question. On demande si le résumé est: très pertinent, assez pertinent, peu pertinent ou non pertinent.

L'évaluation comporte aussi une valeur d'acceptabilité sur la qualité générale du résumé. Nous avons demandé d'attribuer une valeur d'acceptabilité entre 0 et 5 pour chaque résumé (0 pour un résumé inacceptable et 5 pour un texte acceptable) sur la qualité du texte de résumé. Les résumés avec valeur 3 jusqu'à 5 sont considérés acceptables.

Dans la table 2, nous présentons les résultats obtenus pour l'évaluation des 120 jugements où, pour chaque question, nous avons calculé le pourcentage de réponses positives données à cette question. Une réponse positive signifie que le résumé contient assez d'informations sur le point en question. Par exemple dans la deuxième colonne, les résumés produits par le moyenne de toutes les méthodes ont couvert les informations sur la présentation des parties en litige (Q1) dans 47% des cas. LetSum a donc très bien répondu aux exigences des avocats pour des résumés automatiques. Ses résultats sont très proches de ceux des résumés manuels et sa performance est supérieure à celle des autres systèmes commerciaux *Microsoft Word* et *Pertinence Mining*, y compris le système de recherche *MEAD*.

Notre méthode *StartEnd*, basée sur la position des segments, a également donné de bons résul-

Résumé	Lisibilité 0-3	Cohérence 0-3	Pertinence 0-3	Acceptabilité 0-5
Humain	2,00	1,95	2,15	3,43
LetSum	2,30	**2,30**	**2,25**	3,43
StartEnd	**2,40**	2,20	2,10	**3,68**
MEAD	2,15	1,90	2,05	3,23
MsWord	1,65	1,25	1,60	2,63
Per. Mining	1,40	1,10	1,40	2,23
Moyenne	1,98	1,78	1,93	3,10

Table 3: Résultats d'évaluation extrinsèque selon les valeurs qualitatives entre 0 et 3 sur lisibilité, cohérence et pertinence des phrases, valeur d'acceptabilité est entre 0 et 5 sur la qualité générale du résumé

tats. Notre heuristique pour les positions des segments était appropriée, même si elle diffère du baseline utilisé normalement pour les articles journaux. Par le comportement du système *MEAD*, spécialisé pour les articles des journaux, on peut voir que les questions qui possèdent les réponses placées au début du document sont bien répondues alors que le recouvrement des informations clés sur les questions avec réponses dans d'autres positions dans le texte n'est pas satisfaisant. Les systèmes commerciaux comme *Microsoft Word* et *Pertinence Mining* ont les scores les plus faibles dans l'évaluation, car ils produisent des résumés génériques qui ne satisfont pas vraiment les utilisateurs dans un domaine spécifique comme droit.

La table 3 montre les résultats de l'évaluation de la qualité du résumé. Pour les trois critères, lisibilité, cohérence et pertinence des phrases du résumé, les valeurs sont entre 0 et 3. La qualité des résumés produits par LetSum est supérieure à celle des autres méthodes. La lisibilité du résumé de LetSum est jugé clair, la cohérence est évaluée très bonne et les pertinences des phrases sont mesurées très pertinentes pour les besoins des avocats. Les condensés rédigés par un humain sont jugés bons en cohérence (et non pas très bons) parce qu'ils sont en style télégraphique alors que LetSum et les autres systèmes font l'extraction de phrases.

Au point de vue d'acceptabilité du résumé, les résumés de LetSum sont jugés de niveau équivalent à celui des résumés écrits par les arrêtistes des cours. Encore une fois la méthode de positions des phrases dans le jugement des très bons scores pour ce critère d'évaluation. Il faut noter que dans cette partie de l'évaluation il y a peu de différence entre le système StartEnd et LetSum, un système nettement plus élaboré. Ceci peut en partie s'expliquer par le fait que nous n'avons pas considéré le format tabulaire produit par LetSum basé sur l'analyse thématique du texte qui distingue notre méthode.

4 Conclusion

Le domaine juridique est un vaste domaine avec un grand besoin pour le résumé automatique. Au Canada, il y a 30 000 avocats et aux États-Unis plus de 300 000 avocats susceptibles de rechercher de la jurisprudence. Toutes les synthèses de jurisprudence se font manuellement par des juristes. Lors qu'un résumé est disponible, le juriste a une idée du contenu de la décision et il lui est plus facile de savoir si elle a un potentiel de pertinence. Chaque résumé peut sauver, dans la consultation d'une liste de résultats de recherche, deux minutes à la personne qui fait la recherche. Une recherche typique dans CanLII donne plus de trente résultats, on pourrait donc

sauver une heure environ. Comme un avocat-recherchiste facture au moins 100$ de l'heure à son client, et que plusieurs recherches peuvent être requises pour un seul dossier, pour 20 recherches, il y aura donc 20 heures d'économies, 2 000$ sur un seul cas. L'utilisation des résumés automatiques économise du temps, des coûts et des expertises. Ces économies de ressources protègent les intérêts du gouvernement et de la population en tant que des clients attendant de recevoir un service juridique.

Nous avons développé LetSum, le premier système complet pour le résumé de textes juridiques en anglais. Il est basé sur l'identification de la structure thématique et présente le résumé sous forme d'une fiche de résumé augmentant ainsi la cohérence et la lisibilité du résumé. Dans les différentes étapes de notre étude, nous avons cherché à maximiser la précision de notre analyse en vue de diminuer les erreurs, car les textes de lois sont très précieux. L'excellente évaluation de LetSum est le témoin de la validité de notre approche. En faisant ressortir les points essentiels des jugements, nous espérons avoir rendu la justice plus accessible à tous et aussi aider la société.

Remerciements

Nous tenons à remercier l'équipe LexUM du laboratoire d'informatique juridique du Centre de recherche en droit public de la faculté de droit de l'Université de Montréal pour leur collaboration. La recherche présentée ici est soutenu financièrement par le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG).

Références

FARZINDAR A. (2004). Développement d'un système de résumé automatique de textes juridiques. In *TALN-RECITAL'2004*, p. 39–44, Fès, Maroc.

FARZINDAR A. (2005). *Résumé automatique de textes juridiques.* PhD thesis, Université de Montréal et Université de Paris4-Sorbonne.

FARZINDAR A., LAPALME G. & DESCLÉS J.-P. (2004). Résumé de textes juridiques par identification de leur structure thématique. *Traitement Automatique des Langues (TAL), Numéro spécial sur: Le résumé automatique de texte : solutions et perspectives*, **45**(1), 39–65.

LEHMAM A. (1995). *Le résumé automatique à fragments indicateurs: RAFI.* PhD thesis, Université de Nancy-II, Nancy, France.

LIN C.-Y. & HOVY E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of 2003 Language Technology Conference (HLT-NAACL 2003)*, p. 150–157, Edmonton, Canada.

MANI I., HOUSE D., KLEIN G., HIRSHMAN L., ORBST L., FIRMIN T., CHRZANOWSKI M. & SUNDHEIM B. (1998). *The TIPSTER SUMMAC Text Summarization Evaluation.* Rapport interne MTR 98W0000138, The Mitre Corporation, McLean, Virginia.

RADEV D., OTTERBACHER J., QI H. & TAM D. (2003). Mead reducs: Michigan at duc 2003. In *DUC03*, p. 160–167, Edmonton, Alberta, Canada: Association for Computational Linguistics.

SPARK-JONES K. & GALLIERS J. R. (1995). *Evaluating Natural Language Processing Systems: An Analysis and Review.* Number 1083 in Lecture Notes in Artificial Intelligence. Springer.

Utilisation de la structure morpho-syntaxique des phrases dans le résumé automatique : compression de phrases narratives

Mehdi Yousfi-Monod, Violaine Prince
LIRMM - CNRS - Université Montpellier 2, UMR 5506
161 rue Ada, 34392 Montpellier Cedex 5 - France
{yousfi, prince}@lirmm.fr

Mots-clefs : résumé automatique, compression de phrases, analyse syntaxique

Keywords: automatic summarization, sentence compression, syntactic analysis

Résumé Nous proposons une technique de résumé automatique de textes par contraction de phrases. Notre approche se fonde sur l'étude de la fonction syntaxique et de la position dans l'arbre syntaxique des constituants des phrases. Après avoir défini la notion de constituant, et son rôle dans l'apport d'information, nous analysons la perte de contenu et de cohérence discursive que la suppression de constituants engendre. Nous orientons notre méthode de contraction vers les textes narratifs. Nous sélectionnons les constituants à supprimer avec un système de règles utilisant les arbres et variables de l'analyse morpho-syntaxique de SYGFRAN [Cha84]. Nous obtenons des résultats satisfaisants au niveau de la phrase mais insuffisants pour un résumé complet. Nous expliquons alors l'utilité de notre système dans un processus plus général de résumé automatique.

Abstract We propose an automated text summarization through sentence compression. Our approach uses constituent syntactic function and position in the sentence syntactic tree. We first define the idea of a constituent as well as its role as an information provider, before analyzing contents and discourse consistency losses caused by deleting such a constituent. We explain why our method works best with narrative texts. With a rule-based system using SYGFRAN's morpho-syntactic analysis for French [Cha84], we select removable constituents. Our results are satisfactory at the sentence level but less effective at the whole text level. So we explain the usefulness of our system in a more general automatic summarization process.

1 Introduction

La quantité d'informations disponibles sur Internet ou au sein de certaines entreprises, administrations et laboratoires ne cesse de croître. Ce phénomène rend la recherche d'information de plus en plus difficile. Le résumé automatique, visant à réduire considérablement la taille de ces données, apparaît comme une des solutions permettant, non seulement de faciliter cette recherche en présentant un texte pertinent de plus petite taille, mais aussi de rendre plus rapide le choix d'acceptation de la pertinence ou non d'un texte par rapport à une requête.

La suppression des phrases estimées les moins pertinentes est une technique majoritaire parmi l'ensemble des résumeurs automatiques actuels. Ces approches travaillent à un niveau de granularité grossier : la phrase. Pourtant dans de nombreux textes narratifs, certaines phrases sont longues et peuvent réunir à la fois des passages importants et d'autres moins importants. Pour gagner en contraction, il devient donc nécessaire de pénétrer dans les phrases afin d'éliminer les constituants[1] les moins importants. L'idée centrale de notre recherche est de traquer les limites de la contraction de textes par compression de phrases sans perte majeure d'information. L'originalité de notre approche est de se baser conjointement sur la fonction syntaxique et la position dans l'arbre syntaxique des constituants de la phrase pour sélectionner les constituants supprimables. Nous ne tenons pas compte du contexte ni du cotexte d'une phrase dans son analyse : seules les informations syntaxiques présentes dans la phrase sont utilisées.

Dans la prochaine section, nous survolons les principales approches sur le résumé automatique puis nous traitons de deux méthodes basées sur la compression de phrases (section 2) ; nous présentons ensuite notre approche (section 3), nous continuons en illustrant l'efficacité de notre système par une expérimentation basée sur une application prototype appliquée à un texte du genre conte (section 4) et enfin nous discutons sur les résultats de cette expérimentation et sur les perspectives envisagées (section 5).

2 La compression de phrases

Une grande partie des techniques de résumé automatique procède par extraction de segments textuels. Ces méthodes sont fondées sur l'hypothèse « qu'il existe, dans tout texte, des *unités textuelles saillantes* » [Min04]. Ces dernières représentent des points focaux, qui, soit expriment l'apport sémantique du texte, soit permettent de le représenter dans sa globalité. Dès lors, le résumé par extraction cherchera à repérer ces unités saillantes et proposera un texte de taille plus petite que le document initial qui garderait majoritairement ces unités. Nous faisons également l'hypothèse de l'existence de ces unités ainsi que de leur intérêt pour le résumé. Ce seront les constituants dits gouverneurs, définis en section 3.2, qui correspondront à ces unités saillantes. Parmi ces techniques de résumé, une majorité utilise l'extraction de phrase clés [Luh58, BE97, GMCK00, BN00] pour produire le résumé final. Dans cet article nous nous intéressons uniquement au résumé intra-phrase et plus précisément à la compression de phrases.

[KM02] aborde le problème de la compression de phrases en utilisant un modèle de canal bruyant (*noisy-channel model*) qui consiste à faire l'hypothèse (1) : la phrase à comprimer

[1] Nous appelons *constituants* les syntagmes des phrases, c'est-à-dire toute unité de la phrase à laquelle on peut attribuer une fonction. Par exemple, prenons le groupe nominal "un médecin de famille". Il est composé de deux constituants : un groupe nominal "un médecin" et un groupe nominal prépositionnel "de famille". Ce dernier a un rôle de modificateur du premier.

fût autrefois courte et l'auteur y a ajouté des informations supplémentaires (le bruit). Le but est alors de retrouver ces informations pour les supprimer. Les auteurs utilisent un modèle probabiliste de type modèle de Bayes qu'ils entraînent sur un corpus de documents avec leur résumé. Le moteur d'apprentissage a pour but de sélectionner les mots à conserver dans la phrase comprimée. Une faible probabilité sera attribuée à une phrase comprimée lorsque cette dernière sera incorrecte grammaticalement ou aura perdu certaines informations comme la négation. D'après leur évaluation, les résultats sont assez concluants. Relativement aux compressions réalisées par des êtres humains, une légère perte d'importance et de justesse grammaticale est observée.

[Gre98] utilise la nature des syntagmes et propositions pour estimer leur importance, puis supprime les moins importants pour produire les phrases compressées. La coherence obtenue est évidemment faible mais suffisante pour l'application souhaitée qui est la réduction de textes télégraphiques destinés à être lus par les mal voyants.

Ces deux approches ne prennent pas en compte les informations sur la fonction syntaxique et la position dans l'arbre syntaxique des constituants des phrases. Ces informations pourraient être grandement utiles dans l'aide au choix des constituants à supprimer.

[Lin03] a évalué la qualité d'un résumé produit par extraction de phrases clés puis compression des phrases extraites. L'auteur conclut, d'après les résultats de ses expérimentations, qu'on ne peut pas se fier à une compression strictement basée sur la syntaxe des phrases pour améliorer la qualité des résumés produits par extraction. Cependant, étant donné que l'auteur n'utilise qu'une seule méthode (celle de [KM00]) pour comprimer les phrases, nous ne sommes pas d'accord sur sa conclusion généralisée à l'ensemble des méthodes de compression. Ce que nous concluons c'est que la méthode de compression utilisée, qui, en pratique, mélange à la fois paradigme statistique, apprentissage, technique de "noyage" (dans le bruit) et structure syntaxique, ne satisfait pas les contraintes de conservation du contenu. Notre approche diffère grandement de celle de [KM00] sur au moins deux points : nos règles de compression sont produites manuellement, en relation avec des modèles linguistiques, puis mises en œuvre, et non inférées automatiquement de façon calculatoire. De plus, nous ne faisons pas l'hypothèse de départ (1) de [KM02] qui pour nous est très discutable.

3 La compression par élagage de l'arbre syntaxique

Le point de départ de notre approche fût l'intuition que **la fonction syntaxique et la position dans l'arbre syntaxique des constituants des phrases jouaient un rôle conséquent dans l'importance de ces constituants pour la compréhension d'un texte.** Cette intuition prend ses racines dans l'analyse grammaticale logique enseignée depuis longtemps et dont on trouve des manuels connus (citons Grévisse [Gre97] pour mémoire). En effet, ne sont pas toujours indispensables pour comprendre le sens principal de la phrase, certains épithètes, certains compléments circonstanciels, etc. Par exemple, dans la phrase « *Un chat gros et laid mange une souris.* », le groupe adjectival épithète "gros et laid" peut être supprimé sans nuire réellement à la compréhension et à l'intérêt.

Une autre approche se basant sur la fonction syntaxique est celle de [LBM04] qui travaille à un niveau de granularité très fin, nettement inférieur à la proposition. Dans le système des auteurs, les fonctions syntaxiques des syntagmes sont extraites par un système à base de règles. Une forme logique des phrases est produite et représentée par un arbre dont les noeuds sont les syntagmes (ou des variables si des informations sont manquantes) et les arêtes les fonctions. À

partir des relations entre les syntagmes, un graphe du document est créé sur lequel l'algorithme Pagerank [BP98] est appliqué pour évaluer l'importance de chaque noeud. Les noeuds les plus importants sont ensuite extraits et fournis à un module de génération de phrases qui produit le résumé final. Leur système utilise la fonction syntaxique des syntagmes mais pas la structure syntaxique des phrases[2], ceci laisse au module de génération la lourde tâche de produire des phrases syntaxiquement et sémantiquement cohérentes. Notre système ne fait que supprimer des sous-arbres de l'arbre syntaxique, ceci évite de tomber dans ces problèmes d'incohérence.

Notre approche nécessite un outil d'analyse morpho-syntaxique des phrases (section 3.1) et une étude sur l'importance des constituants relativement à leur fonction syntaxique et leur position dans l'arbre syntaxique (section 3.2). Nous présentons l'architecture de notre système dans la section 3.3.

3.1 L'analyseur morpho-syntaxique

Nous utilisons l'analyseur morpho-syntaxique du français SYGFRAN, basé sur le système opérationnel SYGMART, tous deux définis dans [Cha84]. SYGFRAN utilise un ensemble de règles de transformations d'éléments structurés, basées sur les règles de la grammaire française, qui permettent de transformer une phrase (texte brut) en un arbre syntaxique (élément structuré) enrichi d'informations sur les constituants. Cet analyseur a les avantages suivant :

- la rapidité : la complexité d'analyse est en $O(k * n * log_2(n))$ où k est le nombre de règles et n la donnée textuelle. Il s'agit d'une limite supérieure, car l'analyseur étant structuré en plusieurs grammaires ordonnées, le facteur multiplicatif réel est beaucoup plus petit que k. Cela dit, même ainsi, plus le texte est important, plus k est petit devant n. Aujourd'hui, SYGFRAN analyse un corpus de 220000 phrases en moins d'une demi-heure (avec un Pentium IV 2,4Ghz, 4734 Bigomips, 1Go Ram).
- la robustesse : SYGFRAN parvient à obtenir une structure correcte pour au moins 30 % de l'ensemble des différents cas de syntaxe des phrases du français, pour les autres cas, **SYGFRAN fournit une analyse partielle mais exploitable**.
- la production d'un arbre syntaxique : la plupart des systèmes actuels d'analyse syntaxique ne réalisent qu'un simple marquage linéaire, ceux qui produisent un arbre sont très peu robustes à l'égard de l'ensemble des constructions syntaxiques existantes.

SYGFRAN prend en entrée du texte brut et produit une structure parenthésée, correspondant à l'arbre morpho-syntaxique de chaque phrase du texte, dans laquelle de nombreuses variables sont renseignées sur les différents natures, fonctions syntaxiques, formes canoniques, catégories grammaticales, temps, modes, genres, nombres, etc. des constituants.

3.2 Fonction et Position

Le test de suppression des constituants est abordé par de nombreux ouvrages sur la grammaire française pour aider à la détermination de la fonction syntaxique d'un constituant. Le test est validé si la phrase résultante reste grammaticalement cohérente. Cependant, les textes linguistiques traitant de l'importance des constituants dans la phrase selon leur fonction syntaxique sont beaucoup plus rares. Des recommandations sont fournies par les linguistes, mais pas de règle fondamentale. Nous avons donc procédé de la manière suivante. **Nous avons considéré**

[2] Les auteurs utilisent une structure logique différente de l'arbre syntaxique des phrases.

ces recommandations comme des hypothèses de travail et nous avons cherché à les étayer empiriquement. Ainsi, Mel'čuk, dans son analyse du français contemporain, parle de fonctions syntaxiques dites de "gouvernement" (à la suite des travaux de Chomsky). Sont **gouverneurs** des constituants considérés comme indispensables à la cohérence grammaticale et sémantique de la phrase. Ainsi, le sujet d'une phrase et son groupe verbal sont gouverneurs sur le plan de la cohérence grammaticale.

Considérons la phrase simple suivante : « *Jean mange une pomme verte.* ». Le sujet "Jean", s'il est supprimé, produit une phrase incohérente. Comme il est atomique, on ne peut pas le réduire. Le verbe "mange" également. Si on supprime le complément d'objet direct, "une pomme verte", on a une phrase grammaticalement cohérente (car le verbe *manger* a une forme intransitive). En revanche, on perd de l'information importante, vu que le verbe n'est pas utilisé ici de manière intransitive. Il est spécifiquement qualifié, il importe donc de lui restituer son complément, sur lequel on regarde si on peut appliquer une fonction de restriction. Dans le constituant "une pomme verte" il y a en réalité deux constituants, qui se divisent à leur tour en gouverneur et non gouverneur. Dans un groupe nominal adjectival, le nom est gouverneur et la restriction "une pomme" par rapport à "une pomme verte" ne perd pas en cohérence grammaticale et ne perd pas sa fonction syntaxique. Ainsi la détermination du constituant *secondaire* se fait par rapport au rôle syntaxique. Trois niveaux de granularité sont considérés, la **phrase** (qui peut comprendre plusieurs propositions), la **proposition** (qui est définie par un sujet, un verbe et éventuellement un ou plusieurs compléments) et le **constituant nominal.**

Voici les ordres d'importance (décroissante) des éléments à chaque niveau de granularité :
- la phrase : la proposition principale, les propositions relatives tenant lieu de complément du verbe, les propositions relatives tenant lieu d'épithète et se trouvant généralement en apposition ;
- la proposition : les sujets et verbes, les compléments d'objet (directs et indirects), les compléments circonstanciels ;
- le constituant nominal : les noms, les compléments de noms, les adjectifs (épithètes).

L'idée est de dire que plus on descend dans la liste (par rapport à une granularité donnée) plus on a de chances de réaliser une compression sans perte de cohérence ni perte d'information. Tout le problème consiste à savoir si on peut supprimer systématiquement ou non des éléments de granularité plus large comme les propositions relatives, si on peut supprimer les moins importants des constituants (les compléments circonstanciels par exemple), si on peut élaguer des constituants nominaux, et si ces actions peuvent être relativement généralisées (grosso modo, à tout type de texte).

Pour cela, à partir de textes de genres variés, nous avons réalisé des tests de suppression de certains constituants en fonction de leur fonction syntaxique (donc plutôt la granularité "moyenne"), en estimant les pertes de cohérence discursive et de contenu important dans les phrases comprimées. Dans les textes du genre article scientifique ou énoncé technique, chaque constituant se révèle avoir beaucoup plus d'importance que dans un texte narratif (roman, conte, ...). La raison est que les auteurs de textes narratifs ajoutent de nombreuses informations à caractère essentiellement descriptif qui aident le lecteur à être transporté dans l'histoire mais qui ne sont pas indispensables à la compréhension du cœur de l'histoire. Alors que dans un article scientifique ou technique, chaque constituant a un rôle important à jouer dans la compréhension du discours. Afin d'évaluer les qualités de la compression par suppression de constituants, nous avons donc cherché à la tester sur des corpus où elle avait un sens, en d'autre termes dans les textes de type **narratif**, en se proposant ultérieurement de tester d'autres paradigmes pour les textes scientifiques ou techniques.

[Man04] aborde la problématique du résumé de textes narratifs, en s'appuyant principalement sur des indices temporels. Il étudie les événements sur trois plans : la scène, l'histoire et l'intrigue, dans le but d'extraire les événements clés, scènes clés, et les intrigues saillantes. Il compte sur les méthodes actuelles (basées sur le marquage lexical, l'étude de la structure rhétorique, l'analyse morpho-syntaxique, ...) et futures pour extraire les indices temporels nécessaires. Notre méthode actuelle ne tient compte que des informations syntaxiques.

En supprimant dans une première passe les constituants les plus secondaires on obtient un résumé dont le contenu important est bien conservé mais dont la taille est grande. La compression peut alors consister à plusieurs passes jusqu'à obtenir un rapport spécifique (taille/pertes) du résumé produit. Chaque constituant est supprimé par élagage de l'arbre syntaxique. Après une première passe, les arbres syntaxiques obtenus se révèlent être de bons représentants des originaux. Leur représentativité se dégrade sensiblement après chaque passe.

Nous avons noté trois catégories de constituants susceptibles d'être supprimés selon leur fonction syntaxique et leur position : les compléments circonstanciels, les épithètes et les appositions. Comme on peut le voir, ils sont de granularité moyenne. Les appositions, lorsqu'elles se transforment en propositions relatives (complément de nom) deviennent de granularité plus importante, et augmentent de ce fait le taux de compression obtenu.

Les compléments circonstanciels. De manière générale, ce sont les CC de *temps* et de *but* qui répondent aux questions les plus importantes, à savoir "Quand ?" et "Dans quel but ?". Les CC de *lieu* (questions "Où ?") ont leur importance principalement au début du texte, lorsque le décor est posé. Ceux de *manière* (questions "Comment ?") et de *cause* (questions "Comment est-ce arrivé ?") sont peu importants dans une majorité des cas. La fréquence d'apparition des autres CC (*comparaison*, *condition*, *conséquence*, *opposition*, *mesure*, ...) étant assez faible, leur suppression n'aboutit fréquemment qu'à une petite perte de contenu. Certains gérondifs fonctionnent comme des propositions subordonnées circonstancielles, nous les supprimons aussi. L'importance des CC varie aussi selon la nature du verbe de la proposition. Dans le cas d'un CC de lieu placé après le verbe "être", la suppression ne sera pas possible. Enfin nous avons remarqué qu'un CC situé dans une phrase interrogative était très important car la question porte généralement sur lui.

Les épithètes. Les adjectifs et groupes adjectivaux ont une fonction d'épithète. D'une manière comparable aux CC, lorsqu'un épithète est placé après le verbe "être", et plus généralement après un verbe d'état, son importance s'accroît considérablement, rendant la suppression impossible. Enfin, nous avons noté que lorsque l'épithète était placé dans un groupe nominal dans lequel le déterminant était un article défini, alors sa suppression était difficile. Ceci est dû au fait que l'article défini est utilisé pour parler d'une entité particulière et que les épithètes du nom permettent de différencier cette entité des autres. Certaines propositions relatives ont aussi une fonction d'épithète. Les relatives constituent, d'après Mann et Thompson [MT88], des informations sur le contexte, elle ne sont donc pas indispensables.

Les appositions. L'apposition peut avoir des natures variées, elle peut être :
- un groupe nominal (« *Jean, le gourmand, aime les bonbons.* »),
- un pronom (« *Jean doit manger lui-même les bonbons.* »),
- une proposition relative (« *Jean, qui aime les bonbons, a beaucoup de caries.* »),
- une proposition participale présent (« *Jean, aimant les bonbons, a beaucoup de caries.* »),
- une proposition participale passé (« *Jean, aimé des enfants, fera un bon père.* »),
- une proposition infinitive (« *Jean, manger des légumes, cela m'étonnerait !* »).

Dans les trois premiers cas, les constituants se suppriment sans difficulté. Les propositions par-

ticipales sont aussi sujettes à la suppression, mais une perte un peu plus importante de contenu est à noter. Dans le dernier cas, la suppression paraît difficile car la proposition infinitive apporte systématiquement une information importante qui vient compléter le sujet.

3.3 Architecture

L'architecture de notre système est présentée en figure 1. Du texte source sont produits les arbres syntaxiques correspondant au résultat de l'analyse faite par SYGFRAN. Ensuite, le module de sélection/coloration de segments textuels utilise les informations suivantes pour effectuer la sélection : le texte source, les arbres syntaxiques et les variables/valeurs fournis par SYGFRAN, le seuil du rapport taille/pertes à ne pas dépasser fourni par l'utilisateur ou défini par le type d'application et l'ensemble des règles de sélection des constituants pour effectuer les différentes passes de sélection des constituants jusqu'à satisfaction du rapport taille/pertes. Les constituants sélectionnés sont ensuite supprimés.

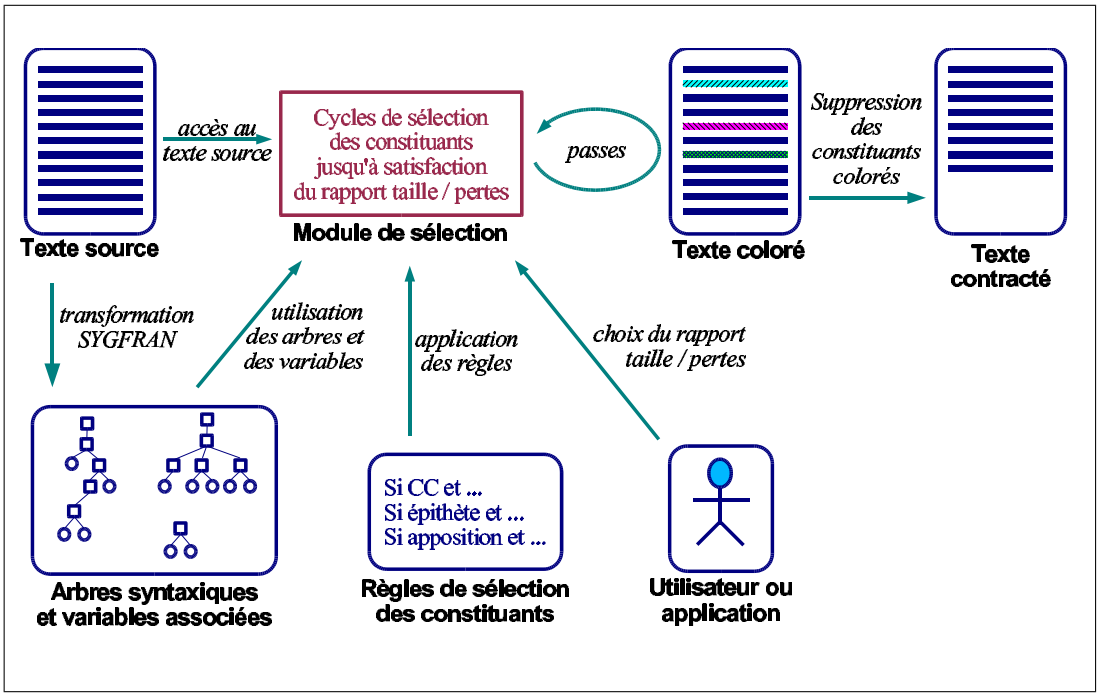


FIG. 1 – Du texte source au texte contracté : notre système de compression de phrases

4 Expérimentations

Nous avons réalisé un programme prototype afin de pouvoir mesurer l'efficacité d'une telle approche. Nous avons défini un système utilisant des règles simples, basées sur les résultats de notre étude expérimentale (section 3.2). Chaque règle possède un nom auquel on associe un ensemble de couples (clé,valeur). Chaque nom représente un type de constituant susceptible d'être supprimé. Les couples (clé,valeur) sont les contraintes qu'un constituant doit respecter pour être sélectionné à la suppression. Notre système actuel possède trois types de contraintes :

une sur la valeur de la variable du constituant fournie par SYGFRAN (par exemple, le constituant doit être un complément circonstanciel), une sur la position du constituant par rapport à un autre constituant relativement à un nœud père spécifique (par exemple, le constituant ne doit pas être à droite d'un verbe d'état) et une sur la position du constituant par rapport à un antécédent possédant une valeur spécifique à une clé (par exemple, le constituant ne doit pas être un sous-constituant d'une phrase interrogative).

Notre prototype actuel n'effectue qu'une passe. Nous comptons créer par la suite des règles paramétrables afin de gérer plusieurs rapports de taille/pertes dans la production du résumé. La première phase consiste à colorier les constituants susceptibles d'être ôtés par la suite. Une couleur est attribuée à chaque type de constituant. Ainsi il est aisé d'estimer la qualité des règles sur le texte en cours avant de supprimer réellement ces constituants. Dans la seconde phase, les segments textuels colorés sont supprimés pour obtenir le résumé final. Nous avons utilisé comme texte de test un conte haïtien. La principale raison de ce choix est que SYGFRAN produit une syntaxe correcte pour l'intégralité des phrases de ce texte. Le résultat de la coloration de la première moitié de ce texte est présenté en figure 2.

MAUI PART À LA RECHERCHE DE SES PARENTS.
À partir de ce soir-là, Maui fut le favori de sa mère : même s'il faisait des bêtises, elle ne le grondait pas. Quand ses frères protestaient, il se moquait d'eux parce qu'il savait avoir la protection de sa mère. Mais pendant son absence, il devait faire attention à ne pas dépasser les limites, sinon il risquait d'être puni par eux au cours de la journée.
Une nuit, Maui imagina un tour à jouer à sa mère afin de découvrir où elle allait. Une fois tous les autres endormis sur leurs nattes, il se releva et fit le tour de la maison, examinant les grands stores tressés qui la fermaient pour la nuit. Partout où filtrait la clarté d'une étoile, il bouchait vite l'ouverture avec des étoffes d'écorce et calfeutrait même les fentes avec des roseaux. Puis il déroba le manteau, la ceinture et la couronne de sa mère et les cacha en se disant qu'il en aurait besoin plus tard. Maui reprit alors sa place sur les nattes et décida de rester éveillé. La longue nuit passa lentement sans que sa mère ne bouge.
Quand vint le matin, pas un rai de lumière ne put percer pour éveiller les dormeurs. Bientôt ce fut l'heure où le soleil grimpait au-dessus de l'horizon. D'habitude Maui pouvait distinguer dans la pénombre les formes des pieds de ses frères à l'autre bout de la maison, mais ce matin il faisait trop noir. Et sa mère continuait à dormir.
Au bout d'un moment elle bougea et marmonna : "Quelle sorte de nuit est-ce donc pour durer si longtemps ?" Mais elle se rendormit parce qu'il faisait aussi noir qu'au cœur de la nuit dans la maison. Finalement elle se réveilla en sursaut et se mit à chercher ses vêtements. Courant de tous côtés, elle arracha ce que Maui avait fourré dans les fentes. Mais c'était le jour ! Le grand jour ! Le soleil était déjà haut dans le ciel ! Elle s'empara d'un morceau de tapa pour se couvrir et se sauva de la maison, en pleurant à la pensée d'avoir été ainsi trompée par ses propres enfants . Sa mère partie, Maui bondit près du store qui se balançait encore de son passage et regarda par l'ouverture. Il vit qu'elle était déjà loin, sur la première pente de la montagne. Puis elle s'arrêta, saisit à pleines mains un arbuste de tiare Tahiti, le souleva d'un coup : un trou apparut, elle s'y engouffra et remit le buisson en place comme avant.
Maui jaillit de la maison aussi vite qu'il put, escalada la pente abrupte, trébuchant et tombant sur les mains car il gardait les yeux fixés sur l'arbuste de tiare. Il l'atteignit finalement, le souleva et découvrit une belle caverne spacieuse qui s'enfonçait dans la montagne.

Légende : compcir (complément circonstanciel), phger (proposition au gérondif), phrel (proposition relative), gadj (groupe adjectival).

FIG. 2 – Coloration d'un texte, d'après notre méthode de compression de phrases

5 Discussion sur les résultats et perspectives

Avec le jeu de règles actuel, notre approche nous a permis d'éliminer environ 34 % du texte complet. Nous constatons une légère perte de contenu et de cohérence discursive, celle-ci reste plus que raisonnable au regard des techniques actuelles de résumé automatique. La cohérence grammaticale, quand à elle, est très bien conservée. Nous estimons que les règles peuvent encore être affinées, mais les données linguistiques dans ce domaine sont très limitées. Pour ce texte, SYGFRAN nous fournit des arbres syntaxiques corrects, mais les valeurs des variables ne sont pas systématiquement justes et complètes. Pour les CC, SYGFRAN ne spécifie actuellement la sémantique de l'objet que pour ceux de temps et de lieu.

Pour le constituant "afin de découvrir où elle allait" du deuxième paragraphe, nous possédons l'information que c'est un CC mais pas que c'est un CC de but. Ce genre de constituant devrait être conservé. Dans le cas du constituant "D'habitude" du troisième paragraphe, SYGFRAN ne détecte pas que c'est un CC de temps, c'est pourquoi nous le sélectionnons à tort à la suppression. Idem pour "Finalement" au quatrième paragraphe. L'évolution des règles de SYGFRAN permettra de gérer de tels cas.

Les règles de sélection des constituants à supprimer peuvent être affinées davantage selon la fonction des constituants et surtout selon le genre des textes. Nous comptons, à cet effet, effectuer des expérimentations sur plus de textes touchant à des genres plus variés. Cependant, la compression de phrases ne suffit pas à produire un résumé d'une taille convenable dans la plupart des cas d'applications. Comme nous l'avons vu, elle est aussi fortement dépendante du genre de texte. Nous considérons donc notre approche intra-phrase comme une des tâches à effectuer lors de la production d'un résumé automatique, en complément avec d'autres approches qui travaillent à un niveau de granularité supérieur ou égal aux phrases.

6 Conclusion

Bien que le problème du résumé automatique ait déjà été abordé par de nombreux scientifiques depuis presque 50 ans [Luh58], l'approche que nous avons adoptée est novatrice. En effet, les approches actuelles du résumé automatique utilisent des informations telles la fréquence des termes, les relations lexicales entre les termes, les étiquettes sur la nature des constituants fournis par des *POS tagger* (lemmatiseurs), les probabilités d'un constituant d'apparaître dans un résumé d'après des moteurs d'apprentissage, la structure rhétorique du texte, cependant, aucune d'entre elles n'utilise conjointement **la fonction syntaxique et la position dans l'arbre syntaxique des constituants.**

Ces informations n'ont pas été réellement exploitées jusqu'à présent car elles ne peuvent être extraites qu'avec des analyseurs morpho-syntaxiques fonctionnant avec un niveau suffisant. Ce niveau n'a été atteint que récemment en traitement automatique des langues, parce qu'il est fort coûteux en temps de calcul. Le système opérationnel SYGMART est l'un de ces outils. En outre, il ajoute à l'analyse des constituants de nombreuses informations concernant les relations entre constituants, ce que peu d'autres analyseurs proposent.

Notre approche a débuté par une étude sur l'importance des constituants dans une phrase. Le critère de suppression a été l'évaluation de la perte de contenu et de cohérence que la suppression de ces constituants engendre. Le critère de sélection est celui de la fonction syntaxique et

de la position dans l'arbre syntaxique des constituants. Les textes narratifs (romans, contes, ...) se sont révélés être les plus adéquats pour une telle approche. Nous avons alors modélisé une compression de phrases basée sur la suppression de ces constituants. La création d'un système de règles basé sur notre modélisation nous a permis de tester la faisabilité d'une telle approche. Nous sommes passés par une étape de coloration des constituants en fonction des règles qui les avaient sélectionnés, afin d'estimer la pertinence de chaque règle. Notre méthode nous a permis de supprimer environ 34 % du texte de test, tout en conservant une très bonne cohérence grammaticale. Nous avons conclu que notre compression a son utilité dans un processus plus large de résumé automatique.

Références

[BE97] R. Barzilay and M. Elhadad. Using lexical chains for text summarization. In *Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS'97)*, Madrid, Spain, 1997. ACL.

[BN00] Branimir K. Boguraev and Mary S. Neff. Lexical cohesion, discourse segmentation and document summarization. In *RIAO-2000*, Paris, April 2000.

[BP98] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *WWW7 : Proceedings of the seventh international conference on World Wide Web 7*, pages 107–117, Amsterdam, The Netherlands, The Netherlands, 1998. Elsevier Science Publishers B. V.

[Cha84] Jacques Chauché. Un outil multidimensionnel de l'analyse du discours. In *Coling'84*, pages 11–15, Standford University, California, 1984.

[GMCK00] J. Goldstein, V. Mittal, J. Carbonell, and M. Kantrowitz. Multi-document summarization by sentence extraction. In *Hahn et al.[15]*, pages 40–48, 2000.

[Gre97] Maurice Grevisse. *le Bon Usage – Grammaire française*. édition refondue par André Goosse, DeBoeck-Duculot, Paris – Louvain-la-Neuve, 13e édition, ISBN 2-8011-1045-0, 1993-1997.

[Gre98] Gregory Grefenstette. Producing intelligent telegraphic text reduction to provide audio scanning service for the blind. In *In AAAI symposium on Intelligent Text Summarisation*, pages 111–117, Menlo Park, California, 1998.

[KM00] Kevin Knight and Daniel Marcu. Statistics-based summarization - step one : Sentence compression. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 703–710, Sapporo, Japan, 2000.

[KM02] Kevin Knight and Daniel Marcu. Summarization beyond sentence extraction : a probabilistic approach to sentence compression. *Artificial Intelligence archive*, 139(1) :91–107, Jully 2002.

[LBM04] Vanderwende Lucy, Michele Banko, and Arul Menezes. Event-centric summary generation. In *In Document Understanding Conference at HLT-NAACL*, Boston, MA, 2004.

[Lin03] Chin-Yew Lin. Improving summarization performance by sentence compression - a pilot study. In *Proceedings of the Sixth International Workshop on Information Retrivial with Asian Language (IRAL 2003)*, Sapporo, Japan, Jully 2003.

[Luh58] H.P. Luhn. The automatic creation of literature abstracts. Journal of research and development, IBM, 1958.

[Man04] Inderjeet Mani. *Narrative Summarization*, volume 45/1. 2004.

[Min04] Jean-Luc Minel. *Le résumé automatique de textes : solutions et perspectives*, volume 45/1. 2004.

[MT88] William C. Mann and Sandra A. Thompson. Rhetorical structure theory : toward a functionnal theory of text organization. In *Research Report RR-87-190, USC/Information Sciences Institute*, pages 243–281, Marina del Rey, CA, 1988.

TALN 2005, Dourdan, 6-10 juin 2005

Amélioration de la segmentation automatique des textes grâce aux connaissances acquises par l'analyse sémantique latente.

Yves Bestgen

Centre pour l'étude du texte et du discours – PSOR - Université catholique de Louvain

Place du Cardinal Mercier, 10 - B1348 Louvain-la-Neuve Belgique

yves.bestgen@psp.ucl.ac.be

Mots-clés : Segmentation automatique de textes, Analyse sémantique latente (ASL)

Keywords: Automatic text segmentation, Latent semantic analysis (LSA)

Résumé Choi, Wiemer-Hastings et Moore (2001) ont proposé d'employer l'analyse sémantique latente (ASL) pour extraire des connaissances sémantiques à partir de corpus afin d'améliorer l'efficacité d'un algorithme de segmentation des textes. En comparant l'efficacité du même algorithme selon qu'il prend en compte des connaissances sémantiques complémentaires ou non, ils ont pu montrer les bénéfices apportés par ces connaissances. Dans leurs expériences cependant, les connaissances sémantiques avaient été extraites d'un corpus qui contenait les textes à segmenter dans la phase de test. Si cette hyperspécificité du corpus d'apprentissage explique la plus grande partie de l'avantage observé, on peut se demander s'il est possible d'employer l'ASL pour extraire des connaissances sémantiques génériques pouvant être employées pour segmenter de nouveaux textes. Les deux expériences présentées ici montrent que la présence dans le corpus d'apprentissage du matériel de test a un effet important, mais également que les connaissances sémantiques génériques dérivées de grands corpus améliorent l'efficacité de la segmentation.

Abstract Choi, Wiemer-Hastings and Moore (2001) proposed to use latent Semantic Analysis to extract semantic knowledge from corpora in order to improve the accuracy of a text segmentation algorithm. By comparing the accuracy of the very same algorithm depending on whether or not it takes into account complementary semantic knowledge, they were able to show the benefit derived from such knowledge. In their experiments, semantic knowledge was, however, acquired from a corpus containing the texts to be segmented in the test phase. If this hyper-specificity of the training corpus explains the largest part of the benefit, one may wonder if it is possible to use LSA to acquire generic semantic knowledge that can be used to segment new texts. The two experiments reported here show that the presence of the test materials in the training corpus has an important effect, but also that the generic semantic knowledge derived from large corpora clearly improves the segmentation accuracy.

1 Améliorer la segmentation des textes par l'adjonction de connaissances sémantiques complémentaires?

Pendant les dix dernières années, de nombreuses méthodes ont été proposées pour segmenter automatiquement des textes en fonction des thèmes qui les composent sur la base de la cohésion lexicale. La distinction principale entre ces méthodes réside dans le contraste entre les approches basées exclusivement sur l'information contenue dans le texte à segmenter comme la cohésion lexicale (par exemple, Choi, 2000 ; Hearst, 1997 ; Heinonen, 1998 ; Kehagias, Pavlina, Petridis, 2003 ; Utiyama, Isahara, 2001), et celles qui reposent sur des connaissances sémantiques complémentaires extraites de dictionnaires et de thesaurus (par exemple, Kozima 1993 ; Lin, Nunamaker, Chau, Chen, 2004 ; Morris, Hirst, 1991), ou des collocations observées dans de grands corpus (Bolshakov, Gelbukh 2001 ; Choi *et al.*, 2001 ; Ferret, 2002 ; Kaufmann, 1999 ; Ponte, Croft, 1997). Selon leurs auteurs, les méthodes qui utilisent des connaissances supplémentaires apportent une réponse aux problèmes posés par les phrases qui relèvent du même thème tout en ne partageant aucun mot commun ou par la présence de synonymes et d'hyperonymes. Des arguments empiriques en faveur de ces méthodes ont été récemment présentés par Choi *et al.* (2001) dans une étude basée sur l'analyse sémantique latente (ASL : Latent semantic analysis, Latent semantic indexing, Deerwester *et al.,* 1990), une technique statistique d'extraction d'espaces sémantiques à partir de corpus qui permet l'estimation de la similarité sémantique entre des mots, des phrases ou des paragraphes. En comparant l'efficacité du même algorithme selon qu'il prend en compte ou non ces connaissances sémantiques complémentaires, Choi *et al.* (2001) ont mis en évidence l'avantage dérivé de telles connaissances.

Toutefois, les implications de l'étude de Choi *et al.* pour la segmentation des textes et, plus généralement, pour l'utilisation de l'ASL dans le traitement automatique du langage sont rendues peu claires en raison de la méthodologie qu'ils ont employée. Dans leurs expériences, les connaissances sémantiques ont été extraites d'un corpus qui contenait les textes qui ont été segmentés dans la phase de test. On peut donc se demander si la plus grande partie des bénéfices obtenus par l'ajout de connaissances sémantiques n'est pas due à cette hyperspécificité du corpus d'apprentissage (c.-à-d. inclure le matériel de test). Si c'est le cas, cela met en question la possibilité d'employer l'ASL pour extraire des connaissances sémantiques génériques pouvant être utilisées pour segmenter de nouveaux textes. A priori, le problème ne semble pas très important, parce que Choi *et al.* ont utilisé un grand nombre de petits échantillons de test pour évaluer leur algorithme, chaque échantillon ne représentant en moyenne que 0.15% du corpus d'apprentissage. La présente étude montre, toutefois, que la présence du matériel de test dans le corpus d'apprentissage a un effet important, mais également que les connaissances sémantiques génériques dérivées de grands corpus améliorent nettement l'efficacité de l'algorithme de segmentation. Cette conclusion est issue de deux expériences dans lesquelles la présence ou l'absence du matériel de test dans le corpus d'apprentissage pour l'ASL est manipulée. La première expérience est basée sur le matériel employé par Choi *et al.*, un petit corpus de 1.000.000 de mots. La deuxième expérience est basée sur un corpus beaucoup plus grand (25.000.000 mots). Avant de présenter ces expériences, l'algorithme et l'utilisation par Choi *et al.* de l'ASL dans ce cadre sont décrits.

2 Les deux versions de l'algorithme de Choi

L'algorithme de segmentation proposé par Choi (2000) se compose des trois étapes habituellement présentes dans les procédures de segmentation basées sur la cohésion lexicale. Premièrement, le document à segmenter est divisé en unités textuelles minimales, habituellement les phrases. Ensuite, un indice de similarité entre chaque paire d'unités prises deux par deux est calculé. Chaque valeur brute de similarité est réexprimée sous une forme ordinale en prenant la proportion de valeurs voisines qui sont plus petites qu'elle. Pour finir, le document est segmenté répétitivement selon les frontières entre les unités qui maximisent la somme des similarités moyennes à l'intérieur des segments ainsi constitués.

L'étape la plus intéressante pour la présente étude est celle qui calcule les similarités interphrases. La procédure initialement proposée par Choi (2000), C99, reposait exclusivement sur l'information contenue dans le texte à segmenter. Chaque phrase est représentée par un vecteur construit selon le modèle vectoriel classique (Manning, Schütze, 1999, pp. 539ff) et la similarité entre deux phrases est calculée au moyen de la mesure de cosinus entre les vecteurs correspondants. Dans une première évaluation basée sur la procédure décrite ci-dessous, Choi a montré que son algorithme était plus efficace que plusieurs autres approches telles que *TextTilling* (Hearst, 1994), *Segmenter* (Kan, Klavans, McKeown, 1998) et le *Maximum-probability segmentation algorithm* de Utiyama et Isahara (2001).

Choi *et al.* (2001) ont proposé d'améliorer la mesure de similarité inter-phrases en prenant en compte les proximités sémantiques entre les mots estimées sur la base de l'analyse sémantique latente (ASL). Brièvement, l'ASL s'appuie sur la thèse qu'il est possible d'estimer la similarité sémantique entre des mots en analysant les contextes dans lesquels ces mots apparaissent (Deerwester *et al.* 1990 ; Degand, Spooren, Bestgen, 2004 ; Landauer, Dumais 1997). La première étape d'une analyse sémantique latente consiste en la construction d'un tableau lexical contenant les fréquences d'occurrence de chaque mot dans chacun des documents, un document pouvant être une phrase, un paragraphe, un texte, ... Pour extraire les dimensions sémantiques, ce tableau lexical subit une décomposition en valeurs singulières, une sorte d'analyse factorielle qui extrait les dimensions orthogonales les plus importantes. Après cette étape, chaque mot est représenté par un vecteur de poids indiquant sa force d'association avec chacune des dimensions. Ceci permet de mesurer la proximité sémantique entre deux mots quelconques en utilisant, par exemple, la mesure de cosinus entre les vecteurs correspondants. La proximité entre deux phrases (ou toutes autres unités textuelles), même lorsque ces phrases ne font pas partir du corpus initial, peut être estimée en calculant un vecteur pour chacune de ces phrases -- correspondant à la somme pondérée des vecteurs des mots qui les composent -- et puis en calculant le cosinus entre ces vecteurs (Deerwester *et al.* 1990). Choi *et al.* (2001) ont montré que l'utilisation de cette procédure pour calculer les similarités inter-phrases produit des performances supérieures à celles enregistrées au moyen de la version précédente de l'algorithme (basé seulement sur la répétition de mots).

3 Expérience 1

Le but de cette expérience est de déterminer l'impact de la présence du matériel de test dans le corpus d'apprentissage de l'ASL sur les résultats obtenus par Choi *et al.* (2001). Est-ce que les

connaissances sémantiques extraites d'un corpus qui n'inclut pas le matériel de test améliorent également l'efficacité de la segmentation ?

3.1 Méthode

Cette expérience est basée sur la méthodologie développée par Choi (2000). Cette méthodologie a été également employée par plusieurs auteurs pour évaluer l'efficacité de leur système de segmentation (Brants, Chen, Tsochantaridis, 2002 ; Ferret, 2002 ; Kehagias *et al.*, 2003 ; Utiyama, Isahara, 2001). La tâche consiste à retrouver les frontières entre des textes concaténés. Chaque échantillon de test est une concaténation de dix segments de textes. Chaque segment est composé des *n* premières phrases d'un texte aléatoirement choisi dans deux sous-sections du corpus de Brown. Pour l'expérience, j'ai utilisé le matériel de test le plus général proposé par Choi (2000) dans lequel la taille des segments dans chaque échantillon varie aléatoirement de 3 à 11 phrases. Il est composé de 400 échantillons.

L'expérience vise à comparer l'efficacité de l'algorithme selon que le matériel de test est inclus dans le corpus d'apprentissage de l'ASL (condition *non autonome*) ou qu'il ne l'est pas (condition *autonome*). Un espace sémantique *non autonome*, correspondant à celui utilisé par Choi *et al.*, a été construit en utilisant l'entièreté du corpus de Brown comme corpus d'apprentissage. Quatre cents espaces *autonomes* différents ont été construits, un pour chaque échantillon de test, en retirant à chaque fois du corpus de Brown uniquement les phrases qui composent cet échantillon.

Pour extraire l'espace sémantique par l'ASL et pour appliquer l'algorithme de segmentation, une série de paramètres ont dû être fixés. Tout d'abord, les paragraphes ont été utilisés comme documents pour construire le tableau lexical parce que Choi *et al.* ont observé que de telles unités de taille moyenne étaient plus efficaces que des unités plus courtes comme les phrases. Les mots repris dans la liste de mots-outils (*stoplist*) de Choi ont été supprimés, ainsi que ceux qui n'apparaissaient qu'une seule fois dans l'ensemble du corpus. Les mots n'ont pas été tronqués en fonction de leur racine (*stemming*), suivant en cela la procédure de Choi *et al.* (2001). Pour établir l'espace sémantique, la décomposition en valeurs singulières a été réalisée par le programme SVDPACKC (Berry, 1992 ; Berry *et al.*, 1993), et les 300 premiers vecteurs singuliers ont été conservés. En ce qui concerne l'algorithme de segmentation, j'ai utilisé la version dans laquelle le nombre de frontières à trouver est imposé et fixé ici à neuf. Un masque de 11 x 11 a été employé pour la transformation ordinale, comme recommandé par Choi (2000).

3.2 Résultats

L'efficacité de la segmentation a été évaluée au moyen de l'indice utilisé par Choi *et al.* (2001) : le taux Pk d'erreur de segmentation (Beeferman, Berger, Lafferty, 1999) qui indique la proportion de phrases qui sont incorrectement classées comme appartenant au même segment ou incorrectement classées comme appartenant à des segments différents.

Les résultats[1] sont présentés dans la Figure 1. Les espaces autonomes donnent lieu à des performances plus faibles que l'espace non autonome, comme le confirme le test *t* pour échantillon appareillé (chaque échantillon de test étant utilisé comme une observation) qui est significatif pour un alpha plus petit que 0.0001. L'algorithme C99, qui n'utilise pas l'ASL pour estimer les similarités entre les phrases, produit un Pk de 0.13 (Choi *et al.*, 2001, tableau 3, ligne 3 : *no stemming*). Il s'avère donc que, si la condition *autonome* est meilleure que C99, l'avantage est très faible.

	Pk
Non autonome	0.084 (0.005)
Autonome	0.120 (0.006)

Figure 1 : Taux d'erreur et variance (entre parenthèses) pour les conditions non autonome et autonome.

Avant de conclure que la présence du matériel de test dans le corpus d'apprentissage de l'ASL a fortement modifié l'espace sémantique, une explication alternative doit être considérée. La perte d'efficacité en condition *autonome* pourrait être due au fait qu'il y a systématiquement légèrement moins de mots indexés dans les espaces sémantiques autonomes que dans l'espace *non autonome*. La suppression de chaque échantillon de test a entraîné la perte en moyenne de 23 mots différents sur un total de 25.847 mots qui sont indexés dans l'espace *non autonome*. Dans les espaces *autonomes*, ces mots ne sont pas disponibles pour estimer la similarité entre les phrases, tandis qu'ils sont utilisés dans l'espace *non autonome*. Afin de déterminer si ce facteur peut expliquer la différence d'efficacité, une analyse complémentaire a été effectuée sur l'espace *non autonome* dans laquelle, pour chaque échantillon de test, uniquement les mots présents dans l'espace *autonome* correspondant ont été pris en compte. De cette manière, seules les relations sémantiques peuvent jouer. Comparé à l'espace *non autonome* complet, je n'ai observé presque aucune diminution d'efficacité, le taux d'erreur Pk passant de 0.084 à 0.085 dans la nouvelle analyse. Ce résultat indique que ce ne sont pas les mots choisis pour le calcul des proximités qui importent, mais les relations sémantiques dans les espaces.

4 Expérience 2

L'expérience 1 a été menée sur le corpus d'apprentissage de Choi *et al.* (2001), un corpus de 1.000.000 de mots issus de textes de genres et de thèmes très différents. La petite taille du corpus et la diversité des textes pourraient avoir affecté les résultats de deux manières. D'abord, l'impact de la présence du matériel de test dans le corpus dépend probablement de ces caractéristiques du corpus. Retirer les premières phrases d'un texte devrait avoir moins d'effet si le corpus contient

[1] Ce taux d'erreur est en fait légèrement meilleur que celui obtenu par Choi et al. (2001), la différence pouvant être due à plusieurs facteurs tels que le prétraitement du corpus de Brown (identification des mots et des paragraphes) ou la fonction de pondération appliquée aux fréquences brutes, qui était ici la formule de pondération décrite dans Landauer, Foltz, et Laham (1998).

beaucoup de textes sur des thèmes similaires. En second lieu, un corpus plus volumineux permettrait probablement l'extraction d'un espace sémantique plus stable et plus efficace. Ceci pourrait produire une plus grande différence entre la version "ASL" de l'algorithme et celle qui n'utilise pas de connaissances sémantiques supplémentaires (C99). Pour ces raisons, une deuxième expérience a été menée sur la base d'un corpus beaucoup plus volumineux, comprenant les articles parus durant les années 1997 et 1998 dans le journal de langue française belge *Le Soir* (approximativement 52.000 articles et 26.000.000 mots). Dans ce corpus, chaque échantillon du matériel de test correspond en moyenne à 0.0066% du corpus complet. Cette deuxième expérience a également permis de comparer les conditions *non autonome* et *autonome* à une condition *antérieure* basée sur les articles parus dans le même journal, mais pendant les années 1995 et 1996 (approximativement 50.000 articles et plus de 22.000.000 mots). Cette condition nous informera à propos de la possibilité d'employer l'ASL pour extraire des connaissances sémantiques plus génériques, puisque le corpus d'apprentissage de l'ASL est antérieur aux textes à segmenter. Il faut toutefois noter que ces connaissances étant extraites de la même source journalistique, les qualifier d'indépendantes seraient nettement excessifs.

4.1 Méthode

Le matériel de test a été extrait du corpus 1997-1998 suivant les directives données dans Choi (2000). Il se compose de 100 échantillons de dix segments, dont la longueur varie aléatoirement de 3 à 11 phrases. Trois types d'espace d'apprentissage pour l'ASL ont été construits. L'espace *non autonome* est basé sur l'entièreté du corpus 1997-1998. Cent espaces *autonomes* différents ont été construits comme décrit dans l'expérience 1. Enfin, un espace *antérieur* a été établi à partir du corpus 1995-1996. Les paramètres utilisés pour extraire les espaces sémantiques sont identiques à ceux employés dans l'expérience 1 sauf que, pour réduire la taille des tableaux lexicaux. les articles, et non les paragraphes, ont été utilisés comme documents et les mots ont été lemmatisés au moyen de TreeTagger (Schmid 1994).

4.2 Résultats

Globalement, les résultats sont similaires à ceux obtenus lors de la première expérience. Comme le montre la Figure 2, les espaces autonomes donnent lieu à des performances plus faibles que l'espace non autonome et, comme attendu, l'espace antérieur donne lieu à des performances encore plus faibles.

	Pk
Non autonome	0.074 (0.004)
Autonome	0.084 (0.005)
Antérieure	0.098 (0.005)

Figure 2 : Taux d'erreur et variance (entre parenthèses) pour les conditions non autonome, autonome et antérieure.

Toutefois, il est important de noter que l'algorithme C99, qui n'est pas basé sur l'analyse sémantique latente, produit un taux d'erreur Pk de 0.155, soit une valeur nettement plus mauvaise que celles obtenues avec les espaces *autonomes* (0.084) et avec l'espace *antérieur* (0.098). Ceci confirme l'utilité des connaissances sémantiques extraites de grands corpus pour estimer les similarités interphrases.

Par rapport à la première expérience, l'écart entre les conditions *non autonome* et *autonome* est beaucoup plus faible, passant de 0.036 pour l'expérience 1 à 0.01 pour l'expérience 2. Cet écart demeure néanmoins statistiquement significatif ($t(99) = 3.17$; $p = 0.002$). Bien que plus grand, l'écart entre les conditions *autonome* et *antérieure* (0.014), est statistiquement juste significatif ($t(99) = 2.04$; $p = 0.045$). La Figure 3 montre que la condition *autonome* surpasse la condition *antérieure* dans 46 échantillons, alors que l'inverse se produit dans 35 échantillons, les 19 échantillons restants ne montrant aucune différence entre ces deux conditions.

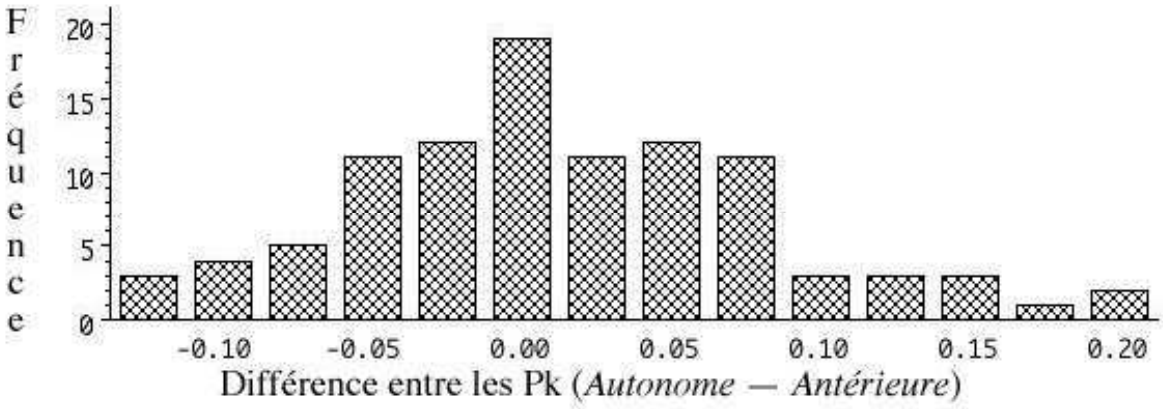


Figure 3 : Distribution des différences en Pk entre les conditions *Autonome* et *Antérieure*

On voit donc que l'avantage de la condition *autonome* sur la condition *antérieure* est principalement dû à quelques échantillons de test pour lesquels la condition *autonome* est nettement plus efficace. Rappelons également que la condition *antérieure* donne lieu à des résultats nettement meilleurs que ceux obtenus lorsque la segmentation s'effectue sans le recours à des connaissances sémantiques complémentaires.

5 Conclusion

Les deux expériences rapportées ici montrent que la présence du matériel de test dans le corpus d'apprentissage de l'ASL augmente l'efficacité de l'algorithme de segmentation même lorsqu'un corpus de plus de 25.000.000 mots est utilisé. Elles montrent également que l'utilisation de connaissances sémantiques indépendantes améliore l'efficacité de la segmentation et que ceci s'observe même lorsque ces connaissances sont extraites d'années antérieures de la même source. Cette observation souligne la possibilité de constituer par analyse sémantique latente des connaissances sémantiques plus ou moins génériques, c'est-à-dire, des connaissances qui peuvent être utilisées pour traiter de nouvelles données, comme cela a été récemment proposé dans la recherche de l'antécédent d'une anaphore, dans un système de reconnaissance de la parole ou en

traduction automatique (Bellegarda, 2000 ; Klebanov, Wiemer-Hastings, 2002 ; Kim, Chang, Zhang, 2003). Une question à laquelle la présente étude ne répond pas concerne la possibilité d'utiliser un corpus tiré d'une autre source, telle qu'un autre journal. Bellegarda (2000) a observé, en reconnaissance automatique de la parole, qu'un tel espace sémantique est moins efficace. Cependant, évaluer la proximité sémantique entre deux phrases est probablement moins affecté par la source du corpus que de prédire le prochain mot d'un énoncé.

Récemment, plusieurs auteurs ont proposé des algorithmes de segmentation, basés principalement sur la programmation dynamique, qui égalent ou même surpassent les résultats de Choi (Ji, Zha, 2003, Kehagias *et al.,* 2003 ; Utiyama, Isahara, 2001). Ces algorithmes ne s'appuient pas sur des connaissances sémantiques supplémentaires. Les résultats de la présente étude suggèrent que leur efficacité pourrait encore être améliorée en prenant en compte de telles connaissances. Enfin, d'autres techniques que l'ASL ont été proposées pour extraire des connaissances sémantiques à partir de grands corpus tel pASL (Brants *et al.*, 2002). L'analyse sémantique latente étant relativement simple à mettre en pratique grâce à la disponibilité de programmes très puissants tel que SVDPACKC (Berry *et al.*, 1993), son avantage principal est qu'elle est employée par une communauté de plus en plus large de chercheurs.

Une limitation importante de ce travail réside dans la tâche employée pour évaluer l'efficacité de l'algorithme de segmentation. Identifier les frontières entre des textes concaténés est une tâche artificielle et certainement moins complexe que de localiser les changements de thèmes à l'intérieur de textes. Le fait que la procédure et le matériel conçu par Choi soient devenus une sorte de "standard" employé par une série de chercheurs pour évaluer leur algorithme ne suffit pas à la légitimer. Il serait donc utile de confirmer les conclusions de la présente étude dans une tâche de segmentation intratexte.

Remerciements

Y. Bestgen est chercheur qualifié du Fonds National de la Recherche (FNRS). Cette recherche est financée par le projet FRFC n° 2.4535.02 et par une “Action de Recherche concertée” du Gouvernement de la Communauté française de Belgique.

Références

BEEFERMAN D., BERGER A., LAFFERTY J. (1999). Statistical models for text segmentation, *Machine Learning, Vol.* 34, pp. 177–210.

BELLEGARDA J. (2000). Large vocabulary speech recognition with multispan statistical language models. *IEEE Transactions on Speech and Audio Processing*, Vol. 8, pp. 78-84.

BERRY M. (1992). Large scale singular value computation. *International jornal of Supercomputer Application*, Vol. 6, 13-49.

BERRY M., DO T., O'BRIEN G., KRISHNA V., VARADHAN S. (1993). SVDPACKC: Version 1.0 user's guide, Tech. Rep. CS-93-194, University of Tennessee, Knoxville, TN, October 1993.

BOLSHAKOV I., GELBUKH A. (2001). Text segmentation into paragraphs based on local text cohesion. In *Proceedings of Text, Speech and Dialogue (TSD-2001)*, 158–166.

BRANTS T., CHEN, F., TSOCHANTARIDIS I. (2002). Topic-based document segmentation with probabilistic latent semantic analysis. In *Proceedings of CIKM'02*, 211-218

CHOI F. (2000). Advances in domain independent linear text segmentation, In *Proceedings of NAACL-00*, 26–33.

CHOI F., WIEMER-HASTINGS P., MOORE J. (2001) Latent semantic analysis for text segmentation, In *Proceedings of NAACL'01*, 109–117.

DEERWESTER S., DUMAIS S., FURNAS G., LANDAUER T., HARSHMAN R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, Vol. 41, pp. 391-407.

DEGAND L., SPOOREN W., BESTGEN Y. (2004). On the use of automatic tools for large scale semantic analyses of causal connectives. In *Proceedings of ACL 2004 Workshop on Discourse Annotation*, 25-32.

FERRET O. (2002). Using collocations for topic segmentation and link detection. In *Proceedings of COLING 2002*, 260-266.

HEARST M. (1997). TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, Vol. 23, pp. 33–64.

HEINONEN O. (1998). Optimal multi-paragraph text segmentation by dynamic programming. In *Proceedings of 17th International Conference on Computational Linguistics (COLING-ACL'98)*, 1484-1486.

JI X., ZHA H. (2003). Domain-independent text segmentation using anisotropic diffusion and dynamic programming. In *Proceedings of SIGIR 2003*, 322-329.

KAN M., KLAVANS J., MCKEOWN K. (1998). Linear segmentation and segment significance. In *Proceedings of the 6th International Workshop of Very Large Corpora*, 197-205.

KAUFMANN, S. (1999). Cohesion and collocation: using context vectors in text segmentation. In *Proceedings of ACL'99*, 591–595.

KEHAGIAS A., PAVLINA F., PETRIDIS V. (2003). Linear text segmentation using a dynamic programming algorithm. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, 171-178

KIM Y., CHANG J., ZHANG B. (2003). An empirical study on dimensionality optimization in text mining for linguistic. Knowledge Acquisition. In *Proceedings of PAKDD 2003*, 111–116.

KLEBANOV B., WIEMER-HASTINGS P. (2002). Using ASL for pronominal anaphora resolution. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, 197-199.

KOZIMA H. (1993). Text segmentation based on similarity between words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, 286-288.

LANDAUER T., DUMAIS S. (1997). A solution to Plato's problem : the latent semantic analysis theory of acquisition, induction and representation of knowledge, *Psychological Review*, Vol. 104, pp. 211–240.

LANDAUER T., FOLTZ P., LAHAM D. (1998). An Introduction to latent semantic analysis. *Discourse Processes*, Vol. 25, pp. 259-284.

LIN M., NUNAMAKER J., CHAU, M., CHEN H. (2004). Segmentation of lecture videos based on text: A method combining multiple linguistic features. In *Proceedings of the 37th Hawaii International Conference on System Sciences*.

MANNING C., SCHÜTZE H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.

MORRIS J., HIRST G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics,* 17: 21-42.

PEVZNER L., HEARST M. (2002). A Critique and improvement of an evaluation metric for text segmentation, *Computational Linguistics,* Vol 28, pp. 19-36

PONTE J., CROFT W. (1997). Text segmentation by topic. *Proceedings of the 1st European Conference on Research and Advanced Technology for Digital Libraries*, 120-129.

SCHMID H. (1994). Probabilistic Part-of-speech tagging using decision trees. P*roceedings of International Conference on New Methods in Language Processing* .

UTIYAMA M., ISAHARA H. (2001). A Statistical model for domain-independent text segmentation. *Proceedings of ACL'2001*, 491–498.

Détection Automatique de Structures Fines de Texte

Nicolas Hernandez et Brigitte Grau
LIMSI/CNRS - LIR – Université de Paris-Sud
BP 133, F-91403 ORSAY CEDEX (France)
Hernandez|Grau@limsi.fr

Mots-clefs : Navigation intra-documentaire, analyse thématique, structures du discours, relations discursives, subordination et coordination, parallélisme lexico-syntaxico-sémantique, modèle d'apprentissage, analyses linguistiques

Keywords: Text browsing, topic analysis, text structures, discursive relations, subordination and coordination, lexical, syntactic and semantic parallelism, learning model, linguistic analysis

Résumé Dans ce papier, nous présentons un système de Détection de Structures fines de Texte (appelé *DST*). DST utilise un modèle prédictif obtenu par un algorithme d'apprentissage qui, pour une configuration d'indices discursifs donnés, prédit le type de relation de dépendance existant entre deux énoncés. Trois types d'indices discursifs ont été considérés (des relations lexicales, des connecteurs et un parallélisme syntaxico-sémantique) ; leur repérage repose sur des heuristiques. Nous montrons que notre système se classe parmi les plus performants.

Abstract In this paper, we present a system which aims at detecting fine-grained text structures (we call it *DST*). Based on discursive clues, DST uses a learning model to predict dependency relations between two given utterances. As discourse clues, we consider lexical relations, connectors and key phrases, and parallelism. We show that our system implements an improvement over current systems.

1 Introduction

Comme le souligne l'annonce du 14 décembre 2004 de la société Google de numériser et de rendre disponible en ligne 15 millions de livres appartenants à 5 des plus célèbres bibliothèques anglo-saxonnes du monde[1], le besoin d'accéder facilement et rapidement au contenu d'un document électronique est plus que jamais un enjeu d'actualité.

Dans ce papier, nous nous intéressons à la détection de l'organisation du contenu informationnel d'un document textuel. De nombreux travaux (principalement au sein de la comunauté de résumé automatique) ont montré l'intérêt d'appréhender la structure d'un texte : afin de manipuler des unités de texte de différentes granularités (i.e. différents degrés informationnels), de fournir un contexte à une information ciblée, de permettre une navigation intra-documentaire, etc. (Moens & Busser, 2001; Choi, 2002; Couto *et al.*, 2004).

En particulier nous nous focalisons sur la micro-structure d'un texte (niveau phrastique voire propositionnel). Nous affichons ainsi une complémentarité aux approches globales tout en offrant la possibilité de raffiner leur modèle. En effet qu'elles supposent une organisation plate et linéaire du flot d'informations communiqué (Hearst, 1997; Choi, 2002), ou bien une organisation plus riche en arbres (Moens & Busser, 2001; Couto *et al.*, 2004), les approches globales sont généralement fondées sur des mesures de cohésion lexicale (notamment à travers le suivi de chaînes lexicales) qui souffrent d'un manque de précision quant à la délimitation des unités de texte (appelées segment). De plus elles prennent rarement en compte dans leur analyse les phénomènes discursifs locaux (e.g. annonces thématiques – e.g. "Les points que nous allons traiter sont :", structures énumératives, transitions, etc.).

Notre approche se situe parmi les travaux qui proposent de rechercher le point d'attache optimal d'un énoncé entrant dans la structure en cours de construction. Parmi les approches existantes, Marcu (1999) propose un système pour la détection automatique de la structure rhétorique d'un texte, Choi (2002) s'intéresse à une structuration thématique fine, Kruijff-Korbayová & Kruijff (1996) analysent le discours en terme de progression thématique. Ces systèmes constituent de sérieuses avancées mais requièrent encore la prise en compte de plus d'indices discursifs et de modèles plus souples pour appréhender les différents mécanismes de structuration du discours.

Dans ce papier, nous présentons un système de Détection de Structures fines de Texte (appelé *DST*). DST utilise un modèle prédictif obtenu par un algorithme d'apprentissage qui, pour une configuration d'indices discursifs donnés, prédit le type de relation de dépendance existant entre deux énoncés. L'originalité principale de notre approche est de proposer un modèle Théorique simplifié de la Structure du Discours. En effet, nous nous intéressons seulement au rapport structurel élémentaire liant deux énoncés (relation de subordination, de coordination, et absence de relation) indépendamment d'un éventuel étiquetage sémantico-rhétorique de la relation[2]. Le fait de dissocier le modèle de dépendance de la recherche du point d'attache de l'énoncé entrant nous permet d'envisager différents algorithmes de structuration. Une de nos particularités techniques est de proposer une mesure pour appréhender le parallélisme syntaxico-sémantique de deux énoncés, indice discursif peu considéré jusqu'à présent. Nous avons travaillé sur des articles scientifiques en anglais mais notre démarche est adaptable à d'autres langues comme le français.

[1] New York Public Library, University of Michigan, Stanford, Harvard (USA), Oxford (GB).

[2] Cette tâche sera abordée ultérieurement.

2 L'accès au contenu

Le processus de compréhension requiert d'une part *d'identifier des unités discursives (informationnelles, intentionnelles, ayant une mise en forme visuelle, ou autres)* et d'autre part *d'établir des relations entre ces unités.* Cette reconnaissance de la cohérence peut nécessiter des connaissances sémantiques et pragmatiques, non disponibles dans le texte. Néanmoins nous partons du postulat qu'il est possible de mettre en place des analyses automatiques à partir des indices du discours (chaînes lexicales, connecteurs, introducteurs de cadres, etc.) pour permettre de reconnaître cette cohérence.

L'une des caractéristiques majeures transversale à la plupart des théories du discours est la considération d'un modèle de dépendance qui définit la nature de la relation structurelle existante entre deux énoncés en terme de *subordination* ou de *coordination* (Mann & Thompson, 1987; Polanyi, 1988; Virbel, 1989; Asher & Lascarides, 1994). Les différences entre théories viennent de la signification qu'elles donnent à la nature de ces relations, mais aussi des contraintes structurelles d'assemblage des unités discursives. Au niveau de la micro-structure, les chercheurs ont tendance à considérer que l'unité élémentaire de référence est proche de celle de la proposition (Mann & Thompson, 1987; Polanyi, 1988). Afin de faciliter le repérage automatique, nous considérons comme Choi (2002) la phrase syntaxique comme unité élémentaire.

Suivant le genre de texte considéré (expositif, narratif, dialogue, etc.), les théories du discours mettent en évidence un ou plusieurs plans d'organisation de l'information: rhétorique, logico-visuelle, informationnelle, etc. Les interactions entre ces différentes structures sont encore très floues, c'est pourquoi nous avons décidé de nous concentrer sur le plan informationnel que nous considérons comme pertinent pour les textes scientifiques. Notre description du plan informationnel repose sur la théorie de la RST[3] (Mann & Thompson, 1987), le LDM (Polanyi, 1988), et aussi la progression thématique de la phrase au discours (Kruijff-Korbayová & Kruijff, 1996). Globalement cela signifie que la relation entre deux énoncés est déterminée en fonction de leur contenu informationnel indépendamment de l'intention rhétorique de l'auteur.

Dans notre modèle, un énoncé entrant se rattache au discours selon une relation de subordination ou de coordination (ou bien les deux). Un énoncé est interprété en fonction de son thème (ce dont il parle), de son propos (ce qui est dit au sujet du thème) et de sa fonction sémantico-rhétorique. Ces éléments sont identifiés à partir d'indices présents dans l'énoncé et dans son contexte ce qui permet de déduire avec quelles parties du discours il est lié et comment.

Nous illustrons ces relations à l'aide du texte de la figure 1 extrait d'un passage de notre corpus. Les indices discursifs aisément représentables visuellement sont soulignés dans le texte. Les couples d'énoncés (1, 2) et (1, 6) décrivent des relations de subordination. 1 est un modèle classique d'annonce thématique avec un quantifieur *two*, une phrase syntaxiquement incomplète et un caractère de ponctuation annonce ":". Les énoncés 2 et 6, quant à eux, contiennent des marques qui caractérisent des items d'une énumération ("*1.*" et "2."). Ces deux énoncés présentent aussi une relation de coordination explicite l'un envers l'autre, soulignée notamment par un parallélisme syntaxique :

NUM. NOM, *whereby* ADJ NOM be+conj='présent' VERB+conj='participe passé' PREP

Le couple d'énoncés (2, 3) constitue un exemple de subordination où le deuxième énoncé 3 correspond au développement d'un des aspects du premier. Cette subordination est marquée par une progression thématique de rhème en thème (i.e. le terme *importance* qui est repris dans 3). Le couple d'énoncés (4, 5) décrit, quant à lui, un exemple de coordination implicite.

[3] *Rhetorical Structure Theory* (RST), *Linguistic Discourse Model* (LDM).

Two traditional approaches to automatic abstracting are: (1)
1. Extraction, whereby specific sentences are selected from the source text according to some assessment of their importance. (2)
Importance indicators include the concentration of topic-relevant terms [...]; the occurrence of expressions, such as "important", "to sum up"; and the position of the sentence within the text. (3)
This approach is exemplified by Pollock and Zamora's ADAM system [1]. (4)
The problems with this approach are that importance clues are often not reliable, and that the extracted sentences do not always constitute a coherent text, since they often contain cross-references. (5)
2. Summarisation, whereby detailed semantic analysis is applied to the text, and a representation such as a semantic net is produced, from which a summary is then generated. (6)

Figure 1: Exemples de relations de subordination, et de coordination explicite et implicite

En effet, tous les deux sont subordonnés à une même entité, *l'approche en terme d'extraction*, et chacun d'eux en traite de manière indépendante, le premier en présentant un exemple et le second en décrivant les problèmes.

3 Algorithme de structuration "*shift and reduce*"

La structure de texte en arbre unique est une simplification de la réalité, néanmoins nous adoptons une modélisation *hiérarchique* parce qu'elle reste la plus communément rencontrée dans les textes. Notre algorithme de structuration reprend le principe des algorithmes de Marcu (1999) et Choi (2002). Nous l'avons adapté afin de tenir compte de la relation de coordination. Cet algorithme construit une structure hiérarchique du discours dont les arcs sont orientées vers les énoncés entrants toujours attachés sur la frontière droite de l'arbre. Un énoncé entrant coordonné à un énoncé de la structure est considéré comme étant subordonné au même énoncé que l'énoncé auquel il est coordonné. Un nœud factice joue le rôle de père de tous les nœuds.

L'algorithme utilise deux structures de données : une pile qui stocke la branche "frontière droite" de l'arbre en cours de construction (le dernier élément empilé est le point d'attache le plus prioritaire), et une file qui contient la liste des énoncés tels qu'ils sont ordonnés dans le texte et analysés successivement. La pile joue un rôle de mémoire dont chaque élément correspond à une granularité inférieure obtenue dans la structure du discours. L'objectif est d'identifier les énoncés qui sont liés et les relations qu'ils entretiennent.

Algorithme :

1. Si la pile est vide, on défile la file et empile la pile (état initial).
2. Tant que la pile et la file ne sont pas vides, calcul de la relation entre l'élément au sommet de la pile et le premier élément de la file.
   - Si une relation de subordination est détectée, alors l'élément de la file est défilé et empilé (on descend dans la granularité du texte) ;
   - Si une relation de coordination est détectée, alors l'élément au sommet de la pile est dépilé et remplacé par l'élément de la file ;
   - Sinon (aucune relation) l'élément au sommet de la pile est dépilé et écarté (l'idée étant de remonter jusqu'au niveau de dépendance lié à l'élément en tête de file).

4 Indices discursifs

La reconnaissance des relations discursives entre deux énoncés est fondée sur la présence, ou l'absence, d'indices significatifs dans les textes scientifiques : relations lexicales, expressions clefs et parallélisme de construction.

4.1 Relations lexicales

Les relations lexicales entre deux énoncés sont envisagées selon leur nature sémantique et selon les parties des énoncés concernées (thème ou rhème). Nous utilisons un module de Construction de Chaînes Lexicales (CCL) pour les repérer. Celui-ci est fondé sur une variante de l'algorithme de (Barzilay & Elhadad, 1997). CCL recherche les relations entre les lemmes associés aux paires de mots étudiés en tenant compte de la distance sémantique entre ces mots ainsi que de leur distance dans le texte. Le mot le plus fréquent au sein d'une chaîne constitue son élément représentatif. Nous considérons :

- Les relations morphologiques : deux mots appartenant à la même famille morphologique[4], indépendamment de leur catégorie lexicale ;
- Les relations utilisées pour référer à un même objet du discours, telles que la synonymie, l'hyperonymie et l'hyponymie, la méronymie et l'holonymie, trouvées dans WordNet ;
- Les relations d'antonymie trouvées grâce à WordNet ou à la présence de préfixes tels que *dis-, in-, un-, non-, under-, im-, a-, de-, ir-, anti-* sur les mêmes lemmes. Nous construisons des chaînes lexicales spécifiques à ce type de relation.

Etant données deux phrases constituant le contexte d'études, des chaînes lexicales sont calculées entre les deux phrases, globalement, et entre les différentes combinaisons des parties constituant le thème et le rhème des deux phrases. La distinction entre les parties thématique et rhématique d'une phrase est réalisée selon une heuristique robuste de découpage de la phrase en deux par rapport au verbe le plus proche de son milieu.

La présence d'un lien lexical entre les rhèmes de deux énoncés traduit généralement une subordination du deuxième énoncé vis-à-vis de l'énoncé précédent (e.g. une élaboration ou une reformulation). Une progression linéaire, de rhème en thème, correspond aussi au même type de subordination (e.g. une annonce thématique). Une relation contrastive peut dénoter une coordination. Dans tous les autres cas, la présence ou l'absence d'une relation lexicale constitue un indice supplémentaire qui pourra se combiner avec les suivants.

4.2 Expressions clefs (essentiellement des connecteurs)

Notre liste d'expressions clefs est issue en partie de la liste de méta-descripteurs acquise automatiquement par Hernandez & Grau (2003), et de l'analyse de notre corpus. Nous l'avons aussi complétée à partir des mots clefs fournis par Choi (2002) pour lesquels nous réassignons la relation (subordination ou coordination) en fonction de nos observations personnelles.

En raison du nombre d'exemples que compte notre corpus (1190 couples de phrases) par rapport au nombre de marques que nous avons retenu (178), nous n'avons pas choisi de considérer chacune des marques comme une caractéristique distincte, au contraire de Choi dont le modèle

[4]Nous utilisons la base CELEX (*www.ldc.upenn.edu/readme_files/celex.readme.html*).

compte 19 marques. Nous avons opté pour une pré-classification de celles-ci en 5 classes en fonction de leur comportement pour structurer le discours et réduit ainsi la complexité du nombre d'indices. Les classes rassemblent des marques ayant un même "comportement" structurel vis-à-vis de la subordination et de la coordination entre deux énoncés. Les classes que nous avons définies sont les suivantes :

- Initie : marque le premier item d'une liste d'items (suppose une coordination) : "*former, first, on the one hand, '1.', 'a)', begin, start*" ;
- Continue : Coordonne mais n'initie pas une liste et n'en termine pas forcément une : "*second, another, other, also, and, or, however, but, then, in addition, although, etc.*" ;
- Termine : Marque le dernier item d'une liste (suppose une coordination) : "*on the other hand, last, finally, to conclude, to sum up, end, finish, latter, in conclusion, result*" ;
- Subordonné : Apparaît en début d'un énoncé subordonné : "*so, the, this, these, it, he, by this, consequently, for example, for instance, therefore, thus, note that, such, etc.* " ;
- Subordonnant : Apparaît en fin d'un énoncé subordonnant (i.e. en général une annonce) : "*such as, follow, as follow, see below, below, and, :, ?, etc.* ".

Une marque est dans une seule classe sauf si cette dernière permet de la distinguer dans sa définition (e.g. la position dans la phrase pour différencier les marques *subordonnées* des marques *subordonnantes*). Les notions de début et de fin sont relatives à chaque énoncé et correspondent à une distance en nombre de mots exprimée en pourcentage (respectivement fixée à 40% du début ou de la fin). La taille maximale est fixée à 10 tokens.

En plus de ces classes de marques discursives, nous rajoutons une classe de marques désignant la négation (e.g. *aren't, can't, nothing, nobody, rarely, etc.*) et afin de prendre en compte les formes passives, et l'inversion des parties thématiques et rhématiques qui en découle, nous considérons la présence du verbe "être" suivi directement d'un autre verbe comme une caractéristique. Par la suite, nous appellerons ces dernières caractéristiques les *indices syntaxiques*. Au final, la distribution de nos 178 marques se répartit ainsi : 7 marques pour la classe Initie, 38 pour la classe Continue, 9 pour la classe Termine, 62 pour la classe Subordonnée, 30 pour la classe Subordonnant, 31 marques de négation et 1 marque du verbe *être*.

4.3 Parallélisme

Le parallélisme de construction entre deux énoncés rend compte d'une importance égale (lien de coordination) (Hernandez, 2004). Il se traduit par a) des similarités des constituants à différents niveaux paradigmatiques (lemme, trait sémantique, catégorie grammaticale, fonction syntaxique) ; b) une similarité syntagmatique qui s'exprime à la fois par une similarité dans l'ordre des constituants parallèles et par une similarité dans les écarts de distance entre ces mêmes constituants.

Afin de calculer le *degré de parallélisme* entre deux énoncés, nous réduisons la complexité du problème d'abord en homogénéisant les entités du discours (chaque mot est remplacé par l'élément représentatif de la chaîne lexicale à laquelle il appartient). Ensuite, chaque structure syntaxique hiérarchique est remplacée par une liste plate, qui correspond à une notation préfixée de l'arbre (les nœuds internes, qui sont des étiquettes, sont placés avant les feuilles, qui sont les lemmes). Cette liste est obtenue à partir du résultat d'analyse fourni par l'analyseur statistique de Charniak (1997)[5], en supprimant les niveaux de parenthèses.

[5]Nous utilisons la version 2001, développée pour l'anglais à l'université de Brown.

Pour tout couple de phrases donné, le système calcule un degré de parallélisme entre toutes les séquences extraites de chacune des phrases, comportant le même nombre d'items similaires, au minimum deux, différents ou non, placés dans leur ordre d'apparition dans les phrases. Par exemple, les phrases *cabcad* et *acba* partagent 4 constituants : *c*, *a* deux fois et *b*. Une fois supprimés les constituants non similaires (i.e. *d*), on extrait de la première phrase *caba* et *abca*, et de la deuxième *acba*. On ne tient pas compte des éléments différents, qui peuvent être insérés n'importe où dans les phrases. Le parallélisme est fondé sur des constructions similaires d'éléments similaires. La mesure que nous avons définie s'inspire des mesures de distances d'édition entre des séquences de caractères. Chaque constituant est identifié de manière unique par sa position. Plus un constituant est distant de son symétrique dans l'autre séquence, plus les séquences comparées diffèrent. Elle est définie par la formule suivante :

$$degreDeParallelisme(s_m, s_n) = \sum_{i=1}^{l(s)}((\frac{D(s) - d(x_i)}{D(s)}))$$

avec x_i, le i^{eme} constituant de la séquence s_m, $l(s)$, la longueur des séquences comparées, $D(s)$, la distance maximale possible entre un constituant d'une séquence s et son constituant parallèle i.e. $D(s) = l(s) - 1$, et d, la distance effective d'un constituant courant de la séquence s_m et son constituant parallèle. Le degré de parallélisme d'un couple d'énoncés correspond au degré maximal obtenu pour les séquences extraites de ces énoncés.

5 Apprentissage des relations discursives

Afin de reconnaître les relations discursives, nous avons décidé d'opter, de même que Marcu (1999), pour un apprentissage par arbre de décision qui possède l'avantage d'être compréhensible par tout utilisateur (si la taille de l'arbre produit est raisonnable) et d'avoir une traduction immédiate en terme de règles de décision. Nous avons utilisé le classifieur C4.5 fourni dans le logiciel WEKA[6]. Les caractéristiques que nous venons de décrire sont au nombre de 22 et sont repérées automatiquement dans le corpus.

5.1 Données

Afin de constituer un ensemble de couples de phrases et de relations correspondantes, nous avons manuellement annoté un corpus de 5 documents anglais appartenant au domaine de la linguistique informatique. Ils font tous entre 8 et 10 pages et sont au format pdf. L'un d'eux est en simple colonne. De fait ils couvrent la période 1998 et 1999 et aucun d'eux ne partage de références communes. Ces articles sont Mitkov (COLING-ACL'98), Kan et al. (WVLC'98), Green (ACL'98), Sanderson et al. (SIGIR'99) et Oakes et al. (IRSG'99).

L'annotation a consisté à indiquer pour chaque phrase du texte les relations de subordination et de coordination explicite existant avec une phrase se trouvant en amont dans le texte ; ces deux types de relations pouvant exister pour une même phrase. Le principe de dépendance que nous avons suivi consiste à toujours resituer un énoncé vis-à-vis de la thématique globale puis d'analyser si localement il n'y a pas des dépendances plus fortes. Chaque couple d'énoncés que nous avons liés est décrit par une décision, D, concernant le type de relation qui les unit. Ces

[6]Cette boîte à outils est disponible à l'URL suivante *www.cs.waikato.ac.nz/ml/weka.*

couples sont ensuite représentés par l'ensemble des caractéristiques discursives, C, que nous avons précédemment définies. Sur un total de 1038 phrases[7], 1190 couples exemples, (C, D), ont été constitués. Ils se répartissent en 632 couples liées par une relation de subordination, 285 instances "coordination" et 273 instances décrivant une absence de relation. Les instances décrivant une absence de relation ont été engendrées automatiquement en considérant les couples d'énoncés contigus ne possédant pas de relation entre eux. En comparaison Choi utilise un corpus d'apprentissage de 754 exemples.

5.2 Résultats

De part la quantité de nos données d'apprentissage (relative au coût en temps d'annotation de corpus), nous adoptons une technique d'évaluation par validation croisée sur 10 partitions. Son principe consiste à partitionner le corpus d'apprentissage en un certain nombre de parts égales et d'utiliser tour à tour une partie comme ensemble d'exemples de test et les autres comme ensemble d'exemples d'entraînement. La moyenne des taux d'erreur correspond au *taux d'erreur global*.

		Expériences		
coordination_et_subordination **approche de base de 53,10%**		**Progression thématique**	**Expressions clefs**	**Progression thématique Et Expressions clefs**
Ensemble de base		52,68%	57,31%	56,13%
Caractéristique ajoutée	*cohésion lexicale*	52,68%	57,31%	**57,05%**
	antonymie	52,43%	56,89%	55,79%
	indices syntaxiques	**54,70%**	**58,57%**	55,71%
	# de mots communs	52,43%	57,31%	56,47%
	degré de parallélisme	52,43%	56,97%	55,12%
Toutes les caractéristiques		54,62%	57,05%	55,21%
seulement_la_subordination **approche de base de 69,83%**		**Progression thématique**	**Expressions clefs**	**Progression thématique Et Expressions clefs**
Ensemble de base		69,83%	73,14%	73,59%
Caractéristique ajoutée	*cohésion lexicale*	69,83%	72,26%	73,70%
	antonymie	69,83%	73,14%	73,70%
	indices syntaxiques	**72,48%**	**76,35%**	**75,02%**
	# de mots communs	69,83%	72,81%	74,03%
	degré de parallélisme	69,83%	73,59%	74,25%
Toutes les caractéristiques		70,16%	75,13%	**75,02%**

Table 1: Précisions de DST dans la prédiction de relation

Nous avons réalisé deux jeux d'expérience : le premier en considérant toutes les relations de notre modèle (subordination, coordination et absence de relation), le deuxième en ne considérant plus que la relation de subordination et l'absence de relation. Ce dernier jeu d'expériences nous permet de comparer nos résultats avec ceux de Choi (2002). Pour simplifier la présentation de ces jeux d'expériences par la suite, nous omettrons la relation "absence de relation" dans leur désignation. Pour chacun des jeux nous proposons de comparer les résultats sur deux ensembles d'indices de base distincts auxquels on ajoute tour à tour telle ou telle caractéristique pour observer les améliorations éventuelles. Les performances des deux ensembles combinés sont aussi considérées. Ces deux *ensembles de base* sont : 1) les caractéristiques décrivant la

[7]Les phrases ont été détectées à l'aide des caractères de ponctuation puis corrigées manuellement.

progression thématique (de thème en thème, de thème en rhème, de rhème en thème et de rhème en rhème) ; 2) les caractéristiques fondées sur les *expressions clefs* : les classes Initie, Termine, Continue, Subordonné et Subordonnant. Les caractéristiques individuelles que nous ajoutons sont : les liens lexicaux autre que d'antonymie (appelés par la suite "cohésion lexicale"), les liens lexicaux d'antonymie, les indices syntaxiques (*be* et négation), le degré de parallélisme et le nombre de mots communs (approche simplifiée de notre mesure du parallélisme).

Afin de positionner l'apport des différents apprentissages, nous comparons leur performance vis-à-vis d'une *approche de base* qui correspond à la prédiction de la classe majoritaire dans le corpus d'apprentissage (c'est-à-dire qu'elle correspond au taux d'erreur si l'on assigne tous les exemples à cette classe).

La table 1 décrit les résultats que nous obtenons respectivement lorsque l'on considère les relations de coordination et de subordination, puis lorsque l'on ne considère plus que la relation de subordination. Les valeurs en gras correspondent à des précisions maximales.

Le premier constat que nous faisons est que nous obtenons des résultats compris entre 60% et 75% similaires à ceux de Choi (2002) et de Marcu (1999). Plus particulièrement, nous obtenons *des résultats supérieurs à ceux obtenus par Choi* dans des configurations expérimentales similaires : 76,35% contre 73,61% pour nos meilleures performances de précision respectives.

Par rapport aux approches de base, les meilleurs sous-ensembles de caractéristiques augmentent la précision de plus de 5% pour chacun des jeux d'expériences. Il existe néanmoins des sous-ensembles qui détériorent les performances et les résultats sont en général moins bon pour le jeu *coordination_et_subordination*.

Les meilleurs résultats que nous obtenons sont à partir de l'ensemble de base composé de caractéristiques fondées sur les expressions clefs.

Les résultats obtenus avec les caractéristiques fondées sur des liens lexicaux quels qu'ils soient, combinées ou non, sont bien en dessous de ceux que l'on pouvait espérer. Pour le jeu *coordination_et_subordination* les expériences menées à partir de l'ensemble de base "progression thématique" déteriorent pour la plupart la précision de l'approche de base. Pour le jeu *seulement_la_subordination*, la précision des expériences à partir de l'ensemble de base "progression thématique" reste inchangée par rapport à l'approche de base. Le gain notable de l'ensemble "progression thématique" vient lorsqu'il est combiné à l'ensemble "expressions clefs".

Un *gain inattendu est celui apporté par le couple de présence du verbe être ou d'une négation.* Ce résultat requiert un retour au texte pour déterminer un phénomène discursif éventuel.

Enfin, lorsque l'on compare *les caractéristiques "nombre de mots pleins communs" et "degré de parallélisme" les différences sont légères mais mettent en avant le degré de parallélisme.*

6 Conclusion

Notre approche du discours enrichit le modèle de Choi (2002) qui ne considère que la relation de subordination. Nous considérons en plus la relation de coordination ce qui nous permet de modéliser plus finement le discours.

Le système de Marcu (1999) se situe à un degré supérieur de complexité dans le sens où il cherche à reconnaître l'opération de structuration à réaliser en fonction du contexte et de la

configuration structurelle en cours. Marcu fait des hypothèses très fortes sur le type de structure et d'attachements possibles. En comparaison, le fait de dissocier le modèle de dépendance de la structuration nous permet de fixer indépendamment les contraintes de structuration, et par là d'appréhender plus largement les différents phénomènes de structuration du discours (i.e. des structures autres que hiérarchiques orientées vers la frontière droite). Ce type de modélisation peut ainsi être utilisé pour analyser par exemple des dialogues.

En utilisant l'algorithme "*shift and reduce*", nous obtenons une structure hiérarchique proche de celle d'une structure décrite par une analyse RST (correspondance entre les plans informationnelles et intentionnelles). La différence majeure survient au niveau de la nucléarité des relations unissant les énoncés.

Parmi nos perspectives nous envisageons d'enrichir notre modèle avec la relation de subordination dirigée vers l'aval du texte, ainsi que de nouveaux indices (comme ceux de mis en forme visuelle) qu'ils se trouvent dans les énoncés considérés ou dans leur contexte.

Références

Nicholas Asher et Alex Lascarides. Intentions and information in discourse. 1994.

Regina Barzilay et Michael Elhadad. Using lexical chains for text summarization. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, Madrid, Spain, July 11 1997.

Eugene Charniak. Statistical parsing with a context-free grammar and word statistics. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence AAAI*, Menlo Park, 1997. MIT Press.

Freddy Y. Y. Choi. *Content-based Text Navigation.* PhD thesis, Department of Computer Science, University of Manchester, 2002.

Javier Couto, Olivier Ferret, Brigitte Grau, Nicolas Hernandez, Agata Jackiewicz, Jean-Luc Minel, et Sylvie Porhiel. RÉgal, un système pour la visualisation sélective de documents. *La présentation d'information sur mesure, Numéro Spécial de RIA*, pages 481–514, 2004.

Marti A. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, March 1997.

Nicolas Hernandez et Brigitte Grau. Automatic extraction of meta-descriptors for text description. In *RANLP*, Borovets, Bulgaria, 10-12 September 2003.

Nicolas Hernandez. Un indice de structuration de texte combinant finesse et disponibilité au niveau global et local. In *ATALA*, La Rochelle, France, 22 juin 2004.

Ivana Kruijff-Korbayová et Geert-Jan M. Kruijff. Identification of topic-focus chains. In S. Botley, J. Glass, T. McEnery, et A. Wilson, editors, *DAARC96*, volume 8, pages 165–179. July 17-18 1996.

William C. Mann et Sandra A. Thompson. Rhetorical structure theory: A theory of text organisation. Technical report isi/rs-87-190, Information Sciences Intitute, June 1987.

Daniel Marcu. A decision-based approach to rhetorical parsing. In *The 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 365–372, Maryland, June 1999.

M.-F. Moens et R. De Busser. Generic topic segmentation of document texts. In *ACM SIGIR*, pages 418–419, New York, 2001.

Livia Polanyi. A formal model of the structure of discourse. *Journal of Pragmatics*, 12:601–638, 1988.

J. Virbel. The contribution of linguistic knowledge to the interpretation of text structure. In J. André, V. Quint, et R. Furuta, editors, *Structured Documents*, pages 161–181. Cambridge University, 1989.

Paradocs : un système d'identification automatique de documents parallèles

Alexandre Patry et Philippe Langlais
Laboratoire de Recherche Appliquée en Linguistique Informatique
Département d'Informatique et de Recherche Opérationnelle
Université de Montréal
C.P. 6128, succursale Centre-ville
H3C 3J7, Montréal, Québec, Canada
{patryale,felipe}@iro.umontreal.ca

Mots-clefs : Corpus parallèles, apprentissage automatique, traduction automatique

Keywords: Parallel documents, machine learning, machine translation

Résumé Les corpus parallèles sont d'une importance capitale pour les applications multilingues de traitement automatique des langues. Malheureusement, leur rareté est le maillon faible de plusieurs applications d'intérêt. Extraire de tels corpus du Web est une solution viable, mais elle introduit une nouvelle problématique : il n'est pas toujours trivial d'identifier les documents parallèles parmi tous ceux qui ont été extraits. Dans cet article, nous nous intéressons à l'identification automatique des paires de documents parallèles contenues dans un corpus bilingue. Nous montrons que cette tâche peut être accomplie avec précision en utilisant un ensemble restreint d'invariants lexicaux. Nous évaluons également notre approche sur une tâche de traduction automatique et montrons qu'elle obtient des résultats supérieurs à un système de référence faisant usage d'un lexique bilingue.

Abstract Parallel corpora are playing a crucial role in multilingual natural language processing. Unfortunately, the availability of such a resource is the bottleneck in most applications of interest. Mining the web for such a resource is a viable solution that comes at a price : it is not always easy to identify parallel documents among the crawled material. In this study we address the problem of automatically identifying the pairs of texts that are translation of each other in a set of documents. We show that it is possible to automatically build particularly efficient content-based methods that make use of very little lexical knowledge. We also evaluate our approach toward a front-end translation task and demonstrate that our parallel text classifier yields better performances than another approach based on a rich lexicon.

1 Introduction

De nos jours, les corpus de *documents parallèles* (ensemble de documents exprimant le même contenu dans le même ordre) jouent un rôle crucial dans les applications multilingues de traitement automatique des langues (Véronis, 2000). Aligné au niveau des phrases, une tâche pouvant être accomplie avec fiabilité (Langlais *et al.*, 1998), un corpus parallèle s'avère très utile aux concordanciers bilingues (Macklovitch *et al.*, 2000) et est la pierre angulaire de la plupart des systèmes commerciaux de mémoire de traduction. Aligné au niveau des mots, une tâche maintenant bien maîtrisée (Brown *et al.*, 1993), un corpus parallèle peut servir à plusieurs applications telles que la traduction automatique, la désambiguïsation de mots ou l'extraction d'information translinguistique.

Malheureusement, il existe assez peu de *corpus parallèles* (ensemble de documents parallèles) riches et bien organisés comme le sont par exemple les *Hansards* canadiens (anglais/français), les débats parlementaires de Hongkong (anglais/chinois), les transcriptions des débats du parlement européen[1] (EUROPARL, disponibles en onze langues) ou encore les transcriptions des débats parlementaires du Nunavut (anglais/inuktitut)[2].

S'il existe également des ressources telles que la Bible qui sont traduites dans de nombreuses langues (mais pas nécessairement organisées en corpus parallèle), il n'en reste pas moins que la rareté des corpus parallèles demeure le goulot d'étranglement pour plusieurs applications d'intérêt. Plusieurs solutions ont été proposées pour palier leur absence. Il est par exemple possible d'extraire automatiquement des corpus parallèles à partir du Web (Ma & Liberman, 1999; Kraaij *et al.*, 2003; Resnik & Smith, 2003). Il est également possible de tirer profit de *corpus comparables* (corpus traitant du même sujet sans nécessairement être parallèles) (Munteanu *et al.*, 2004), voire même d'utiliser des corpus n'ayant aucune affinité (Rapp, 1999). D'autres misent à plus long terme sur des outils informatiques simplifiant la gestion des données parallèles (Hajlaoui & Boitet, 2004).

Dans cet article, nous nous intéressons à la détection des documents parallèles dans un corpus bilingue (par exemple extrait d'un site Web) à l'aide d'invariants lexicaux (par exemple données chiffrées, entités nommées, ponctuations). Cette idée était à la base d'un algorithme d'alignement bilingue de phrases décrit par Simard *et al.* (1993) ; nous montrons ici qu'elle s'applique à notre problème.

Nous décrivons en section 2 notre méthodologie et présentons les différentes métriques utilisées. Nous montrons en section 3 que notre approche permet d'identifier sans faute les paires parallèles d'une partie du corpus EUROPARL. Nous évaluons également notre approche à travers une tâche de traduction automatique et mesurons des performances supérieures à celles d'une approche faisant usage d'un lexique bilingue riche (section 4). Nous discutons en section 5 de travaux connexes et présentons en section 6 nos conclusions.

2 Méthodologie

Nous considérons dans cette étude que nous disposons de deux ensembles de documents: un ensemble $\mathcal{S}$ contenant les documents d'une langue source et un ensemble $\mathcal{T}$ contenant ceux

[1] `http://www.europarl.eu.int/home/default_fr.htm`
[2] `http://www.inuktitutcomputing.ca/NunavutHansards/`

d'une langue cible. Ces documents peuvent par exemple provenir du Web (Kraaij *et al.*, 2003) et leur langue peut avoir été identifiée automatiquement, comme ce sera le cas dans les expériences de la section 4.

Le problème que nous résolvons consiste à déterminer le sous-ensemble du produit Cartésien $\mathcal{S} \times \mathcal{T}$ qui contient les paires de documents parallèles. Nous ne faisons pas usage dans cette étude d'informations externes aux documents comme leur nom ou leurs balises structurelles, ce qui exclut l'usage d'heuristiques basées sur les noms de fichiers comme celles décrites dans (Resnik & Smith, 2003). Cette contrainte ne découle pas d'une pensée puriste, mais correspond à notre volonté d'évaluer objectivement différentes métriques n'utilisant que le contenu des documents. Ces caractéristiques externes pourraient cependant être incorporées facilement à notre approche.

L'identification des paires de documents parallèles est réalisée en deux étapes: le pointage de toutes les paires du produit Cartésien $\mathcal{S} \times \mathcal{T}$ et la classification de chacune d'elles comme parallèle ou non. Les différents pointages utilisés sont décrits dans la section 2.1 et l'algorithme de classification dans la section 2.2.

2.1 Métriques

Trois différentes familles de métriques sont utilisées pour mesurer le parallélisme de deux documents. La mesure de cosinus et la distance d'édition normalisée utilisent certaines des unités lexicales des documents: les séquences de chiffres (NOMBRE), certaines ponctuations (PUNCT) et les entités nommées (ENTITÉ). Les ponctuations que nous avons considérées sont les parenthèses, les crochets et les guillemets. De plus, nous avons considéré comme une entité nommée tout mot commençant par une lettre majuscule mais ne débutant pas une phrase. Ces types d'unités lexicales sont relativement indépendants des langues considérées. La troisième famille de métriques utilise la sortie d'un aligneur de textes au niveau des phrases pour juger du parallélisme de deux documents.

Mesure de cosinus (COS) Nous avons repris l'idée proposée par Nadeau et Foster (2004) et représenté un document par différents vecteurs où chaque dimension correspond à une unité lexicale et chaque coordonnée à la fréquence de cette unité dans le document. Dans nos expériences, chaque document est représenté par trois vecteurs: un pour les nombres, un pour les ponctuations et un pour les entités nommées. Un exemple d'une telle représentation est présenté en figure 1. La similarité entre deux documents est mesurée par la *mesure de cosinus* entre leur représentation vectorielle, mesure populaire en extraction d'information.

Distance d'édition normalisée (EDIT) La représentation vectorielle ne tient pas compte de l'ordre des unités lexicales dans le document, information qui peut être pertinente ici. Nous proposons de représenter un document par trois séquences d'unités lexicales (NOMBRE, PUNCT, ENTITÉ). Le parallélisme de deux documents peut ainsi être mesuré en comparant leurs séquences (voir la figure 1).

Pour mesurer la similarité de deux séquences, nous utilisons la distance d'édition (Levenshtein, 1966), qui compte le nombre minimal d'opérations nécessaires pour transformer la première séquence en la seconde (les opérations permises sont l'insertion, la suppression ou la substitution

d'une unité lexicale). Nous la normalisons ensuite par la longueur de la plus longue des deux séquences.

> Approximately **60**% very roughly, **60**% to **40**%, when the **60**% is paid by the tenant and **40**% is approximately paid by the Government subsidy.
>
> apiqqutiqaqqaujunga akunialuk, angiqqaugaluarakku $**60** milian kaivainnaqtuq kiinaujaqarvingmut, kisianittauq tusaqtitauvalliaqqaugama, takuvallialiqtugu $**39** milian **807** tausan ammalu taanna angiqtauguni taikkuali amiakkujut $**60** milianut tikillugu kisumut atuqtaugajaqpat ?

FIG. 1 – Si nous devions comparer les nombres dans les deux documents ci-haut (extraits anglais et inuktitut tirés des débats parlementaires du Nunavut), les représentations vectorielles utilisées pour la mesure de cosinus seraient $(0_{39}, 2_{40}, 3_{60}, 0_{807})$ et $(1_{39}, 0_{40}, 2_{60}, 1_{807})$. Alors que les représentations séquentielles pour mesurer la distance d'édition normalisée seraient $<60, 60, 40, 60, 40>$ et $<60, 39, 807, 60>$.

Scores d'alignements Une autre source d'information permettant de mesurer le parallélisme de deux documents est la sortie d'un aligneur de textes au niveau des phrases. Nous avons utilisé l'aligneur JAPA (Langlais *et al.*, 1998) qui produit une séquence d'alignements et un score global mesurant le *coût* de l'alignement produit. Les alignements qu'il produit sont de type m-n ($m, n \in \{0, 1, 2\}$) où m et n sont respectivement le nombre de phrases sources et le nombre de phrases cibles impliquées dans l'alignement.

Nous retenons cinq pointages: le ratio d'alignements *1-0* ou *0-1*, le ratio d'alignements *1-1*, le ratio d'alignements *1-2* ou *2-1*, le ratio d'alignements *2-2* et le score global d'alignement. Nous nommerons dorénavant les quatre ratios M-N et le score global COÛT. Intuitivement, le résultat de l'aligneur sur deux documents parallèles devrait contenir plusieurs alignements de type *1-1* et devrait être de faible coût.

2.2 Identification des paires parallèles

Chaque paire de documents est décrite par un ensemble de pointages. Une première approche pour identifier celles qui sont parallèles consiste à ajuster manuellement des seuils sur ces pointages, une tâche délicate ne se généralisant pas nécessairement bien. Nous avons plutôt utilisé AdaBoost (Y.Freund & Schapire, 1999), un algorithme d'apprentissage. Cet algorithme prend en entrée un ensemble de paires de documents, leurs pointages et leur *étiquette* (parallèle ou non) et produit à partir de cet ensemble d'entraînement une fonction classant une paire comme parallèle ou non à partir de ses pointages.

AdaBoost est un algorithme d'apprentissage itératif combinant plusieurs *classificateurs faibles* (classificateur juste plus d'une fois sur deux) en un classificateur plus robuste. À chaque itération, un classificateur faible est entraîné à reconnaître l'étiquette de toutes les paires de documents (à partir des pointages) en accordant plus d'importance à celles qui ont été moins bien étiquetées par les classificateurs faibles précédents. Les itérations se poursuivent jusqu'à ce qu'un classificateur faible ait un ratio d'erreur supérieur ou égal à 50% ou jusqu'à ce qu'un nombre maximal (fixé à l'avance) d'itérations ait été atteint. Le classificateur retourné par AdaBoost fait voter les différents classificateurs faibles afin de déterminer si une paire est parallèle

ou non.

Dans nos expériences, nos classificateurs faibles étaient des réseaux neuronaux (Bishop, 1996) à une couche cachée de cinq neurones et après quelques expériences informelles, nous avons décidé de borner le nombre d'itérations d'AdaBoost à 75. L'entraînement et les tests ont été réalisés à l'aide du logiciel PLEARN [3].

3 Expérience contrôlée

EUROPARL est un corpus parallèle tiré de la transcription des débats parlementaires européens s'étant tenus entre avril 1996 et septembre 2003 (Koehn, 2002). Les débats parlementaires européens sont traduits en onze langues, mais nous nous sommes concentrés sur les traductions anglaises et espagnoles. Notre corpus était composé de 487 textes anglais et de 487 textes espagnols ayant en moyenne environ 2800 phrases chacun.

3.1 Protocole d'évaluation

Parce que les paires de documents parallèles sont bien identifiées dans EUROPARL, les différentes configurations ont été comparées sur la base de leur *précision*, de leur *rappel* et de leur *f-mesure* (moyenne harmonique de la précision et du rappel). La précision (resp. rappel) est le ratio du nombre de paires vraiment parallèles que le classificateur a identifiées sur le nombre total de paires que le classificateur a identifiées (resp. sur le nombre total de paires parallèles dans le corpus). La précision indique la qualité de l'ensemble des paires trouvées et le rappel sa couverture.

Les différentes configurations ont été évaluées à l'aide d'une *validation croisée en cinq étapes*. Le produit cartésien $\mathcal{S} \times \mathcal{T}$ a été partitionné aléatoirement en cinq sous-ensembles de même taille. Ensuite, cinq expériences ont été lancées en testant chaque fois sur un sous-ensemble différent et en entraînant avec les paires ne faisant pas partie de ce sous-ensemble de test.

3.2 Système de référence (LEXIQUE)

Pour mettre en contexte les performances de nos différents classificateurs, un système de référence utilisant un *lexique bilingue* a été mis au point. Le lexique bilingue qui a été utilisé contient plus de 70 000 entrées et provient du projet PYTHOÑOL[4], qui vise à aider les locuteurs anglais à apprendre l'espagnol.

Un document est représenté par l'ensemble de ses *mots rares* (dans le cadre de ce projet, les mots rares sont ceux n'apparaissant qu'une seule fois dans le document) présents dans le lexique bilingue. Chaque document source est ensuite apparié avec le document cible partageant avec lui le plus grand nombre de mots rares.

[3] `http://plearn.sourceforge.net`
[4] `http://sourceforge.net/projects/pythonol/`

Configuration							Performances (%)		
COS	EDIT	NOMBRE	PUNCT	ENTITÉ	COÛT	M-N	précision	rappel	f-mesure
	√	√	√	√			100	100	100
√	√	√	√	√	√	√	99.8	99.8	99.8
	√	√					98.3	99.8	99.0
	√	√			√		96.6	99.8	98.1
					√		85.8	99.8	92.1
						√	65.6	99.4	77.1
					√	√	49.3	99.4	62.7
√		√	√	√			24.6	99.2	38.7
√		√					12.4	98.9	21.8

TAB. 1 – Précision, rappel et f-mesure de différentes configurations d'entraînement du classificateur. Notez que valeurs rapportées sont des moyennes sur les cinq étapes de la validation croisée.

3.3 Résultats

Nous avons entraîné des classificateurs sur plusieurs combinaisons des pointages décrits dans la section 2.1. Leurs performances sont présentées dans la Table 1. La meilleure de nos configurations et le système de référence ont tous deux obtenus des résultats parfaits.

Les meilleurs performances des métriques basées sur la distance d'édition semblent confirmer l'hypothèse selon laquelle l'ordre des unités lexicales est importante pour l'identification des documents parallèles. Il est à noter que le seul usage de la distance d'édition sur les nombres amène une f-mesure de 99%, ce qui suggère que les nombres sont de très bons indices de parallélisme pour ce genre de corpus. En effet, les débats parlementaires contiennent plusieurs nombres stables comme des dates, des numéros de lois ou encore les comptes de votes.

On observe également que les configurations utilisant les pointages d'alignements n'amènent pas de bons résultats. L'usage des ratios de types d'alignements donne en particulier une f-mesure moyenne inférieure d'au moins 20% aux meilleures configurations et ont été instables dans les différentes étapes de la validation croisée.

4 Tâche réelle

Nous avons montré dans la section précédente qu'il était possible d'identifier parfaitement les paires parallèles d'un corpus bilingue comme EUROPARL. Nous voulons maintenant mesurer si des performances satisfaisantes peuvent être obtenues dans un contexte d'utilisation plus représentatif. Nous avons pour cela aspiré le site Web de la *Pan American Health Organization*[5]. Bien qu'en principe simple, cette tâche s'est avérée particulièrement délicate (nombreux formats propriétaires, absence d'une nomenclature pour nommer et identifier les différentes ressources bilingues).

Le corpus résultant, PAHO, totalise 6878 documents dont 2523 ont été identifiés comme étant anglais (et 4355 comme espagnols) par SILC[6], l'outil que nous avons utilisé pour identifier la

[5] `http://www.paho.org`.
[6] `http://rali.iro.umontreal.ca`.

langue de chaque document. Au total, ce corpus compte plus de 10 millions de paires potentielles. Chaque document contient en moyenne environ 180 phrases. Une inspection informelle du corpus a révélé que plusieurs de ces documents sont identiques ou très similaires et que certains sont bilingues.

4.1 Protocole d'évaluation

Pour cette expérience, nous avons mesuré l'impact de nos différents extracteurs de paires parallèles sur une tâche de traduction automatique (TA) de l'espagnol vers l'anglais. Deux raisons majeures ont mené à ce choix. Premièrement, l'identification de documents parallèles n'a d'intérêt que dans un cadre applicatif donné ; la traduction étant l'application bilingue par excellence. Deuxièmement, nous ne connaissons pas les documents parallèles du corpus PAHO ce qui complique les calculs de précision et de rappel.

Le moteur de traduction que nous utilisons ici est un moteur probabiliste état de l'art (Koehn *et al.*, 2003). L'avantage d'un tel choix réside dans le fait que l'obtention d'un tel système est entièrement automatique une fois un corpus parallèle identifié.

Afin d'évaluer les traductions produites, nous avons téléchargé 520 nouvelles phrases du site de la *Pan American Health Organization* avec leur traduction. Pour les mêmes raisons d'automaticité, nous mesurons la qualité de nos traductions à l'aide de quatre métriques couramment utilisées en TA: deux taux d'erreurs au niveau des phrases (SER) et des mots (WER) et deux mesures de précision n-grammes (BLEU et NIST) calculées par le script `mteval`[7].

Les deux premières métriques varient entre 0 et 100 où 0 représente une traduction parfaite. SER (pour *Sentence-Error-Rate*) est le ratio des phrases produites par le moteur de TA qui sont différentes de la référence. WER (pour *Word-Error-Rate*) calcule la distance d'édition normalisée entre les mots de la traduction produite et ceux de la traduction de référence. BLEU et NIST comptent le nombre de séquences partagées entre la traduction automatique et la traduction de référence en donnant plus d'importance aux séquences plus longues. Le score BLEU varie entre 0 et 1 (où 1 est le score de la référence) alors que le score NIST n'est pas normalisé[8].

En plus de l'évaluation à l'aide d'un moteur de TA, la précision (voir la section 3.1) de chaque configuration a été calculée manuellement.

4.2 Résultats

Nous avons comparé les performances de notre moteur de TA lorsqu'il est entraîné sur quatre corpus parallèles différents. Le corpus COS-TOUS a été généré à l'aide de la mesure de cosinus sur les nombres, sur les ponctuations et sur les entités nommées (ligne 8 de la Table 1). Le corpus EDIT-TOUS a été obtenu à l'aide de la configuration ayant obtenue les meilleurs résultats sur EUROPARL, la distance d'édition normalisée sur les nombres, sur les ponctuations et sur les entités nommées (ligne 1 de la Table 1). Le corpus LEXIQUE a été produit à l'aide du système de référence (basé sur l'utilisation d'un lexique bilingue). Finalement, le corpus COS-TOUS ∪ EDIT-TOUS est l'union de COS-TOUS et de EDIT-TOUS. Une inspection de ces deux corpus nous a en effet révélé qu'ils ne partagent que 229 paires de documents.

[7] Disponible à l'adresse `http://www.nist.gov/speech/tests/mt/mt2001/resource`.

[8] Son calcul sur la référence produit dans notre cas une valeur de 13.11.

Corpus parallèle	N	SER	WER	NIST	BLEU	précision
COS-TOUS ∪ EDIT-TOUS	494	99.42	60.02	5.3125	0.2435	99.0
LEXIQUE	529	99.42	61.67	5.1989	0.2304	89.2
EDIT-TOUS	390	99.42	61.53	5.1342	0.2290	99.0
COS-TOUS	333	99.23	62.23	5.1629	0.2256	99.7

TAB. 2 – Performances de notre moteur de TA lorsque entraîné sur les corpus parallèles retournées par différentes configurations où N est le nombre de paires identifiées comme étant parallèles.

Les classificateurs identifiant les paires parallèles ont été entraînés sur les paires du corpus EUROPARL. Pour chaque configuration, les paires partageant un document ont été rejetées (ce sont les paires les plus incertaines). Les performances de traduction des moteurs probabilistes correspondant sont présentées en Table 2.

Contrairement à nos expériences sur EUROPARL, la mesure de cosinus et la distance d'édition ont des performances comparables. Cela pourrait s'expliquer par la plus petite taille des documents de PAHO (rendant l'ordre des caractéristiques moins important) et par l'étape de suppression des paires partageant un document. Nous observons que les performances du moteur entraîné sur le corpus COS-TOUS ∪ EDIT-TOUS sont meilleures que celles du moteur entraîné sur le corpus LEXIQUE, et ce même si ce dernier contient plus de paires. Ce résultat est particulièrement intéressant puisqu'il montre qu'il n'est pas nécessaire de réunir un lexique bilingue. Un autre résultat encourageant est la forte précision de tous nos classificateurs (99% ou plus).

5 Travaux connexes

Ce travail a été inspiré de celui de Nadeau et Foster (2004). Les auteurs ont proposé l'utilisation de la mesure de cosinus sur les nombres, les ponctuations, les entités nommées et le nombre de paragraphes pour détecter les paires de documents parallèles d'un corpus bilingue de communiqués. Ils ont montré qu'à l'aide d'un filtre sur les dates de publication, ils pouvaient identifier les documents parallèles du *Groupe Canada NewsWire*[9] avec une grande précision.

Nous avons étendu cette idée de trois façons. Premièrement nous avons validé l'utilisation d'unités lexicales invariantes sur des corpus de natures différentes. Deuxièmement, nous avons montré que l'ordre de ces caractéristiques est porteur d'information. Finalement, nous avons testé l'impact de cette approche sur une tâche concrète: la traduction automatique.

Notre travail, même si mené de façon indépendante, partage des points communs avec celui de Munteanu *et al.* (2004). Les auteurs ont montré qu'un moteur de traduction pouvait bénéficier d'un corpus parallèle extrait automatiquement de corpus comparables. L'approche qu'ils ont proposée est analogue à la nôtre: ils entraînent un classificateur (dans leur cas par une approche de *maximum entropie*) pour identifier les paires de phrases en relation de traduction (alors que nous travaillons au niveau du document). Ils font cependant l'hypothèse qu'un corpus parallèle est disponible afin d'entraîner un modèle de traduction qu'ils utiliseront ensuite pour aligner les phrases au niveau des mots. Nous pensons que cette approche est complémentaire à la nôtre. Notre approche serait plus adaptée pour les corpus où nous savons à priori qu'ils contiennent

[9] http://www.newswire.ca

plusieurs documents parallèles.

6 Conclusions et travaux futurs

Nous avons présenté une approche complètement automatique permettant d'identifier les paires de documents parallèles d'un corpus bilingue et ce, à l'aide d'un nombre restreint d'informations lexicales. Nous avons plus précisément étudié l'usage de certains invariants lexicaux comme les nombres, certaines ponctuations et les entités nommées. Nous avons montré que cette approche amenait des résultats comparables (voire supérieurs) à une approche de référence faisant usage d'un lexique bilingue riche.

L'un des avantages majeurs de notre approche est sa souplesse que nous devons à l'utilisation d'un algorithme d'apprentissage à la fois simple à mettre en place et efficace. Il est donc tout à fait possible d'étendre la liste des traits (pointages) que nous avons utilisés pour représenter nos documents. Ajouter comme trait le nombre d'entrées d'un lexique bilingue que partagent deux documents serait par exemple particulièrement aisé.

Nous travaillons actuellement sur l'amélioration de deux limitations du système proposé. Premièrement, nous avons considéré systématiquement dans cette étude toutes les paires du produit cartésien entre l'ensemble des documents sources et cibles. Cela impose des temps de traitement qui peuvent vite devenir prohibitifs. Certaines heuristiques conservatrices peuvent être appliquées pour limiter l'espace de recherche des paires de documents parallèles. Nous pouvons par exemple éliminer les paires de documents dont le rapport de longueur est anormalement grand ou faible (Kraaij *et al.*, 2003).

Deuxièmement, nous aimerions vérifier l'efficacité de l'approche si seulement une partie de chaque document est inspectée (par exemple les premières phrases). Cela diminuerait le temps de calcul des pointages et par le fait même accélérerait le processus au complet.

Remerciements

Nous voudrions remercier Leila Arras et Marie Ouimet pour avoir mis à notre disposition le corpus PAHO.

Références

BISHOP C. M. (1996). *Neural networks for pattern recognition.* Oxford University Press.

BROWN P. F., PIETRA S. A. D., PIETRA V. J. D. & MERCER R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, **19**(2), 263–311.

HAJLAOUI N. & BOITET C. (2004). PolyphraZ : a tool for the quantitative and subjective evaluation of parallel corpora. In *Proc. of the International Workshop on Spoken LanguageTranslation*, p. 123–129, Kyoto, Japan.

KOEHN P. (2002). Europarl: A multilingual corpus for evaluation of machine translation. Draft.

KOEHN P., OCH F. J. & MARCU D. (2003). Statistical phrase-based translation. In *Proceedings of the Second Conference on Human Language Technology Reasearch (HLT)*, p. 127–133, Edmonton, Alberta,

Canada.

KRAAIJ W., NIE J.-Y. & SIMARD M. (2003). Embedding web-based statistical translation models in cross-language information retrieval. *Computational Linguistics*, **29**(3), 381–419.

LANGLAIS P., SIMARD M. & VERONIS J. (1998). Methods and practical issues in evaluating alignment techniques. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 711–717, Montréal, Quebec, Canada.

LEVENSHTEIN V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Sov. Phys. Dokl.*, **6**, 707–710.

MA X. & LIBERMAN M. (1999). Bits: A method for bilingual text search over the web. In *Machine Translation Summit VII*, Kent Ridge Digital Labs, National University of Singapore.

MACKLOVITCH E., SIMARD M. & LANGLAIS P. (2000). Transsearch: A free translation memory on the world wide web. In *Second International Conference On Language Resources and Evaluation (LREC)*, volume 3, p. 1201–1208, Athens Greece.

MUNTEANU D. S., FRASER A. & MARCU D. (2004). Improved machine translation performance via parallel sentence extraction from comparable corpora. In *HLT-NAACL*, p. 265–272.

NADEAU D. & FOSTER G. (2004). Real-time identification of parallel texts from bilingual news feed. In *CLINE 2004*, p. 21–36: Computational Linguistics in the North East.

RAPP R. (1999). Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th conference on Association for Computational Linguistics*, p. 519–526: Association for Computational Linguistics.

RESNIK P. & SMITH N. A. (2003). The web as a parallel corpus. *Computational Linguistics*, **29**, 349–380. Special Issue on the Web as a Corpus.

SIMARD M., FOSTER G. F. & ISABELLE P. (1993). Using cognates to align sentences in bilingual corpora. In *CASCON '93: Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research*, p. 1071–1082: IBM Press.

J. VÉRONIS, Ed. (2000). *Parallel Text Processing, Alignment and Use of Translation Corpora*. Kluwer Academic.

Y.FREUND & SCHAPIRE R. (1999). A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, **14**(5), 771–780. Appearing in Japanese, translation by Naoki Abe.

Traduction automatique statistique avec des segments discontinus

Michel Simard∗, Nicola Cancedda∗, Bruno Cavestro∗,
Marc Dymetman∗, Eric Gaussier∗, Cyril Goutte∗,
Philippe Langlais†, Arne Mauser‡, Kenji Yamada⋆

(∗) Xerox Research Centre Europe (XRCE)
prenom.nom@xrce.xerox.com

(†) Laboratoire RALI, Université de Montréal
felipe@iro.umontreal.ca

(‡) Lehrstuhl für Informatik VI, RWTH Aachen
arne.mauser@kullen.rwth-aachen.de

(⋆) USC Information Science Institute
kyamada@isi.edu

Mots-clefs : traduction automatique statistique, segments discontinus, modèles log-linéaires

Keywords: statistical machine translation, discontinuous phrases, log-linear models

Résumé Cet article présente une méthode de traduction automatique statistique basée sur des segments non-continus, c'est-à-dire des segments formés de mots qui ne se présentent pas nécéssairement de façon contiguë dans le texte. On propose une méthode pour produire de tels segments à partir de corpus alignés au niveau des mots. On présente également un modèle de traduction statistique capable de tenir compte de tels segments, de même qu'une méthode d'apprentissage des paramètres du modèle visant à maximiser l'exactitude des traductions produites, telle que mesurée avec la métrique NIST. Les traductions optimales sont produites par le biais d'une recherche en faisceau. On présente finalement des résultats expérimentaux, qui démontrent comment la méthode proposée permet une meilleure généralisation à partir des données d'entraînement.

Abstract This paper presents a phrase-based statistical machine translation method, based on non-contiguous phrases, i.e. phrases with *gaps*. A method for producing such phrases from a word-aligned corpora is proposed. A statistical translation model is also presented that deals with such phrases, as well as a training method based on the maximization of translation accuracy, as measured with the NIST evaluation metric. Translations are produced by means of a beam-search decoder. Experimental results are presented, that demonstrate how the proposed method allows to better generalize from the training data.

1 Introduction

L'évolution des modèles et des méthodes et la prolifération des corpus parallèles ont, depuis peu, poussé les approches statistiques à l'avant-plan de la recherche en traduction automatique. Bien qu'on retrouve toujours au coeur de ces approches le cadre général qui a motivé les propositions initiales de l'équipe IBM (Brown et al.1993), on a pu observer des transformations importantes au cours des dernières années. La plus remarquable est sans doute le passage du niveau des mots à celui de *segments* de longueur variable[1] (Och et al.1999; Marcu and Wong2002; Tillmann and Xia2003). Alors que les modèles traditionnels prenaient pour unité de base le mot, les modèles "segmentaires" reconnaissent le rôle primordial que jouent dans la langue les expressions combinant plusieurs mots, et l'importance de les traduire en bloc. C'est bien sûr le cas des multitermes, qu'on rencontre plus fréquemment dans les domaines techniques et spécialisés, mais aussi des expressions idiomatiques, des locutions, et de tout un ensemble de phénomènes de la langue générale.

Mais le succès des approches segmentaires ne s'explique pas uniquement par l'importance et la fréquence de ces phénomènes linguistiques. En fait, l'utilisation de segments de plus d'un mot améliore la qualité des traductions, même lorsque ces segments n'ont pas de réel statut linguistique. Face à la rareté des événements sur lesquels se fonde l'estimation des nombreux paramètres d'un modèle de traduction, le concepteur se retrouve souvent devant un choix difficile, entre des estimations peu fiables et un lissage plus ou moins arbitraire. À défaut de résoudre ce dilemme, l'emploi d'unités plus longues représente l'application d'un principe intuitif : lorsqu'on a vu un long segment de texte en langue-source souvent traduit d'une certaine façon, il y a tout lieu de croire que cette traduction est préférable à toute autre qu'on pourrait obtenir de façon compositionnelle. En somme, les modèles segmentaires incorporent dans un cadre statistique l'intuition derrière la traduction automatique basée sur les exemples et, à la limite, les mémoires de traduction. Finalement, les segments de plusieurs mots contribuent à résoudre le problème du choix lexical face aux ambiguïtés de la langue-source. Alors que le mot anglais *bank* se traduit presque systématiquement par *banque* en français, il suffit d'avoir observé que *river bank* a été traduit par *rive*, ne fût-ce que quelques fois, pour produire la bonne traduction.

Les modèles segmentaires existants ne traitent que des segments constitués de mots contigus. Nous proposons ici un modèle capable de gérer des *segments discontinus*, c'est-à-dire des expressions formés de mots qui ne sont pas nécéssairement contigus, tant dans la langue-source que dans la langue-cible. La suite de cet article est ainsi structuré : en section 2, nous discutons des motivations pour traiter les segments discontinus, et présentons une méthode pour obtenir de telles unités, à partir d'un corpus d'entraînement; le modèle de traduction log-linéaire conditionnel que nous avons adopté fait l'objet de la section 3; nous décrivons brièvement le décodeur à la section 4; enfin, nous présentons en section 5 les résultats d'expériences que nous avons menées dans le but d'évaluer le potentiel de notre approche.

2 Les segments discontinus

Notre objectif, avec des segments constitués de mots non-contigus est d'améliorer la qualité des traductions produites, d'abord en élargissant la portée des effets mentionnés plus haut de

[1] On utilise couramment le terme *phrase* en anglais, de façon un peu abusive, faut-il souligner.

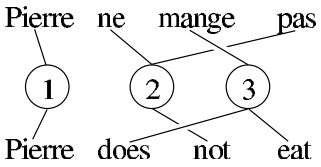


Figure 1: Alignement d'une négation, entre le français et l'anglais.

désambiguïsation lexicale et de traduction basée sur les exemples, mais aussi en prenant compte de nouveaux phénomènes linguistiques. Les verbes à particules, en anglais, constituent un exemple d'un tel phénomène. Dans une phrase comme "Mary *switches* her bedside lamp *off*" ("Marie éteint sa lampe de chevet") les modèles de traductions basés sur les mots sont généralement incapables de rendre compte de l'effet combiné de *switch* et de *off*. Alors qu'ils traitent correctement les locutions inséparables comme *to run out*, les modèles segmentaires existants sont tout aussi impuissants dans ce cas. Notons que ce phénomène ne se limite pas à l'anglais, puisqu'on l'observe également en allemand et dans bien d'autres langues.

Les unités linguistiques non-contiguës ne se limitent pas aux seuls verbes : la négation se forme de façon différente en français et en anglais, et les modèles existants sont incapables de représenter correctement l'alignement de mots complexe qui en résulte (figure 1). D'une façon générale, un modèle autorisant des relations de type *plusieurs-à-plusieurs* permet de rendre compte du fait qu'un même concept peut se voir réalisé par des unités de granularité différente dans différentes langues, sans égard pour la contiguïté.

Au sein d'une bi-phrase, nous appelons *bi-segment* une paire constituée d'un *segment-source* et d'un *segment-cible* : $b = \langle \tilde{f}, \tilde{e} \rangle$. Le segment-source est une suite de mots de la langue-source et de *jokers* (représentés par le symbole $\diamond$); on définit le segment-cible de manière analogue. Par exemple, $\tilde{f} = f_1 \diamond\diamond f_2 f_3$ est un segment-source de longueur 5, constitué d'un mot source, suivi de deux jokers, puis de deux mots-source contigus.

Avec de tels bi-segments, la traduction d'une phrase en langue-source f est produite en combinant les bi-segments $b = \langle \tilde{f}, \tilde{e} \rangle$ d'un ensemble choisi de façon d'une part à recouvrir entièrement la phrase f, et d'autre part à produire une phrase e bien formée dans la langue-cible. La production d'une traduction complète peut être décrite par une suite ordonnée de bi-segments $b_1 ... b_K$: on dépose d'abord le segment-cible $\tilde{e}_1$ du bi-segment b_1, puis chacun des segments subséquents $\tilde{e}_k$ sur la première position "libre", c'est-à-dire soit le joker le plus à gauche, soit l'extrémité droite de la séquence (figure 2) .

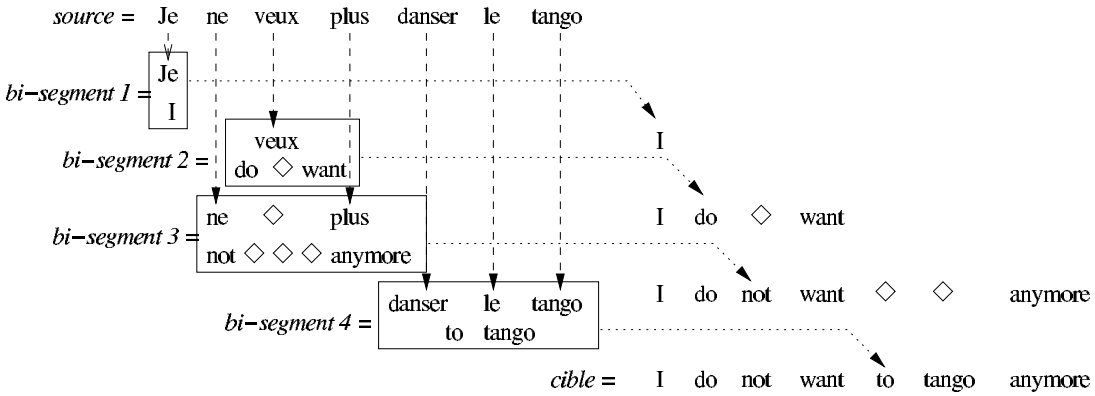


Figure 2: Production d'une traduction par combinaison de bi-segments.

Notre approche nécessite une *banque de bi-segments*, contenant les "briques" qui seront utilisées pour construire les traductions. La constitution d'une telle banque s'effectue en deux étapes : on aligne d'abord les mots d'un corpus bilingue, de façon à obtenir des bi-segments de base; on combine ensuite ces bi-segments, de manière à obtenir des briques de taille et de complexité croissante.

La première étape repose sur l'utilisation de la méthode d'alignement de mots proposée par (Goutte et al.2004). Cette méthode produit des alignements de type *plusieurs-à-plusieurs* entre les mots de la source et de la cible, par le biais d'une partition parallèle des deux textes, vus comme des ensembles de mots. Chaque mot appartient ainsi à un et un seul sous-ensemble dans cette partition, les sous-ensembles correspondants dans la source et la cible constituent ce qu'on appelle des *cepts*, et l'ensemble de ces cepts constitue l'alignement. Chaque cept réunit donc des mots de la source et de la cible, sans aucune contrainte de contiguïté. Dans la figure 1, ces cepts sont représentés par les cercles numérotés 1, 2 et 3.

L'ensemble des cepts observés dans un corpus bilingue constitue naturellement une banque de bi-segments élémentaires, que nous appelons L_1. Partant de là, on peut construire des banques de segments complexes : en combinant deux-à-deux les cepts provenant d'une même paire de phrases, on génère l'ensemble que nous appelons L_2. Par exemple, dans la figure 1, en combinant les cepts 1 et 2, on obtient le bi-segment $\langle$*Pierre ne* $\diamond$ *pas*, *Pierre* $\diamond$ *not*$\rangle$. Les combinaisons de 3 cepts produisent l'ensemble L_3, et ainsi de suite. La taille de ces ensembles croît théoriquement de façon exponentielle avec le nombre de cepts combinés. Comme nous le verrons plus loin, le nombre de bi-segments disponibles affecte directement le temps requis pour produire une nouvelle traduction. C'est pourquoi on aura recours à différentes méthodes de filtrage, visant à ne conserver que les bi-segments les plus susceptibles d'être utiles, en se basant par exemple sur la fréquence des observations dans un corpus de référence.

3 Le modèle de traduction

En traduction automatique statistique, étant donnée une phrase-source $f_1^J = f_1 ... f_J$, on recherche la phrase-cible $e_1^I = e_1 ... e_I$ qui en constitue la traduction la plus probable :

$$\hat{e}_1^I = \text{argmax}_{e_1^I}\{P(e_1^I|f_1^J)\}$$

Notre approche repose sur une modélisation directe de la probabilité a posteriori $P(e_1^I|f_1^J)$ au moyen d'un modèle log-linéaire :

$$P_\lambda(e_1^I|f_1^J) = \frac{1}{Z_{f_1^J}} \exp\left(\sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J)\right)$$

Dans cette équation, la contribution de chacune des *fonctions-attributs* h_m est pondérée par un facteur λ_m, lesquels constituent les paramètres du modèle; $Z_{f_1^J}$ représente un facteur de normalisation propre à la phrase source f_1^J. Il est possible d'introduire des variables additionnelles dans le modèle, de façon à tenir compte de phénomène cachés; on modifie alors les fonctions-attributs pour y incorporer ces variables. Par exemple, notre modèle doit prendre en compte l'ensemble des bi-segments qui est à l'origine d'une traduction; les fonctions-attributs auront donc la forme générale $h_m(e_1^I, f_1^J, b_1^K)$. Le recours à ce genre de modèle est maintenant monnaie courante en traduction automatique (Tillmann and Xia2003; Zens and Ney2003; Och and Ney2004).

Notre modèle repose présentement sur sept fonctions-attributs. h_{bp} est la *fonction-attribut des bi-segments*. Elle représente la probabilité de produire la phrase en langue-cible, étant donné le découpage de la source, tel que prescrit par l'ensemble de bi-segments utilisé, sous l'hypothèse que chaque segment-source génère un segment-cible de façon indépendante du reste de la phrase-source :

$$h_{bp}(e_1^I, f_1^J, b_1^K) = \sum_{k=1}^{K} \log P(\tilde{e}_k|\tilde{f}_k) \quad (1)$$

Les probabilités des segments-cible sont estimées sur la base de décomptes dans un corpus de référence aligné au niveau des mots. Cette fonction-attribut démontre une forte tendance à surestimer la probabilité des bi-segments peu fréquents. C'est pourquoi on introduit également une *fonction-attribut compositionnelle* h_{comp}, qui se calcule de la même façon que h_{bp} dans l'équation (1), sauf que les probabilités des segments-source sont estimées sur la base de probabilités de traduction des mots qui composent le bi-segment, à la manière du modèle IBM-1 (Brown et al.1993) :

$$P(\tilde{e}|\tilde{f}) = \frac{1}{|\tilde{f}|^{|\tilde{e}|}} \prod_{e \in \tilde{e}} \sum_{f \in \tilde{f}} P(e|f)$$

Ici encore, l'estimation des probabilités de traduction lexicales $P(e|f)$ se fonde sur des décomptes dans le corpus d'entraînement.

h_{tl} est la *fonction attribut langue-cible*. Elle repose sur un modèle N-gramme de la langue-cible. Elle ne tient donc compte que de la suite de mots e_1^I résultant de la combinaison des bi-segments.

Deux fonctions-attributs, h_{wc} et h_{bc}, contrôlent respectivement la longueur de la phrase-cible et le nombre de bi-segments ayant servi à produire celle-ci : $h_{wc}(e_1^I, f_1^J, b_1^K) = I$ et $h_{bc}(e_1^I, f_1^J, b_1^K) = K$. Une sixième fonction $h_{reord}(e_1^I, f_1^J, b_1^K)$ mesure le degré de divergence dans l'ordre des mots de la source et de la cible.

Toutes les fonctions ci-dessus font plus ou moins partie de l'arsenal habituel des fonctions-attributs employées en traduction automatique. Une seule fonction, h_{gc} concerne spécifiquement les segments discontinus, et permet au modèle de contrôler dans une certaine mesure la nature des segments qu'il utilise. Cette fonction prend pour valeur le nombre total de jokers apparaissant dans les segments (source ou cible) de b_1^K.

Nous choisissons les valeurs des paramètres λ_m de façon à maximiser la qualité des traductions produites sur un corpus d'entraînement, tel que proposé par (Och2003). À la différence de ce dernier, toutefois, nous avons développé une version de la métrique d'évaluation de traduction NIST (Doddington2002) qui est dérivable par rapport aux λ_m, ce qui ouvre la voie à l'utilisation de méthodes d'optimisation par descente de gradient (Newton, quasi-Newton, etc.). Pour chacune des phrases sources $f_1 ... f_S$ du corpus d'entraînement, notre système de traduction peut produire plusieurs phrases cibles $e_{s,k}$, ordonnées suivant les valeurs de $P_\lambda(e_{s,k}|f_s)$. Nous calculons alors une version de la métrique d'évaluation NIST, dans laquelle la contribution de chaque phrase est pondérée par :

$$w_{s,k}^{\alpha}(\lambda) = \frac{P_\lambda(e_{s,k}|f_s)^\alpha}{\sum_{k'} P_\lambda(e_{s,k'}|f_s)^\alpha},$$

où α est un paramètre de lissage qu'on fixe de manière expérimentale.

À la différence d'une approche par maximum de vraisemblance dans un modèle log-linéaire, qui correspond à un problème convexe et conduit à un minimum global unique, ce genre

d'apprentissage est assez sensible à l'initialisation des paramètres λ. Notre approche consiste alors à utiliser un ensemble d'initialisations aléatoires pour les paramètres, à effectuer l'optimisation pour chaque initialisation, et à choisir le modèle qui donne la meilleure performance.

Finalement, rappelons que cette procédure d'entraînement requiert des traductions multiples pour chaque phrase-source du corpus d'entraînement. En pratique, notre décodeur peut générer une liste des N-meilleures traductions de chaque phrase-source. Toutefois, différentes valeurs initiales des paramètres λ peuvent conduire à des listes différentes. Il est donc judicieux de répéter le processus : décodage des N-meilleures traduction, optimisation de la valeur de λ, re-décodage des N-meilleures traduction avec ces nouveaux paramètres, ré-optimisation de ceux-ci, etc. Afin d'assurer la convergence du processus d'optimisation, il convient de combiner à chaque itération les nouvelles N-meilleures traductions avec celles obtenues lors des itérations précédentes.

4 Le décodage

Notre méthode de décodage repose sur une recherche en faisceau par piles (*beam-search stack decoding*), tel que proposée dans (Koehn2003), que nous avons adaptée aux segments discontinus. La traduction d'une phrase en langue-source est le résultat d'une suite de *décisions*; chacune de celles-ci implique le choix d'un ensemble de positions à couvrir dans la phrase-source et d'un bi-segment adéquat. La traduction finale s'obtient en combinant ces décisions dans l'ordre, comme à la figure 2. Au cours du processus de décodage, les traductions partielles (que nous appelons des *hypothèses*) sont accumulées dans des listes (les *piles*), chacune desquelles regroupe des hypothèses qui recouvrent le même nombre de mots dans la phrase-source. On *étend* une hypothèse en y comblant la première position libre dans la cible (voir la section 3); chaque hypothèse ainsi étendue est stockée dans la pile correspondant au nouveau nombre de mots traduits dans la source.

On associe un score à chaque hypothèse. Ce score est la combinaison d'une composante exacte et d'une composante heuristique : la composante exacte est obtenue en combinant la contribution des valeurs de fonctions-attributs des décisions qui constituent l'hypothèse; la composante heuristique se veut un estimé optimiste du coût nécessaire pour compléter la traduction, tenant compte notamment de la présence de segments discontinus. Chaque pile fait l'objet d'un filtrage, visant à y éliminer les hypothèses les moins prometteuses. Ce filtrage se fonde à la fois sur la valeur du score et sur le nombre d'hypothèses dans la pile.

On trouve la traduction finale dans la "dernière" pile, c'est-à-dire celle correspondant à une couverture totale de la phrase-source. On récupère alors la traduction ayant le meilleur score, et qui constitue une phrase bien formée, c'est-à-dire sans jokers.

5 Évaluation

Nous avons effectué certaines expériences, visant à évaluer le potentiel de notre approche, et en particulier l'apport des bi-segments discontinus. Toutes nos expériences ont porté sur la traduction du français vers l'anglais. Nous avons utilisé des textes provenant du corpus *Aligned*

Corpus	phrases	mots-source	mot-cible
construction des bi-segments	931 000	17,2M	15,2M
entraînement no. 1	250	3646	3295
entraînement no. 2	250	3793	3441
test no. 1	250	3007	2745
test no. 2	250	3238	2949

Table 1: Caractéristiques des corpus utilisés.

nombre max. de jokers	source	cible	source et cible
0	1 047 101	1 224 910	831 034
1	2 232 226	2 448 223	1 959 154
2	3 403 827	3 403 827	3 403 827

Table 2: Distribution cumulative des bi-segments de B_2, en fonction du nombre maximum de jokers dans la source, la cible et les deux.

Hansards of the 36th Parliament of Canada[2]. De cet ensemble de données, nous avons extrait cinq sous-corpus : un corpus de *construction des bi-segments*, deux corpus d'*entraînement* et deux corpus de *test*. Ces sous-corpus ont été extraits des portions dites *training*, *test-1* et *test-2* des Hansard alignés. Pour des raisons d'efficacité, nous nous sommes limités aux phrases de 30 mots et moins, et à des corpus d'entraînement et de test de petite taille. Le tableau 1 résume les principales caractéristiques des corpus.

Nous avons construit des banques de bi-segments, suivant la méthode présentée à la section 2. Cette méthode génère potentiellement un très grand nombre de bi-segments. Or le temps requis pour le décodage croît de façon essentiellement linéaire avec le nombre de bi-segments disponibles. C'est pourquoi il importe de limiter la taille des banques. Pour ces expériences, nous nous sommes donc limités à la combinaison des ensembles L_1 à L_5, c'est-à-dire obtenu de toutes les combinaisons de 1, 2, 3, 4 ou 5 cepts du corpus de construction. Partant de là, nous avons construit deux banques, qui se différencient par le nombre maximal de jokers admis dans les segments source ou cible : les bi-segments de la banque B_0 ne comportent aucun joker (ce sont donc des bi-segments continus), alors que ceux de la banque B_2 comportent au plus 2 jokers dans la source ou la cible. Dans chacune de ces banques, nous avons exclus les bi-segments n'apparaissant qu'une fois dans le corpus, et pour tout segment-source, nous n'avons retenu que les 20 segments-cible les plus fréquents. La distribution cumulative des bi-segments dans la banque B_2 en fonction du nombre de jokers qu'ils comportent est donnée au Tableau 2.

Nous avons ensuite procédé à l'estimation des paramètres du modèle, suivant la méthode de la section 3 : partant de paramètres aléatoires, nous avons produit les 1000 meilleures traductions pour chacune des phrase des corpus d'entraînement. Nous avons effectués ce processus 3 fois, chaque fois partant de paramètres aléatoires différents, pour chacun des 2 corpus d'entraînement, afin de contrôler la stabilité du processus. Pour chacun des ensembles de données d'entraînement résultants, nous avons alors cherché les valeurs de λ_m maximisant le score NIST lissé, à partir de 100 initialisations aléatoires. Pour chacune des banques de bi-segments B_0 et B_2, nous avons effectué 2 itérations de ce processus; comme on peut le voir à la figure 3, le processus converge rapidement.

Les phrases des corpus de test ont ensuite été traduites avec les paramètres optimisés. Nous

[2]Corpus compilé par Ulrich Germann et distribué par le *USC Information Sciences Institute*

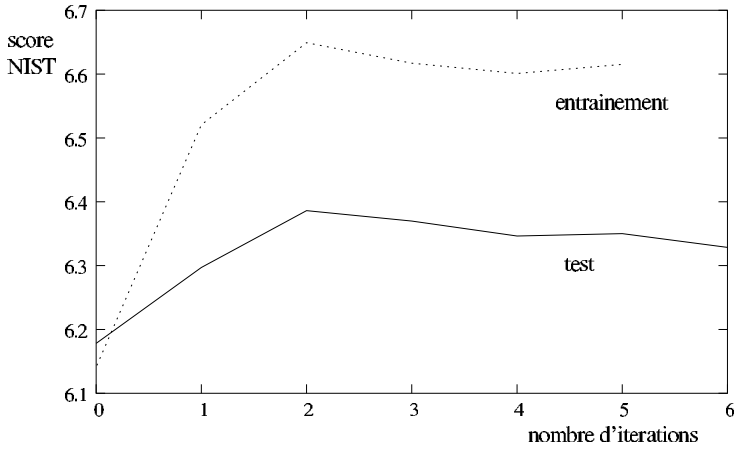


Figure 3: Variation du score NIST en fonction du nombre d'iterations

avons mesuré la qualité des traductions en termes des métriques NIST et BLEU (Papineni et al.2002). À titre de comparaison, nous avons également produit un modèle IBM-4 à partir des données de construction des bi-segments et d'entraînement, à l'aide du système *GIZA++* (Och and Ney2000). Nous avons alors traduit les données de test à l'aide du décodeur *ReWrite* (Germann et al.2001). Les résultats de ces expériences sont rapportés au tableau 3.

Des valeurs supérieures des métriques NIST et BLEU indiquent de meilleures performances; globalement, notre système se comporte donc sensiblement mieux avec la banque B_2 qu'avec B_0, qui produit elle-même des résultats légèrement supérieurs à ceux obtenus avec un modèle IBM-4. Les banques B_0 et B_2 ne diffèrent que par la présence de segments discontinus dans B_2 : c'est donc en allant "piocher" parmi ceux-ci que le modèle arrive à améliorer ses résultats. Ceci semblerait donc supporter notre thèse, que l'utilisation de bi-segments discontinus est bénéfique.

En examinant plus attentivement les traductions produites avec la banque B_2, on constate que les bi-segments discontinus, bien que 3 fois plus nombreux dans la banque que leurs homologues continus, n'ont pas nécessairement la faveur du modèle de traduction. Par exemple, notre système a produit les traductions les plus probables pour les 250 phrases du corpus de test en utilisant 1479 bi-segments, soit 5,92 bi-segments par phrase en moyenne. De ce nombre, seulement 242 sont discontinus, soit moins de 17%, ou 0,96 bi-segment discontinu par phrase. C'est donc dire que dans nombre de situations, notre système préfère encore utiliser des bi-segments continus.

En pratique, les bi-segments discontinus sont utilisés dans des circonstances qui coïncident par-

Corpus	Expérience	*ReWrite*		B_0		B_2	
		NIST	BLEU	NIST	BLEU	NIST	BLEU
test no. 1	1	6,59	0,36	6,63	0,38	6,82	0,39
	2			6,65	0,38	6,83	0,38
	3			6,72	0,38	6,70	0,37
test no. 2	1	6,12	0,31	6,16	0,32	6,20	0,32
	2			6,20	0,32	6,34	0,34
	3			6,14	0,31	6,24	0,32

Table 3: Résultats expérimentaux

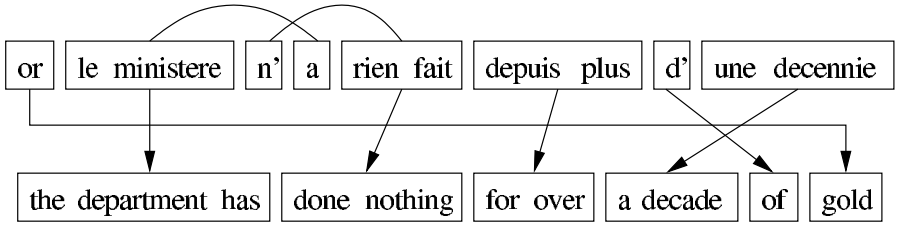


Figure 4: Exemple de traduction avec des bi-segments discontinus

fois avec certains des phénomènes que nous souhaitions voir ainsi traités, mais pas toujours. Et si l'apport des bi-segments discontinus est globalement positif, il reste que ceux-ci introduisent également des problèmes. La figure 4, qui montre un exemple de traduction provenant du corpus de test, tel qu'effectué avec la banque B_2, illustre assez bien la situation. D'une part, on voit comment les bi-segments discontinus permettent de traiter le cas de la négation en français : La combinaison de deux bi-segments ⟨*Le ministère* ⋄ *a*, *the department has*⟩ et ⟨*n'* ⋄ *rien fait*, *done nothing*⟩ permet d'arriver à une traduction assez judicieuse. Par ailleurs, comment expliquer cette mystérieuse apparition en fin de phrase du segment "*of gold*" (en français "*en or*" ou "*d'or*")? D'abord, le système a pris la conjonction de coordination française *or* pour un substantif, qu'il a traduite par *gold*. Il a alors récupéré la préposition *d'*, laissée pour compte dans le bi-segment ⟨*une décennie*, *a decade*⟩, et s'est servie de sa traduction la plus fréquente (*of*) pour introduire ce nouveau substantif.

De telles erreurs sont assez typiques du comportement de notre système dans son état actuel. Deux facteurs en sont vraisemblablement à l'origine. D'abord, nous n'admettons pas dans notre modèle la possibilité de bi-segments dont l'une ou l'autre partie serait vide, qui permettraient, par exemple, de rendre compte de la "disparition" de la préposition *d'* dans le passage à l'anglais. Mais la méthode d'alignement utilisée pour constituer les banques de bi-segments est également en cause ici. En pratique, on constate que les mots-outils qui ne sont pas explicitement traduits sont souvent mal alignés, entraînant la présence de bi-segments "faussement discontinus" dans la banque, par exemple ⟨*devons essayer*, *need* ⋄ *try*⟩ dans laquelle la préposition anglaise *to* est escamotée, ou encore ⟨*soins* ⋄ *santé*, *health care*⟩, dans laquelle c'est le *de* français qui a disparu. De tels bi-segments, combinés à une absence de traitement des insertions et suppressions, entraînent forcément des erreurs de traduction.

6 Conclusions

Nous avons présenté une approche de la traduction automatique statistique par segments de texte discontinus. Une première implantation de cette approche nous a permis de valider le bien-fondé de notre hypothèse de départ, suivant laquelle ces segments discontinus permettraient de mieux représenter certains phénomènes linguistiques, et ainsi de faire meilleur usage des donnés d'apprentissage.

Dans l'implantation actuelle de notre système, le temps requis pour le décodage est encore souvent prohibitif, ce qui ralentit notamment le cycle d'apprentissage des paramètres. Ceci est d'autant plus critique que certaines expériences semblent indiquer que la qualité des traductions produites par notre système aurait beaucoup à gagner d'un volume plus important de données d'entraînement. Nous examinons présentement différentes stratégies d'optimisation du processus de décodage. Mais le nombre de bi-segments disponibles au moment de la traduction d'une

phrase demeure un facteur dominant de complexité. Le rôle relativement mineur que jouent finalement les bi-segments discontinus dans les traductions optimales suggère qu'on pourrait effectuer une sélection plus judicieuse des bi-segments dès l'étape de construction des banques. Une hypothèse qui nous apparaît prometteuse est celle suivant laquelle les bi-segments qui sont réellement utiles sont ceux qui représentent des traductions de nature non-compositionnelles. La construction des banques pourrait donc incorporer une mesure de compositionnalité, par exemple une variante de l'information mutuelle (Lin1999). Par ailleurs, les bi-segments de nos banques sont relativement petits (en moyenne, moins de 4 mots), lorsqu'on les compare à ceux utilisés dans des systèmes comparables (par exemple, jusqu'à 7 mots dans (Och and Ney2004)). Nous envisageons d'incorporer des segments discontinus beaucoup plus grands qui, plutôt que d'être calculés a priori, proviendraient d'une recherche directe dans le corpus d'entraînement. De tels segments, comparables à des repérages approximatifs ("*fuzzy matches*") dans une mémoire de traduction, joueraient alors le rôle de "phrases à trous" dans le processus de décodage.

Références

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

George Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In *Proceedings of the ARPA Workshop on Human Language Technology*.

U. Germann, M. Jahr, K. Knight, D. Marcu, and K. Yamada. 2001. Fast Decoding and Optimal Decoding for Machine Translation. In *Proceedings of ACL'01*, Toulouse, France.

Cyril Goutte, Kenji Yamada, and Eric Gaussier. 2004. Aligning Words Using Matrix Factorisation. In *Proceedings of ACL'04*, pages 503–510.

Philipp Koehn. 2003. *Noun Phrase Translation*. Ph.D. thesis, University of Southern California.

Dekang Lin. 1999. Automatic Identification of Non-compositional Phrases. In *Proceedings of ACL'99*, pages 317–324, College Park, USA, June.

Daniel Marcu and William Wong. 2002. A Phrase-based, Joint Probability Model for Statistical Machine Translation. In *Proceedings of EMNLP'02*, Philadelphia, USA.

F. J. Och and H. Ney. 2000. Improved Statistical Alignment Models. In *Proceedings of ACL'00*, pages 440–447, Hongkong, China, October.

Franz Josef Och and Hermann Ney. 2004. The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics*, 30(4):417–449.

Franz Josef Och, Christoph Tillmann, and Hermann Ney. 1999. Improved Alignment Models for Statistical Machine Translation. In *Proceedings of EMNLP/VLC'99*, College Park, USA.

Franz Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of ACL'03*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evalution of Machine Translation. In *Proceedings of ACL'02*, pages 311–318, Philadelphia, USA.

Christoph Tillmann and Fei Xia. 2003. A Phrase-Based Unigram Model for Statistical Machine Translation. In *Proceedings of HLT-NAACL 2003*, Edmonton, Canada.

Richard Zens and Hermann Ney. 2003. Improvements in Phrase-Based Statistical Machine Translation. In *Proceedings of HLT-NAACL 2003*, Edmonton, Canada.

Alignement de mots par apprentissage artificiel de règles de propagation syntaxique en corpus de taille restreinte

Sylwia Ozdowska (1), Vincent Claveau (2)

(1) ERSS - Université de Toulouse le Mirail
5 allées Antonio Machado
31058 Toulouse Cedex 1
ozdowska@univ-tlse2.fr

(2) OLST - Université de Montréal
CP 6128 succ. Centre-Ville
Montréal, QC, H3C 3J7, Canada
vincent.claveau@umontreal.ca

Mots-clefs : alignement de mots, corpus alignés, apprentissage artificiel, programmation logique inductive, analyse syntaxique

Keywords: word alignment, aligned corpus, machine learning, inductive logic programming, syntactic analysis

Résumé Cet article présente et évalue une approche originale et efficace permettant d'aligner automatiquement un bitexte au niveau des mots. Pour cela, cette approche tire parti d'une analyse syntaxique en dépendances des bitextes effectuée par les outils SYNTEX et utilise une technique d'apprentissage artificiel, la programmation logique inductive, pour apprendre automatiquement des règles dites de propagation. Celles-ci se basent sur les informations syntaxiques connues pour ensuite aligner les mots avec une grande précision. La méthode est entièrement automatique, et les résultats évalués sur les données de la campagne d'alignement HLT montrent qu'elle se compare aux meilleures techniques existantes. De plus, alors que ces dernières nécessitent plusieurs millions de phrases pour s'entraîner, notre approche n'en requiert que quelques centaines. Enfin, l'examen des règles de propagation inférées permet d'identifier facilement les cas d'isomorphismes et de non-isomorphismes syntaxiques entre les deux langues traitées.

Abstract This paper presents and evaluates an effective yet original approach to automatically align bitexts at the word level. This approach relies on a syntactic dependency analysis of the texts provided by the tools SYNTEX and uses a machine-learning technique, namely inductive logic programming, to automatically infer rules called propagation rules. These rules make the most of the syntactic information to precisely align words. This approach is entirely automatic, and results obtained on the data of the HLT evaluation campaign rival the ones of the best existing alignment systems. Moreover, our system uses very few training data: only hundreds of sentences compared to millions for the existing systems. Furthermore, syntactic isomorphisms between the two spotted languages are easily identified through a linguistic examination of the inferred propagation rules.

1 Introduction

L'enjeu que représente l'alignement des corpus parallèles au niveau des mots n'est plus à démontrer : ce dernier trouve ses applications dans des tâches telles que la traduction automatique ou encore la construction de ressources lexicales bi ou multilingues (Véronis, 2000). Il existe principalement deux types d'approches pour aligner des mots : celles à dominante statistique qui s'appuient notamment sur les modèles IBM (Brown *et al.*, 1993), et celles qui tendent à combiner calculs statistiques simples et différentes sources d'information linguistique (Ahrenberg *et al.*, 2000 ; Barbu, 2004). Destinés principalement à la traduction automatique, les systèmes purement statistiques se sont progressivement enrichis en incorporant des données linguistiques issues de l'analyse syntaxique (Lin & Cherry, 2003 ; Ding & Palmer, 2004) et ce afin de mieux prendre en compte les variations systématiques entre les langues impliquées dans le processus de traduction (Dorr, 1994 ; Fox, 2002). L'alignement sur des bases purement syntaxiques a également fait l'objet de travaux : D. Wu (2000) a par exemple proposé une méthode basée sur une analyse en constituants ; S. Ozdowska (2004), dont nous reprenons le cadre expérimental, utilise quant à elle une analyse en dépendances dans le but de proposer une étude contrastive fine des divergences syntaxiques entre le français et l'anglais. Sa démarche a consisté à définir manuellement des règles d'alignement, dites de propagation, qui exploitent les relations de dépendance mises en évidence dans chaque partie d'un corpus parallèle.

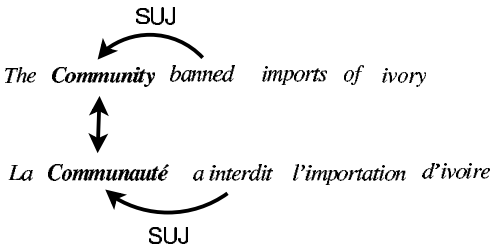
Cet article présente une technique d'alignement proche de cette dernière. Cependant, l'originalité de notre démarche réside dans le fait que les règles de propagation sont acquises de manière automatique en corpus par une technique d'apprentissage artificiel, la programmation logique inductive. Ces règles de propagation, exploitant des informations syntaxiques issues des analyseurs SYNTEX, sont automatiquement inférées à partir d'exemples d'alignements valides. L'objectif de cet article est d'une part de montrer que, contrairement aux approches statistiques, notre technique ne nécessite que très peu de données d'apprentissage. D'autre part, on se propose de vérifier si les règles obtenues et les alignements qu'elles produisent varient en fonction du type de corpus d'apprentissage.

Pour ce faire, nous exposons tout d'abord le cadre méthodologique dans lequel nous avons mené nos travaux. Puis, nous décrivons la technique d'apprentissage automatique des règles de propagation en section 3. Enfin, nous présentons et discutons les résultats obtenus en section 4 avant d'indiquer les perspectives de poursuite de ce travail.

2 Contexte d'expérimentation

2.1 Alignement de mots par propagation syntaxique

L'utilisation de règles de propagation pour aligner des bitextes au niveau des mots a déjà fait l'objet de plusieurs travaux. Ainsi, S. Ozdowska (2004) exploite les relations de dépendance syntaxique dans le processus d'alignement. Elle utilise des règles de propagation syntaxique définies à la main qui, étant donnés deux mots en relation d'équivalence dans un couple de phrases alignées, appelés *couple amorce*, permettent de propager le lien d'alignement à d'autres mots en suivant les relations de dépendance syntaxique connues. Dans l'exemple suivant, il est ainsi possible d'aligner *ban* et *interdire* en exploitant la relation sujet portant sur le couple amorce. Dans cet exemple et les suivants, les couples amorces sont notés en gras.



Chaque règle de propagation est donc décrite en fonction de la relation syntaxique qui sert de base à la propagation et de la direction dans laquelle s'effectue la propagation (et éventuellement des restrictions portant sur les parties du discours des mots concernés). Si nous reprenons l'exemple précédent, la règle de propagation anglais/français utilisée est :
V $\xrightarrow{\text{SUJ}}$ **Nom** / V $\xrightarrow{\text{SUJ}}$ **Nom**
Elle indique que la propagation se fait à partir d'un couple amorce de noms régis (*Community / Communauté)* vers un couple de verbes recteurs (*ban / interdire*) par la relation sujet. Une autre règle de propagation possible est celle qui va du couple amorce de noms régis (*ivory / ivoire*) au couple de recteurs (*imports / importation*) par la relation de préposition :
Nom $\xrightarrow{\text{PREP}}$ **Nom** / Nom $\xrightarrow{\text{PREP}}$ **Nom**

La plupart des règles utilisées dans ce type d'approche ont été définies en accord avec l'hypothèse d'isomorphisme direct entre les langues selon laquelle les structures syntaxiques seraient conservées lors de la traduction, comme dans l'exemple précédent (Hwa *et al.*, 2002). Cependant quelques unes traitent des cas de non-isomorphisme, comme l'alignement de *tax* et *fiscales* dans la biphrase : *tax* ***expenditures*** *have been (...) / les* ***dépenses*** *fiscales demeurent (...)*. Si l'on part du couple amorce de noms recteurs (*expenditures / dépenses*), les structures syntaxiques qui se correspondent dans les deux langues sont (NN représente la dépendance entre deux noms et MOD la dépendance générique tête-modifieur, ici nom-adjectif) :
Nom $\xrightarrow{\text{NN}}$ Nom/**Nom** $\xrightarrow{\text{MOD}}$ Adj

Ce type d'approche permet d'obtenir des alignements qui offrent en général une bonne précision, le rappel se révélant cependant de moins bonne qualité. En effet, le principe d'isomorphisme permet de générer des alignements corrects dans la plupart des cas où il s'applique mais il semble, dans certains cas, trop contraignant. Par ailleurs, ces approches nécessitent une expertise humaine pour écrire ces règles de propagation, ce qui peut se révéler coûteux. C'est ce dernier point que nous proposons de contourner en utilisant une technique d'apprentissage artificiel pour inférer automatiquement des règles de propagation.

2.2 Données d'apprentissage et d'évaluation

Nous avons choisi comme données de référence celles mises à disposition dans le cadre d'une campagne d'évaluation des systèmes d'alignement au niveau des mots notamment pour les paires de langues anglais/français (Mihalcea & Pederson, 2003). En voici la description (Och & Ney, 2000) :
- corpus d'entraînement anglais/français, issu du HANSARD (débats parlementaires), comptant 1.3 million de biphrases. Pour les expériences que nous reportons en section 4, nous n'avons utilisé qu'une portion variant de 10 à 1000 couples de phrases alignées de ce corpus.
- corpus de test constitué de 447 phrases alignées extraites d'une partie différente du HANSARD.

– le jeu de référence contient les alignements effectués par deux annotateurs sur le corpus de test. Chaque lien d'appariement établi s'est vu attribuer la valeur S, s'il s'agissait d'un lien considéré comme non ambigu, ou P dans le cas contraire. La valeur P est choisie en présence d'expressions figées ou de traductions libres. Dans le jeu de référence final, la valeur S est conservée pour les alignements pour lesquels il y a accord inter-annotateurs ; la valeur P est attribuée dans tous les autres cas. La figure 1 présente un exemple de phrase annotée ; les alignements S sont en traits pleins et les P en pointillés. Dans un premier temps, pour évaluer notre approche, nous nous focalisons sur les alignements 1-1 entre mots lexicaux ; les expérimentations décrites en section 4 ne portent donc que sur les annotations S entre mots lexicaux.

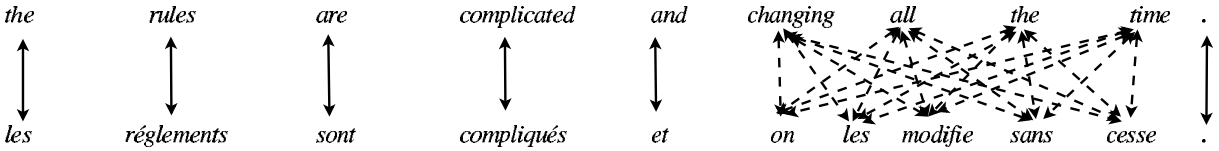


FIG. 1 – Annotation pour la campagne d'alignement HLT

Comme nous l'avons dit précédemment, en plus du HANSARD, les expériences d'inférence de règles de propagation que nous présentons en section 4 sont effectuées sur deux autres corpus. Le premier est un extrait du corpus INRA[1]. Il s'agit d'un corpus spécialisé anglais/français du domaine de la recherche agronomique de 1000 biphrases. Le second est un corpus fourni dans le cadre de la campagne d'évaluation ARCADE (Véronis & Langlais, 2000). Il est constitué de questions-réponses traitées à la Commission Européenne. Là encore, nous n'avons retenu que 1000 biphrases.

Le repérage des relations de dépendance syntaxique dans les trois corpus d'entrainement est effectué indépendamment pour chacune des deux langues par les analyseurs SYNTEX français et anglais (Bourigault & Fabre, 2000). Ces derniers prennent en entrée un texte étiqueté et identifient, pour chaque phrase, des relations telles que sujet, objet direct et indirect, modifieur... Les deux outils sont conçus suivant la même architecture et mettent en oeuvre les mêmes procédures de repérage des relations de dépendance. Par ailleurs, les relations identifiées ainsi que leur représentation sont globalement les mêmes d'une langue à l'autre.

3 Alignement par apprentissage artificiel

Comme nous l'avons déjà dit, l'originalité de notre approche tient au fait que contrairement aux travaux précédemment exposés (Ozdowska, 2004), les règles de propagation ne sont pas données manuellement mais inférées automatiquement. Les deux sous-sections suivantes présentent la technique d'apprentissage artificiel et son utilisation pour inférer ces règles. La technique d'amorçage fournissant automatiquement les exemples nécessaires à cette technique supervisée est décrite en sous-section 3.3.

[1]Nous remercions A. Lacombe, INRA, de nous avoir permis d'utiliser ce corpus.

3.1 Programmation logique inductive

Le principe de notre approche est le suivant : à partir d'exemples de propagations valides au sein de deux phrases alignées, on tente d'apprendre des règles qui les définissent. Pour ce faire, nous utilisons une technique d'apprentissage artificiel supervisée, la programmation logique inductive (PLI). Une présentation approfondie de cette méthode d'apprentissage peut être trouvée dans (Muggleton & De Raedt, 1994), nous n'en donnons ici que les grandes lignes. La PLI permet d'inférer des règles générales (des clauses de Horn) décrivant un concept à partir d'un jeu d'exemples de ce concept E^+ (avec éventuellement des contre-exemples E^-) et un ensemble d'informations externes B, appelées *Background Knowledge*. L'ensemble de règles inférées, appelé classifieur et noté H par la suite, est obtenu en généralisant les exemples en fonction de B.

Quelques conditions imposées à cette tâche d'apprentissage forment le cadre logique de la PLI ($\square$ signifie faux et $\models$ représente l'implication logique) :
- la consistance *a posteriori* impose qu'aucune contradiction n'existe entre B, H et E^+ : $B \wedge H \wedge E^+ \not\models \square$;
- la complétude assure que tous les exemples positifs sont expliqués avec H et les informations du *Background Knowledge*, soit $B \wedge H \models E^+$.

En pratique, les règles composant H sont recherchées à travers un espace d'hypothèses regroupant toutes les règles possibles. Cet espace est organisé hiérarchiquement, ce qui permet de le parcourir efficacement. Une règle de cet espace est retenue si elle maximise un score, généralement défini en fonction du nombre d'exemples (et éventuellement de contre-exemples) qu'elle couvre. La PLI, de par son expressivité (exemples et règles sont exprimés en logique des prédicats), a été utilisée pour de nombreuses tâches d'apprentissage, et notamment en TAL (Cussens & Džeroski, 2000).

3.2 Apprentissage de règles de propagation

Dans notre cas, les règles recherchées sont des propagations et les exemples que nous utilisons sont des phrases alignées comportant des alignements valides ; nous n'utilisons pas de contre-exemples. L'algorithme de PLI que nous utilisons est ALEPH[2]. Dans B sont stockées toutes les informations concernant les dépendances syntaxiques entre mots des phrases exemples et les couples amorces connus. Le formalisme logique de la PLI permet d'encoder facilement ces informations relationnelles. Ainsi, si l'on sait que *companies/entreprises* peuvent être alignés dans l'extrait de biphrase suivant (l'identifiant de chaque mot est noté après les barres obliques) :
... private/id_1_en ***sector****/id_2_en companies/id_3_en*
... les/id_1_fr entreprises/id_2_fr du/id_3_fr ***secteur****/id_4_fr privé/id_5_fr*
on ajoute à E^+ : alignement(id_3_en,id_2_fr). et à B (le nom du prédicat représente le nom de la relation syntaxique, le premier argument représente le recteur et le second le régi) :

```
determinant(id_2_fr,id_1_fr).   prep_de(id_2_fr,id_3_fr).   adjectif(id_2_en,id_1_en).
preposition(id_3_fr,id_4_fr).   adjectif(id_4_fr,id_5_fr).   nom_nom(id_3_en,id_2_en).
amorce(id_2_en,id_4_fr).
```

Une règle qui peut être inférée à partir de cet exemple est :

```
alignement(M_Ang,M_Fr) :- nom_nom(M_Ang,A1), prep_de(M_Fr,F1), preposition(F1,F2),
```

[2] ALEPH est développé par A. Srinivasan et disponible à http://web.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph.

amorce(A1,F2).
Avec les notations précédentes, cette règle s'écrit :
M_Ang $\xrightarrow{NN}$ **A1** / M_Fr $\xrightarrow{PREP\_DE}$ F1 $\xrightarrow{PREP}$ **F2.**
Elle souligne l'équivalence des structures Nom-Nom en anglais avec Nom de Nom en français ; tout couple apparaissant dans une biphrase avec cette structure peut ainsi être aligné.

3.3 Amorçage

Des exemples d'alignements valides sont nécessaires à notre technique d'apprentissage. Cette phase de supervision, pénible si elle était conduite manuellement, est dans notre cas automatisée par une technique dite d'amorçage. Notre approche d'inférence de règles de propagation ne requiert donc finalement aucune intervention humaine ; elle est dite semi-supervisée.

Pour générer ces alignements exemples, ou couples amorces, nous utilisons deux approches complémentaires. Il s'agit d'une part d'une technique statistique classique et d'autre part d'une recherche de cognats. En ce qui concerne la méthode statistique, nous considérons comme couples amorces les paires de mots (anglais/français) apparaissant ensemble dans des phrases alignées de manière statistiquement significative (Ahrenberg *et al.*, 2000) ; la force du lien entre deux mots est calculée par un Jaccard sur les fréquences d'apparition conjointe des deux mots (Ozdowska, 2004). Pour le repérage de cognats, c'est-à-dire de chaînes de caractères identiques ou proches dans les deux langues, la méthode mise au point est similaire à celle décrite dans (Fluhr *et al.*, 2000). Elle consiste à identifier la sous-chaîne maximale commune à deux mots qui cooccurrent dans un couple de phrases alignées.

Par la conjonction de ces deux méthodes, ce sont ainsi en moyenne entre 4 et 6 couples amorces (selon les corpus) par phrase qui sont détectés. Environ 5% des couples amorces se révèlent erronés (*i.e.* mots ne devant pas être alignés) ; ce faible taux ne devrait donc pas gêner le processus d'apprentissage. Chaque couple amorce, allié aux deux phrases alignées dont il est tiré, peut ainsi servir d'exemple pour notre technique d'apprentissage de règles de propagation. Une phrase permet donc de produire autant d'exemples qu'elle comporte de couples amorces.

À partir des exemples obtenus par cette technique, il nous est donc possible d'inférer des règles de propagation à partir de nos trois corpus d'entraînement. Ces règles peuvent ensuite être appliquées à de nouvelles données dans lesquelles on aura préalablement repéré des couples amorces.

4 Résultats

Cette section présente tout d'abord les résultats obtenus par notre approche sur le jeu d'évaluation HLT. Nous décrivons ensuite quelques causes d'erreurs récurrentes et examinons enfin certaines des règles inférées.

4.1 Performances d'alignement

Les trois systèmes d'alignement (*i.e.* les trois ensembles de règles inférées à partir de nos corpus) sont évalués à l'aide des données de la campagne HLT. Leurs performances sont présentées

de manière classique en termes de taux de rappel, taux de précision et f-mesure.

La table 1 présente les résultats des trois systèmes. Pour ces expériences, la phase d'apprentissage a été menée sur 1000 phrases de chaque corpus. À des fins de comparaison, nous indiquons les résultats obtenus par les meilleurs systèmes d'alignement – en terme de f-mesure – ayant participé à la compétition HLT ; il s'agit de Ralign (Simard & Langlais, 2003), XRCE (basé sur GIZA++) (Dejean *et al.*, 2003) et BiBr (Simard & Vogel, 2003), tous les trois utilisant principalement des approches statistiques. Nous indiquons aussi les résultats du système de S. Ozdowska (2004) dans lequel les règles de propagation sont définies manuellement.

Système	HANSARD	ARCADE	INRA	Ozdowska	Ralign	XRCE	BiBr
Précision	**88.51%**	82.65%	86.15%	81.59%	72.54%	55.54%	63.03%
Rappel	60.03%	60.25%	60.73%	58.43%	80.61%	**93.46%**	74.59%
F-mesure	71.54%	69.69%	71.24%	68.10%	**76.36%**	69.68%	68.32%

TAB. 1 – Performances des systèmes d'alignement sur les données HLT

À l'examen de ce tableau, on remarque que les résultats de nos systèmes varient peu en fonction du corpus d'entraînement. D'autre part, les performances de nos trois systèmes sont de niveau comparable à celles des autres. Ils se classent en effet deuxième derrière le système Ralign en terme de f-mesure. Ils jouissent par ailleurs d'une précision très supérieure aux autres systèmes, mais d'un rappel relativement plus bas. Ce rappel s'explique en partie par l'insuffisance de couples amorces et la couverture imparfaite de l'étiquetage syntaxique, ce qui rend certains couples inaccessibles à nos règles de propagation.

On s'intéresse dans un second temps à l'évolution des performances selon la taille des corpus d'entraînement. Pour cela, on fait varier le nombre de phrases servant à produire les exemples pour l'apprentissage. La figure 2 présente les taux de rappel, de précision et la f-mesure obtenus selon le nombre de phrases à partir du corpus HANSARD. Les résultats sont très éloquents : il n'y

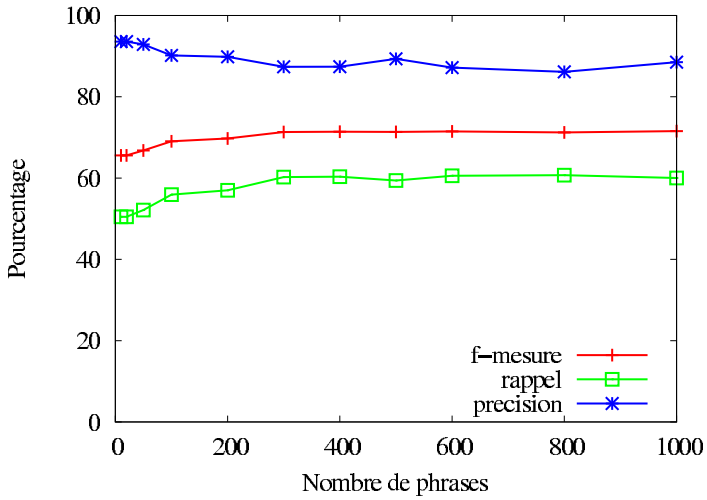


FIG. 2 – Variation des performances selon le nombre de phrases utilisées à l'apprentissage

a quasiment aucune variation de rappel et de précision de 300 à 1000 phrases. En dessous de 300 phrases, la précision augmente sensiblement alors que le rappel décroît. Cela s'explique par le

fait que seules quelques règles de propagation, parmi les plus sûres, sont trouvées. On remarque enfin qu'avec 10 phrases seulement, notre algorithme d'apprentissage est capable d'inférer des règles suffisamment pertinentes pour mener à une f-mesure de 65%. Ces résultats sont donc très positifs, notamment en regard des tailles très restreintes de nos corpus d'entraînement. À titre de comparaison, les systèmes Ralign, XRCE et BiBr utilisent 1.3 million de phrases pour s'entraîner.

4.2 Examen des résultats

Les erreurs d'alignement les plus courantes faites par nos systèmes peuvent se classer en plusieurs grandes catégories. Comme nous l'avons dit précédemment, une grande part des faux négatifs (*i.e.* des alignements non détectés) est due à une trop faible densité de couples amorces en plus d'absences de dépendances au sein de certaines phrases.

En ce qui concerne les faux positifs (*i.e.* des alignements détectés à tort), certains viennent simplement d'erreurs d'étiquetage de SYNTEX (elles-mêmes parfois causées par des erreurs de l'étiqueteur catégoriel utilisé en amont). Par exemple, dans la biphrase *federal government carpenters get $ 6.42/Les menuisiers du gouvernement fédéral touchent $ 6.42*, l'adjectif *federal* a incorrectement été rattaché à *carpenters*, ce qui a provoqué l'alignement incorrect de *carpenter/gouvernement*, tous deux notés recteurs du couple amorce *federal/fédéral*. D'autres erreurs de ce type sont causées par certaines des règles inférées qui ne sont pas assez spécifiques pour éviter de ramener du bruit. C'est notamment le cas des règles manipulant les dépendances objet ou sujet qui, à cause du manque d'informations dont dispose l'algorithme d'apprentissage, ne font pas de différence entre les voix actives et passives. Ainsi, à partir du couple amorce *bring/apporter* dans la biphrase *good legislation has been brought in by Liberal governments / les gouvernements libéraux ont apporté de bonnes mesures législatives*, *gouvernement* et *legislation* ont été alignés à tort. Enfin, des phénomènes de reformulations plus ou moins fidèles lors de la traduction perturbent parfois nos tentatives d'alignement. Ainsi, dans la phrase *the Government must implement the recommendations of the Commissioner of Official Languages/le gouvernement se doit de respecter les recommandations du Commissaire aux langues officielles*, *implement* et *respecter* ont été alignés alors que ce couple n'est pas noté valide dans le jeu de test HLT.

4.3 Règles obtenues

Environ une trentaine de règles de propagation sont obtenues pour chacun des corpus d'entraînement avec 1000 phrases. Il y a peu de différences entre ces règles dans les trois corpus, ce qui explique la proximité des performances observée en section 4.1. Elles sont, pour leur quasi totalité, très similaires à celles proposées par S. Ozdowska. Notons cependant que des règles, non retenues par S. Ozdowska, comme celles exploitant la coordination ou la relation attribut, se révèlent en pratique très productives et expliquent la différence de résultats avec notre approche par apprentissage.

La plupart des règles mettent donc en exergue des isomorphismes connus entre la syntaxe anglaise et française, comme l'alignement des adjectifs modifiant deux noms alignés, ou l'alignement des compléments d'objet direct de deux verbes alignés :

```
alignement(M_Ang,M_Fr) :- adjectif(C,M_Ang), adjectif(D,M_Fr), amorce(C,D).
```

alignement(M_Ang,M_Fr) :- objet(C,M_Ang), objet(D,M_Fr), amorce(C,D).
Ces cas d'isomorphismes parfaits représentent près de 50% des règles de propagation. Certains cas de non-isomorphisme syntaxique sont également trouvés, comme par exemple la construction standard des syntagmes nominaux Nom Nom en anglais et Nom de Nom en français (*cf.* section 3.2). D'autres types de non-isomorphismes peuvent même mener à l'alignement de mots ayant des parties du discours différentes, comme par exemple des noms et des adjectifs :
alignement(M_Ang,M_Fr) :- nom_nom(C,M_Ang), adjective(D,M_Fr), amorce(C,D).

D'une manière générale, il ressort de l'examen de ces règles que la plupart d'entre elles sont des règles de propagation que l'on peut qualifier de génériques. Elles sont effectivement pour une grande partie similaires à celles trouvées manuellement par S. Ozdowska (2004), ce qui confirme la validité de notre processus d'apprentissage. Cependant quelques règles inférées sont plus inattendues – et leur validité peut-être discutée – comme par exemple :
alignement(M_Ang,M_Fr) :- adjectif(M_Fr,C), nom_nom(D,M_Ang), adjectif(D,E), amorce(E,C).
qui permet d'aligner *bargaining* et *négociation* dans la biphrase ... *to have some hang-up with regard to the* ***collective*** *bargaining process*/... *éprouver certains complexes à l'égard de la négociation* ***collective***.

5 Conclusion et perspectives

Nous avons présenté une méthode originale d'alignement de mots basée sur la syntaxe et sur une technique d'apprentissage semi-supervisée. Celle-ci permet d'apprendre automatiquement des règles de propagation à partir d'exemples de couples de mots alignés. Ces exemples sont par ailleurs fournis à l'aide d'une procédure d'amorçage qui confère à notre approche une complète autonomie. Les résultats d'alignement obtenus sont bons et comparables aux meilleurs systèmes d'alignement actuels. De plus, et c'est l'originalité de ce travail, contrairement aux systèmes existants, très peu de données sont nécessaires pour entraîner notre système. On a montré également que les règles de propagation sont relativement génériques et changent peu d'un bitexte à un autre.

Plusieurs perspectives sont ouvertes par ce travail. Concernant la technique d'apprentissage, nous prévoyons d'intégrer les informations catégorielles pour permettre d'inférer des règles ne portant plus seulement sur les dépendances syntaxiques mais aussi sur les parties du discours. Cela permettra d'éviter certaines fausses détections reportées précédemment. L'utilisation d'exemples négatifs, qui permettraient d'empêcher des généralisations excessives et donc des règles de propagation pas assez précises, est également à l'étude. D'un point de vue applicatif, notre méthode étant entièrement automatique, elle peut aisément être adaptée à d'autres paires de langues, pourvu que celles-ci soient suffisamment proches d'un point de vue syntaxique et qu'un analyseur en dépendances existe pour chacune d'elles. Des expériences dans ce sens permettraient d'intéressantes études des cas d'isomorphismes et de non-isomorphismes syntaxiques dans les phrases alignées à travers l'étude des règles de propagation inférées.

Références

AHRENBERG L., ANDERSSON M. & MERKEL M. (2000). *A knowledge-lite approach to word alignment*, In *Parallel Text Processing: Alignment and Use of Translatin Corpora*, chapitre 5, p. 97–138. Kluwer Academic Publishers : Dordrecht.

BARBU A. M. (2004). Simple linguistic methods for improving a word alignment algorithm. In *7th International Conference on the Statistical Analysis of Textual Data, JADT'04*, Louvain-la-Neuve, Belgique.

BOURIGAULT D. & FABRE C. (2000). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de Grammaire*, 25, 131–151. Université Toulouse le Mirail.

BROWN P. F., PIETRA S. A. D., PIETRA V. J. D. & MERCER R. L. (1993). The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2), 263–311.

J. CUSSENS & S. DŽEROSKI, Eds. (2000). *Learning Language in Logic*. Lecture Notes in Artificial Intelligence. Sringer Verlag.

DEJEAN H., GAUSSIER E., GOUTTE C. & YAMADA K. (2003). Reducing parameter space for word alignment. In *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Edmonton, Alberta, Canada.

DING Y. & PALMER M. (2004). Automatic learning of parallel dependency treelet pairs. In *1st International Joint Conference on Natural Language Processing*, Sanya City, Chine.

DORR B. (1994). Machine translation divergences: A formal description and proposed solution. *Computational Linguistics*, 20(4), 597–633.

FLUHR C., BISSON F. & ELKATEB F. (2000). *Parallel Text Alignment Using Crosslingual Information Retrieval Techniques*, chapitre 9. In (Véronis, 2000).

FOX H. J. (2002). Phrasal cohesion and statistical machine translation. In *Empirical Methods in Natural Language Processing, EMNLP'02*, Philadelphia, PA, États-Unis.

HWA R., RESNIK P., WEINBERG A. & KOLAK O. (2002). Evaluating translational correspondence using annotation projection. In *40th Annual Conference of the Association for Computational Linguistics*, Philadelphia, PA, États-Unis.

LIN D. & CHERRY C. (2003). Proalign: Shared task system description. In *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Edmonton, Alberta, Canada.

MIHALCEA R. & PEDERSON T. (2003). An evaluation exercise for word alignment. In *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Edmonton, Alberta, Canada.

MUGGLETON S. & DE RAEDT L. (1994). Inductive Logic Programming: Theory and Methods. *Journal of Logic Programming*, 19-20, 629–679.

OCH F. J. & NEY H. (2000). Improved statistical alignment models. In *38th Annual Conference of the Association for Computational Linguistics*, Hong Kong.

OZDOWSKA S. (2004). Appariement bilingue de mots par propagation syntaxique à partir de corpus français/anglais alignés. In *conférence RECITAL'04*, Fès, Maroc.

SIMARD B. & VOGEL S. (2003). Word alignment based on bilingual bracketing. In *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Edmonton, Alberta, Canada.

SIMARD M. & LANGLAIS P. (2003). Statistical translation alignment with compositionality constraints. In *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Edmonton, Alberta, Canada.

J. VÉRONIS, Ed. (2000). *Parallel Text Processing : Alignment and Use of Translation Corpora*. Dordrecht : Kluwer Academic Publishers.

VÉRONIS J. & LANGLAIS P. (2000). *Evaluation of Parallel Text Alignment Systems. The ARCADE Project*, chapitre 19. In (Véronis, 2000).

WU D. (2000). *Bracketing and Aligning Words and Constituents in Parallel Text using Stochastic Inversion Transduction Grammars*, chapitre 7. In (Véronis, 2000).

Traduction de termes biomédicaux par inférence de transducteurs

Vincent Claveau (1), Pierre Zweigenbaum (2, 3 & 4)

(1) OLST - Université de Montréal
CP 6128 succ. Centre-Ville
Montréal, QC, H3C 3J7, Canada
vincent.claveau@umontreal.ca

(2) AP-HP, STIM/DSI, Hôpital Broussais,
96, rue Didot, 75674 Paris cedex 14
(3) INSERM, U729, 15, rue de l'École de Médecine, 75006 Paris
(4) INaLCO, CRIM, 2, rue de Lille, 75343 Paris cedex 07
pz@biomath.jussieu.fr

Mots-clefs : Traduction automatique de termes, terminologie biomédicale, apprentissage artificiel, inférence de transducteurs

Keywords: Automatic translation of terms, biomedical terminology, machine learning, transducer induction

Résumé Cet article propose et évalue une méthode de traduction automatique de termes biomédicaux simples du français vers l'anglais et de l'anglais vers le français. Elle repose sur une technique d'apprentissage artificiel supervisée permettant d'inférer des transducteurs à partir d'exemples de couples de termes bilingues ; aucune autre ressource ou connaissance n'est requise. Ces transducteurs, capturant les grandes régularités de traduction existant dans le domaine biomédical, sont ensuite utilisés pour traduire de nouveaux termes français en anglais et vice versa. Les évaluations menées montrent que le taux de bonnes traductions de notre technique se situe entre 52 et 67%. À travers un examen des erreurs les plus courantes, nous identifions quelques limites inhérentes à notre approche et proposons quelques pistes pour les dépasser. Nous envisageons enfin plusieurs extensions à ce travail.

Abstract This paper presents and evaluates a method to automatically translate simple terms from French into English and English into French in the biomedical domain. It relies on a machine-learning technique that infers transducers from examples of bilingual pairs of terms; no additional resources or knowledge is needed. Then, these transducers, making the most of high translation regularities in the biomedical domain, can be used to translate new French terms into English or vice versa. Evaluations reported show that our technique achieves good successful translation rates (between 52 and 67%). When examining at the most frequent errors made, some inherent limits of our approach are identified, and several avenues are proposed in order to bypass them. Finally, some perspectives are put forward to extend this work.

1 Introduction

Dans le domaine biomédical, l'évolution rapide des connaissances et la prédominance de l'anglais comme langue de communication rendent cruciales les problématiques de production et de gestion de ressources terminologiques multilingues. Dans ce cadre, la traduction de terminologies existantes (par exemple le thésaurus MeSH), dont fait l'objet cet article, revêt une grande importance. Par ailleurs, outre leur utilité pour les professionnels du domaine, les ressources terminologiques multilingues sont aussi essentielles à beaucoup d'applications du TAL et plus particulièrement pour la traduction automatique. En effet, l'un des problèmes majeurs de cette dernière, lorsqu'elle est appliquée à des textes spécialisés, est l'absence de ressources de traduction (terminologies ou corpus alignés) portant sur le domaine. Ainsi, les expériences menées par P. Langlais et M. Carl (2004) montrent que, dans certains textes, 35% des phrases contiennent au moins un mot inconnu de leur système de traduction généraliste. Les auteurs montrent que ces mots sont en fait des termes du domaine d'étude et soulignent donc l'importance de disposer de ressources terminologiques multilingues pour mener à bien ces tâches de traduction.

Dans cet article, nous présentons et évaluons une méthode automatique tentant de répondre à ces besoins dans un cadre toutefois restreint. Cette méthode doit permettre de produire des traductions de termes simples (*i.e.* composés d'un seul mot) du domaine biomédical d'une langue source dans une langue cible. Les expériences rapportées ici portent sur la traduction du français vers l'anglais et de l'anglais vers le français. Ce travail repose sur deux hypothèses majeures :

1. dans le domaine biomédical, les termes simples en anglais et français sont en majorité morphologiquement proches ;
2. les différences entre termes de chaque langue sont régulières et peuvent être apprises automatiquement.

Ces deux hypothèses tirent parti du fait que les termes biomédicaux français et anglais sont construits sur les mêmes racines grecques et latines, et leurs dérivations très régulières (*e.g. ophtalmorragie/ophthalmorrhagia*, *ophtalmoplastie/ophthalmoplasty*, *leucorragie/leukorrhagia*).

Notre approche s'appuie sur une technique d'apprentissage artificiel qui nous permet d'inférer un classifieur à partir de couples de termes français-anglais traduction l'un de l'autre et morphologiquement proches. C'est ce classifieur qui, étant donnés en entrée des termes français, doit ensuite permettre de produire les termes anglais correspondants ou inversement. Plus précisément, dans notre cas, le classifieur est un transducteur (*cf.* section suivante) et nous utilisons donc une technique existante d'inférence de transducteurs pour le générer à partir d'exemples de couples de termes bilingues. Il est intéressant de noter qu'à part cette phase de supervision, aucune autre connaissance, ni intervention humaine n'est requise.

Peu de travaux se placent dans le cadre de la traduction directe de termes. On peut néanmoins citer les travaux de S. Schulz *et al.* (2004) de traduction de termes biomédicaux du portugais vers l'espagnol fondés sur une analyse morphologique et l'utilisation de règles de réécriture fournies manuellement. Cependant, des problématiques proches sont souvent abordées dans le domaine de la traduction automatique de textes. Ainsi, l'acquisition de cognats (couples de mots bilingues de formes proches) (Fluhr *et al.*, 2000, *inter alia*) s'appuie sur des opérations morphologiques simples (distance d'édition, plus longue sous-chaîne commune) pour aligner des mots dans un corpus bilingue. Les transducteurs sont également parfois utilisés pour la traduction non pas de termes, mais de textes sous forme de chaînes de mots (Knight & Al-Onaizan, 1998) ou d'arbres syntaxiques (Knight & Graehl, 2005) ; les techniques de construction des transducteurs proposées dans ce cadre n'assurent cependant pas la même capacité à traiter des

séquences inconnues que celle que nous présentons ci-après. D'autres travaux reposent quant à eux sur des techniques statistiques de cooccurrences pour trouver des alignements entre mots ou termes dans des corpus alignés (Ahrenberg *et al.*, 2000; Gale & Church, 1991) ou comparables (Fung & McKeown, 1997). Outre le problème de la rareté de corpus spécialisés alignés, ces approches diffèrent de la nôtre en cela qu'il s'agit pour ces auteurs de retrouver une traduction d'un mot dans un texte (mise en relation), alors que nous nous posons dans le cadre plus strict de la traduction (génération). Mentionnons enfin les travaux sur la translittération, notamment du katakana ou de l'arabe (Tsuji *et al.*, 2002; Knight & Graehl, 1998, par exemple). Les techniques utilisées dans ceux-ci sont parfois proches de celle proposée ici, mais ne concernent que la représentation d'imports dans des langues ayant un alphabet différent de la langue source.

La section suivante présente la technique que nous utilisons pour inférer des transducteurs. Nous décrivons ensuite en section 3 la méthodologie employée pour nos expérimentations et les données utilisées. La section 4 détaille d'un point de vue quantitatif et qualitatif les résultats obtenus et nous concluons en donnant quelques perspectives ouvertes par ce travail.

2 Inférence de transducteurs

D'un point de vue général, un transducteur est un outil qui permet d'accepter en entrée des séquences d'un certain langage (langage étant pris ici au sens le plus large) et de produire en sortie les séquences associées dans un autre langage. L'emploi de ce type d'outils à des tâches de traduction en langage naturel semble donc évident. Cependant, en pratique, la complexité des relations entre langue d'entrée et langue de sortie impose deux importantes limites à leur utilisation :

1. l'expressivité des transducteurs n'est pas assez importante pour représenter certains phénomènes complexes de traduction ;
2. la construction des transducteurs pour ce type de tâche est trop complexe pour être menée manuellement.

Ces deux raisons expliquent pourquoi ce type de techniques est en pratique peu employée en dehors de tâches restreintes.

Dans le travail que nous présentons ici, ces deux problèmes cruciaux sont atténués par le fait que notre tâche de traduction est limitée : nous tentons de traduire des termes simples, dans un domaine restreint, entre des langues proches. Par ailleurs, pour ce qui est de la première limite citée ci-dessus, les types de transducteurs que nous utilisons sont relativement expressifs (*cf.* sous-section suivante) et les résultats présentés en section 4 semblent indiquer qu'ils sont adaptés à notre tâche. De plus, ces transducteurs ne sont pas construits manuellement mais inférés automatiquement à l'aide d'un algorithme d'apprentissage artificiel (voir Cornuéjols & Miclet (2002) pour une introduction) développé par J. Oncina (1991) et nommé OSTIA (*cf.* section 2.2). C'est ce dernier point qui fait l'originalité de ce travail et permet de contourner la seconde limite énoncée ci-dessus.

2.1 Transducteurs sous-séquentiels

Les transducteurs inférés par OSTIA — utilisés pour traduire nos termes biomédicaux — sont une extension des transducteurs classiques appelés transducteurs sous-séquentiels. Des définitions formelles de ces objets peuvent être trouvées dans (Oncina *et al.*, 1993) ; nous n'en

donnons ci-dessous que des descriptions générales.

Les transducteurs sont des machines à états finis que l'on peut voir comme des graphes dans lesquels un symbole d'entrée et une séquence de sortie sont associés à chaque arc. Un transducteur a un état initial et un ou plusieurs états finals. Intuitivement, un transducteur est donc un automate auquel on associe des séquences de sortie aux symboles d'entrée sur les arcs. Une séquence d'entrée E est *reconnue* ou *acceptée* s'il existe une suite d'arcs partant de l'état initial et arrivant à un état final telle que les symboles d'entrée, concaténés dans l'ordre de parcours de ces arcs, forment exactement E. La *traduction* d'une séquence d'entrée E correspond à la concaténation, dans l'ordre, de toutes les séquences de sortie des arcs traversés pour la reconnaissance de E.

Un transducteur séquentiel est un transducteur dans lequel tous les états sont finals, et où il est impossible d'avoir deux arcs sortant d'un même état ayant le même symbole d'entrée. Cette dernière propriété est celle qui assure le déterminisme des traductions issues de ces transducteurs. Enfin, un transducteur sous-séquentiel est un transducteur séquentiel dans lequel à chaque état est associée une séquence de sortie. Celle-ci est produite lorsque la séquence d'entrée se termine sur l'état ; c'est ce qui rend les transducteurs sous-séquentiels relativement expressifs.

La figure 1 présente un transducteur sous-séquentiel simple avec les notations habituelles des automates. Il représente la fonction de traduction qui fait correspondre un mot d'entrée vide à la séquence de sortie D, $a(bc)^n$ à $A(BC)^nE$ et $a(bc)^nb$ à $A(BC)^nBF$. En revanche, un mot comme $abca$ n'est pas reconnu, et donc non traduit par ce transducteur.

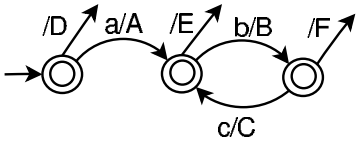


FIG. 1 – Exemple de transducteur sous-séquentiel

Dans notre cas, les séquences d'entrée sont soit les termes biomédicaux en français, vus comme des suites de lettres, et les séquences de sortie sont les termes correspondants en anglais, soit l'inverse.

2.2 Algorithme OSTIA

L'inférence de transducteurs est une technique d'apprentissage artificiel symbolique supervisée. Elle permet d'inférer, c'est-à-dire de produire automatiquement, un classifieur à partir d'exemples de couples séquence d'entrée/séquence de sortie ; dans notre cas, ce sont des couples de termes biomédicaux français-anglais. Ce classifieur est un transducteur qui reconnaît toutes les séquences d'entrée exemples et les traduit correctement en leur séquence de sortie correspondante, mais doit aussi idéalement être capable de produire la sortie correcte d'une chaîne d'entrée inconnue appartenant au même langage que les séquences d'entrée exemples. Les mécanismes de traduction des séquences d'entrée en séquences de sortie doivent donc être suffisamment réguliers pour permettre à l'algorithme d'apprentissage de *généraliser* les exemples ; on parle alors de saut inductif. Dans les travaux que nous présentons ici, cette phase d'inférence est mise en œuvre par l'algorithme OSTIA.

Cet algorithme d'inférence est formellement présenté par J. Oncina (1991), nous n'en décrivons ici que le principe général de fonctionnement. Nous l'illustrons à l'aide d'un exemple :

nous cherchons à apprendre le transducteur précédent (figure 1) avec les six exemples suivants : $\{\epsilon/D, a/AE, ab/ABF, abc/ABCE, abcb/ABCBF, abcbc/ABCBCE\}$ où ϵ représente la chaîne vide. L'algorithme OSTIA se déroule en trois étapes (Oncina, 1998) :

1. un arbre des préfixes de toutes les séquences d'entrée du jeu d'entraînement est construit. Des chaînes vides sont assignées à tous les nœuds et arcs internes de cet arbre, et à chaque nœud feuille est associée la séquence de sortie correspondant à la séquence d'entrée reconnue par cette branche (*cf.* figure 2).
2. tous les préfixes communs des séquences de sortie sont ensuite remontés des feuilles vers la racine de l'arbre (figure 3).

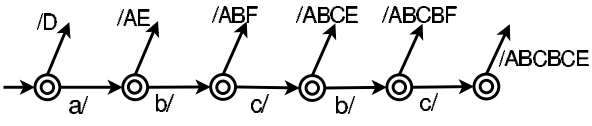


FIG. 2 – Transducteur après l'étape 1

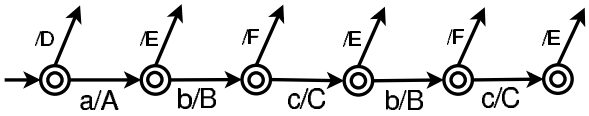


FIG. 3 – Transducteur après l'étape 2

3. enfin, en partant de la racine, tous les nœuds sont considérés deux à deux et fusionnés si le transducteur résultant n'entre pas en contradiction avec les données du jeu d'entraînement (figures 4 et 5). L'ordre de ces tentatives de fusion est généralement indiqué par une fonction heuristique. Il est possible de repousser des séquences de sortie vers les feuilles pour permettre des fusions. Quand plus aucune fusion n'est possible, l'algorithme termine.

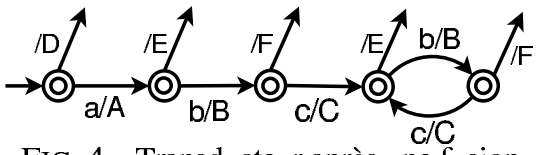


FIG. 4 – Transducteur après une fusion

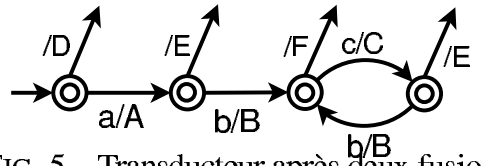


FIG. 5 – Transducteur après deux fusions

C'est bien sûr cette dernière étape qui assure le pouvoir inductif du transducteur, puisqu'elle assure de reconnaître et traduire correctement tous les couples exemples, mais permet également de reconnaître et traduire de nouvelles séquences d'entrée. Il a été montré formellement que cet algorithme converge et produit un transducteur décrivant le concept de traduction représenté par l'échantillon de données pairées constituant le jeu d'entraînement (Oncina, 1991). OSTIA a déjà été appliqué à de nombreuses tâches avec succès, dont la traduction de phrases en langage contrôlé (structures et vocabulaires restreints) (Oncina, 1998).

3 Expérimentations

Nous présentons tout d'abord les données utilisées pour nos expérimentations et la façon dont elles ont été préparées. Nous décrivons ensuite le cadre méthodologique des expériences d'inférence que nous avons menées à partir de ces données.

3.1 Constitution des données d'apprentissage et d'évaluation

Les données que nous utilisons pour évaluer notre technique sont issues d'un dictionnaire médical français en ligne (Dictionnaire Médical Masson, http://www.atmedica.com) contenant pour certaines de ses entrées les termes anglais équivalents. Parmi toutes les entrées, nous ne retenons que celles qui sont des mots simples à la fois en français et en anglais (absence d'espace et

de tiret), ne contenant pas de majuscules (pour éviter les noms propres) et d'une longueur minimale de 8 lettres (pour éviter les acronymes et se focaliser sur des termes techniques, contenant plusieurs morphes). Ce sont ainsi environ 12 000 paires de termes français-anglais qui sont obtenues des 35 000 entrées du dictionnaire.

Pour se focaliser sur les termes qui sont morphologiquement proches, la similarité formelle de chaque paire a été évaluée à l'aide d'une distance d'édition normalisée par la longueur des mots. Les paires sont classées dans une liste selon ce score en ordre décroissant de similarité :

```
1.00|zirconium|zirconium
...
0.93|ophtalmotoxine|ophthalmotoxin
0.93|ophtalmologiste|ophthalmologist
...
0.71|oschéite|oscheitis
0.71|organisé|organized
...
0.12|acouphène|tinnitus
0.11|engelures|chilblain
```

3.2 Méthodologie

Notre technique d'inférence repose entièrement sur les exemples, il est donc nécessaire de ne lui fournir pour l'entraînement que des paires effectivement morphologiquement proches, c'est-à-dire issues de la partie supérieure de la liste triée. Les données de test permettant d'évaluer notre approche peuvent quant à elles être tirées à n'importe quel niveau de la liste, même s'il semble évident qu'on ne peut pas attendre d'un quelconque système des traductions correctes des termes du bas de la liste. Aucun seuil n'apparaissant dans la distribution des mesures de similarité, on propose deux types d'expérience pour tenir compte de cela :

exp. 1. les paires d'entraînement et de test sont issues de la moitié supérieure de la liste ;

exp. 2. les paires d'entraînement sont issues de la moitié supérieure de la liste et celles de test de toute la liste.

Pour chacune des expériences, nous testons les sens de traductions français vers anglais et l'inverse, avec différentes tailles de jeux d'entraînement. Les jeux de test comportent 2000 paires (bien entendu différentes des paires d'entraînement) et les processus d'inférence et de validation sont répétés dix fois et les résultats moyennés. L'expérience 2 est une évaluation dans le pire des cas puisque les transducteurs inférés sont testés sur des paires qui peuvent n'être pas morphologiquement apparentées.

L'unique mesure utilisée pour rendre compte de la performance des transducteurs est la précision des traductions proposées. Cette précision est le ratio de termes de la langue source correctement traduits dans la langue cible (*i.e.* identiques aux termes attendus). Si le terme d'entrée n'est pas reconnu par le transducteur, il est considéré comme mal traduit.

4 Résultats

Cette section détaille les performances obtenues par les expériences présentées ci-avant. Nous en présentons d'abord les résultats d'un point vue quantitatif puis, en sous-section 4.2, nous

proposons un examen plus qualitatif des traductions issues des transducteurs. Enfin, à la lumière de ces résultats, nous mettons en évidence en sous-section 4.3 certaines limites inhérentes à notre approche.

4.1 Précision et complexité des transducteurs inférés

La précision des transducteurs est mesurée pour nos deux expériences selon le nombre d'exemples de couples d'entraînement utilisés par OSTIA. Les figures 6 et 7 présentent les courbes obtenues respectivement pour les expériences 1 et 2, dans les deux sens de traduction. Comme base de comparaison (*baseline*), nous calculons la précision qu'obtiendrait un système de génération de traduction simpliste proposant pour terme cible la même chaîne de caractères que le terme source (traduction à l'identique).

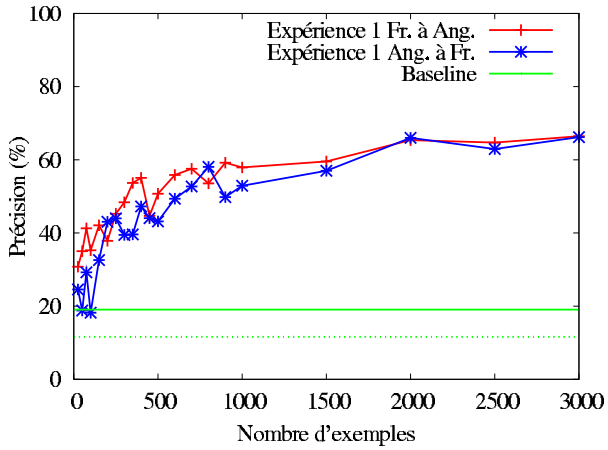


FIG. 6 – Précision selon le nombre d'exemples pour l'exp. 1

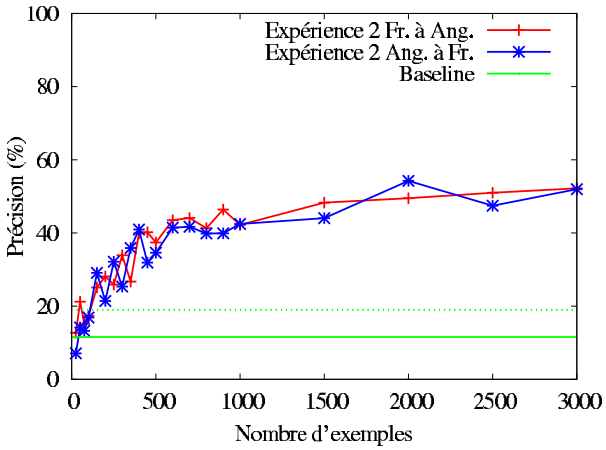


FIG. 7 – Précision selon le nombre d'exemples pour l'exp. 2

On remarque tout d'abord que le sens de traduction influe peu sur la précision. Par ailleurs, les résultats qui ressortent de ces figures sont plutôt bons. Pour 3000 exemples, la précision est d'environ 67% pour l'expérience 1 et 52% pour l'expérience 2, que ce soit du français vers l'anglais ou l'inverse. Comme l'indique la *baseline*, beaucoup de termes sont identiques en français et en anglais, mais les résultats des transducteurs dépassent largement cette base. Le taux de réussite élevé pour l'expérience 2 est particulièrement intéressant puisqu'il représente les résultats de notre technique sans aucune restriction, les tests étant effectués sur des couples parfois morphologiquement non apparentés. Ces résultats confirment donc le bien-fondé de notre approche, et notamment les deux hypothèses présentées en introduction.

Nous mesurons également la variation de la complexité des transducteurs inférés selon le nombre d'exemples. Cette complexité est mesurée en fonction du nombre de nœuds et d'arcs des transducteurs (figure 8) et en temps de calcul de la phase d'inférence (figure 9). Les informations reportées concernent l'expérience 1, dans le sens français vers l'anglais, celles des autres expériences étant similaires. On constate que la complexité arcs/nœuds est quasi linéaire en nombre d'exemples, mais très élevée. La taille de ces transducteurs est telle qu'il n'est d'ailleurs pas possible de les visualiser en totalité. La complexité en temps de calcul est plus problématique car elle augmente de manière exponentielle avec le nombre d'exemples, ce qui peut freiner l'utilisation de cette approche sur de plus larges jeux de données. Néanmoins, la précision se stabilisant assez rapidement, l'emploi de tels jeux de données semble inutile.

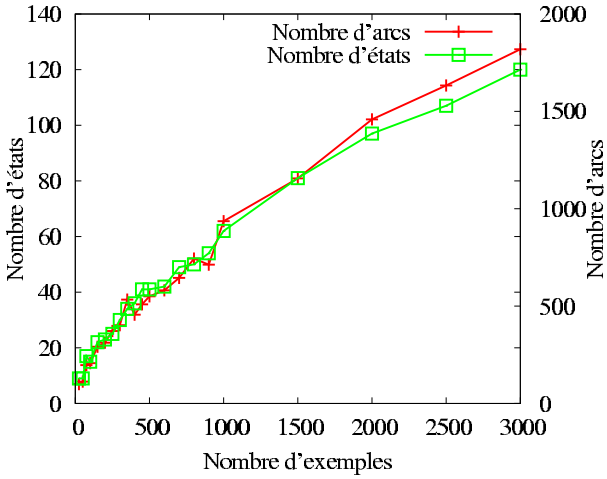


FIG. 8 – Complexité en nœuds et arcs selon le nombre d'exemples (exp. 1)

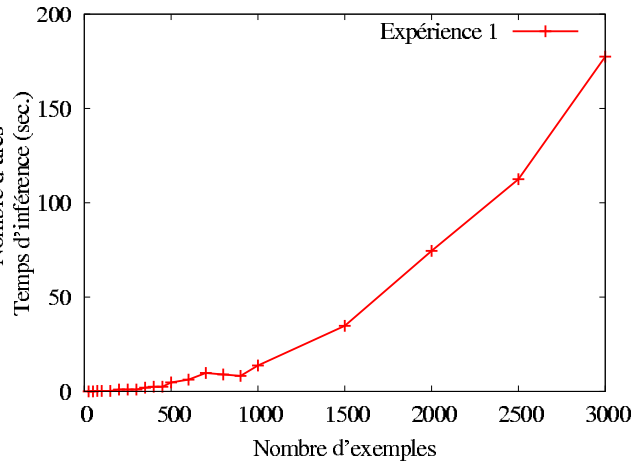


FIG. 9 – Complexité en temps selon le nombre d'exemples (exp. 1)

4.2 Examen des résultats

Les erreurs commises par les transducteurs inférés relèvent de différentes causes. Tout d'abord, certaines sont simplement dues à des termes dont la traduction n'est pas morphologiquement proche du terme d'origine. Comme attendu, ce type d'erreur est plus particulièrement présent dans l'expérience 2 (environ 10% des erreurs).

D'autres erreurs sont dues quant à elles à des exceptions aux régularités de traduction. On constate en effet que certaines traductions de familles de termes, bien que régulières dans leur majorité, présentent quelques cas particuliers. Ceux-ci font que les traductions proposées par les transducteurs inférés sont incorrectes. Par exemple, les termes français en *-rragie* se traduisent généralement en *-rrhagia* (*e.g.*, *stomatorragie/stomatorrhagia*, *pneumorragie/pneumorrhagia*...). Cependant, les couples *hémorragie/hemorrhage* et *pleurorragie/pleurorrhage* font exception à cette règle. Dans les expériences reportées précédemment, lorsqu'ils étaient rencontrés dans le jeu de test, *hémorragie* était (incorrectement) traduit en *hemorrhagia* et *pleurorragie* en *pleurorrhagia*. Il est intéressant de noter que ces termes irréguliers sont généralement d'emploi courant, appartenant presque à la langue générale, de laquelle ils ont certainement acquis cette dérivation particulière.

Enfin, certaines erreurs sont dues à l'apparente proximité de mots relevant de parties du discours ou de classes sémantiques différentes. Par exemple, les adjectifs français en *-ique* se traduisent généralement par un terme anglais en *-ic* (*e.g.*, *spasmolytique/spasmolytic*), alors que les noms avec le même suffixe se traduisent en *-ics* (*thermodynamique/thermodynamics*) ; ou encore, les noms de discipline en *-logie* se traduisent en *-logy* (*cardiologie/cardiology*) et les troubles du langage avec le même suffixe en *-logia* (*dyslogie/dyslogia*).

4.3 Limites de notre approche

Les erreurs de traduction présentées ci-avant mettent en relief certaines limitations de notre approche, ou plus précisément de la technique d'apprentissage utilisée. Tout d'abord, on ne peut apprendre que les régularités de traduction, il est donc normal que les exceptions, imprévisibles par nature, ne puissent pas être apprises. Cependant, ces exceptions, lorsqu'elles se trouvent

dans le jeu d'entraînement, risquent de provoquer de mauvaises inférences et complexifient les transducteurs. Malheureusement, OSTIA ne sait pas repérer ces paires irrégulières et ne permet pas d'apprendre des classifieurs distinguant règles générales et exceptions. Le même problème se pose pour gérer le bruit, c'est-à-dire des paires incorrectement encodées (faute d'orthographe dans l'un des termes ou erreur de traduction). Même si ce cas s'est peu présenté dans nos données, ce critère est à considérer si l'on souhaite employer cette même technique sur des paires de mots obtenues d'une source moins fiable.

Une autre limite de cette technique d'apprentissage est son incapacité à inclure des informations externes lors de sa phase d'inférence. En effet, OSTIA ne s'attache qu'à la suite de lettres composant les mots pour produire un transducteur, alors que, nous l'avons constaté, des informations catégorielles ou sémantiques permettraient de lever des ambiguïtés et d'améliorer les résultats. D'autres techniques d'apprentissage permettent d'adjoindre aisément ce type d'informations supplémentaires, comme la programmation logique inductive (Cornuéjols & Miclet, 2002, chapitre II.2), mais ne sont pas aussi performantes qu'OSTIA pour manipuler des séquences de lettres.

5 Conclusion et perspectives

Cet article présente une technique de traduction de termes simples du domaine biomédical du français vers l'anglais et inversement, s'appuyant sur une technique d'apprentissage artificiel, l'inférence de transducteurs. Les transducteurs sont inférés par l'algorithme OSTIA à partir d'exemples de paires de termes bilingues. Ils permettent ensuite, étant donné un terme dans une langue, de générer le terme correspondant dans une autre langue. Aucune connaissance ou ressource autre que les exemples n'est requise, laissant augurer une bonne portabilité de cette technique à d'autres paires de langues. Les évaluations que nous présentons montrent que cette technique obtient de bons résultats en produisant entre 50% et 66% de traductions correctes selon les expériences. On note de plus que certaines des erreurs de traduction sont dues à des mots largement utilisés, relevant presque de la langue courante. Ces mots ont par conséquent une grande chance d'apparaître dans des dictionnaires ou autres ressources de traduction, et donc ne nécessiteront pas l'emploi de transducteurs pour les traduire.

Beaucoup de perspectives sont ouvertes par ce travail. D'un point de vue technique, on peut notamment envisager d'appliquer cette même technique en considérant les termes non plus comme des séquences de lettres mais des séquences de morphes (*e.g. broncho⊕pleuro⊕pneumo⊕nie*). Des systèmes d'analyse morphologique dérivationnelle et compositionnelle existent déjà pour le domaine biomédical en français (Namer & Zweigenbaum, 2004) et pourraient ainsi servir de première étape à OSTIA. Une autre extension possible porte sur la recherche de traductions de termes complexes (à plusieurs mots). En effet, si l'on dispose de la traduction individuelle de chaque composant d'un terme complexe, on peut construire ou rechercher celle du terme pris dans sa globalité. Mais il faut pour cela tenir compte des variations possibles de ces termes (*virus de la variole/virus variolique*, *variola virus/variolic virus*) (Jacquemin, 2001; Daille, 2003).

D'un point de vue applicatif, nous envisageons d'utiliser et d'évaluer cette approche sur d'autres paires de langues (comprenant notamment l'espagnol, le portugais, l'allemand). Enfin, notre technique pourrait être utilisée pour l'alignement de corpus, les systèmes existants fonctionnant d'autant mieux que des couples de mots en relation de traduction sont connus (Véronis, 2000). Il sera alors intéressant de mesurer le gain de cette approche par rapport à une simple distance d'édition, communément utilisée dans ce contexte. Elle peut même être utilisée pour aligner

directement des ressources terminologiques. Dans ces deux cas, le problème abordé est quelque peu différent puisque, comme évoqué en introduction, il s'agit alors de mettre en relation des termes et non plus de produire des traductions.

Remerciements

Nous tenons à remercier José Oncina pour nous avoir donné accès au code d'OSTIA, et François Coste pour nous avoir fait partager son expérience sur l'inférence de langages réguliers.

Références

AHRENBERG L., ANDERSSON M. & MERKEL M. (2000). *A knowledge-lite approach to word alignment*, chapitre 5. In (Véronis, 2000).

CORNUÉJOLS A. & MICLET L. (2002). *Apprentissage artificiel*. Paris : Eyrolles.

DAILLE B. (2003). Conceptual structuring through term variation. In *Proceedings of the ACL'03 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, p. 9–16, Sapporo, Japon.

FLUHR C., BISSON F. & ELKATEB F. (2000). *Parallel text alignment using crosslingual information retrieval techniques*, chapitre 9. In (Véronis, 2000).

FUNG P. & MCKEOWN K. (1997). A technical word and term translation aid using noisy parallel corpora across language groups. *Machine Translation*, 12(1/2), 53–87.

GALE W. & CHURCH K. (1991). Identifying word correspondences in parallel texts. In *Proceedings of the 4th Darpa Workshop on Speech and Natural Language*, p. 152–157, Pacific Grove, CA, États-Unis.

JACQUEMIN C. (2001). *Spotting and Discovering Terms through NLP*. Cambridge : MIT Press.

KNIGHT K. & AL-ONAIZAN Y. (1998). Translation with Finite-State Devices. In *Third Conference of the Association for Machine Translation in the Americas, AMTA'98*, p. 421–437, Langhorne, États-Unis.

KNIGHT K. & GRAEHL J. (1998). Machine transliteration. *Computational Linguistics*, 24(4), 599–612.

KNIGHT K. & GRAEHL J. (2005). An Overview of Probabilistic Tree Transducers for Natural Language Processing. In *Proceedings of the 6th International Conference, CICLing 2005,*, Mexico, Mexique.

LANGLAIS P. & CARL M. (2004). General-purpose statistical translation engine and domain specific texts: Would it work ? *Terminology*, 10(1), 131–152.

NAMER F. & ZWEIGENBAUM P. (2004). Acquiring meaning for french medical terminology: contribution of morpho-semantics. In *Proceedings of the Conference MEDINFO 2004*, San-Francisco, États-Unis.

ONCINA J. (1991). *Aprendizaje de lenguajes regulares y transducciones subsecuenciales*. Thèse de doctorat, Universidad Politécnica de Valencia, Valence, Espagne.

ONCINA J. (1998). The data driven approach applied to the OSTIA algorithm. In *Proceedings of the Fourth International Colloquium on Grammatical Inference, ICGI'98*, p. 50–56, Ames, États-Unis.

ONCINA J., GARCÍA P. & VIDAL E. (1993). Learning subsequential transducers for pattern regognition interpretation tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(5), 448–458.

SCHULZ S., MARKÓ K., SBRISSIA E., NOHAMA P. & HAHN U. (2004). Cognate mapping - a heuristic strategy for the semi-supervised acquisition of a spanish lexicon from a portuguese seed lexicon. In *Proceedings of the 20th International Conference on Computational Linguistics, COLING'04*, p. 813–819, Genève, Suisse.

TSUJI K., DAILLE B. & KAGEURA K. (2002). Extracting french-japanese word pairs from bilingual corpora based on transliteration rules. In *Proceedings of the Third International Conference on Language Resources and Evaluation LREC'02*, p. 499–502, Las Palmas de Gran Canaria, Espagne.

J. VÉRONIS, Ed. (2000). *Parallel Text Processing*. Dordrecht : Kluwer Academic Publishers.

Traitement automatique de la saillance

Frédéric Landragin
Thales Research & Technology
Domaine de Corbeville
91404 Orsay CEDEX
Frederic.Landragin@thalesgroup.com

Mots-clefs : facteurs de saillance, saillance linguistique, saillance visuelle, principe de primordialité, principe de singularité, structure communicative, méthodes de quantification

Keywords: salience factors, linguistic salience, visual salience, primordiality principle, singularity principle, communicative structure, quantifying methods

Résumé Lorsque nous écoutons un énoncé ou que nous lisons un texte, les phénomènes de saillance accaparent notre attention sur une entité du discours particulière. Cette notion de saillance comprend un grand nombre d'aspects, incluant des facteurs lexicaux, syntaxiques, sémantiques, pragmatiques, ou encore cognitifs. En tant que point de départ de l'interprétation du langage, la saillance fonctionne de pair avec la structure communicative. Dans cet article, notre but principal est de montrer comment aboutir à un modèle computationnel de la saillance, qui soit valable aussi bien pour la saillance linguistique que pour la saillance visuelle. Pour cela, nous retenons une liste de facteurs qui contribuent à rendre saillante une entité. Dans le cas d'une entité du discours, cette approche nous permet de clarifier les rapports entre saillance et structure communicative. Nous définissons nos principes de primordialité et de singularité, puis nous passons en revue les différentes méthodes de quantification de la saillance qui sont compatibles avec ces principes. Nous illustrons alors l'une d'entre elles sur un exemple linguistique et sur un exemple visuel.

Abstract Salience attracts the attention on a particular discourse entity when hearing an utterance or reading a text. Salience is linked to a wide set of aspects from lexical, syntactic, semantic, and pragmatic factors to cognitive factors. Being the starting point of the interpretation process, salience works in close connection with communicative structure. In this article, our main purpose is to show how to tend to a computational model of salience, that can be used for linguistic salience as well as for visual salience. With this aim, we provide a list of factors that contribute to making an entity salient. For a discourse entity, this approach allows us to clarify the links between salience and communicative structure. We define our principles of primordiality and singularity, and we discuss the possible methods to quantify salience that are compatible with these principles. Then we illustrate one of them with a linguistic example and a visual example.

1 Introduction

La saillance intervient fortement lors de l'interprétation d'un énoncé en situation de dialogue ou lors de la compréhension d'un texte : mettant en avant un élément, elle dirige l'attention sur cet élément et rend sa prise en compte prioritaire dans le processus de résolution des références et des coréférences. Elle permet ainsi d'attribuer un référent à un pronom ou à une expression ambiguë telle que « *le N* » lorsque le contexte contient plusieurs objets du type N. La saillance linguistique constitue par exemple une aide à la résolution des anaphores : le N qui vient d'être mentionné est saillant et peut être repris par un pronom (Lappin, Leass, 1994). Dans le dialogue homme-machine sur écran, la saillance visuelle constitue un critère d'identification du N perçu de manière prioritaire et sur lequel est basé l'expression « *le N* » (cf. par exemple (Kievit et al., 2001)). La saillance due à la situation d'interaction permet la compréhension d'un pronom sans antécédent linguistique, comme dans l'exemple de (Isard, 1975) où un enfant tente de caresser un lion à travers une cage et se voit prévenir du danger par « *attention,* ***il*** *risque de te mordre* ». La saillance due au contexte de tâche et à l'historique de l'interaction joue le même rôle de désambiguïsation : « *le N* » peut se comprendre non pas comme « *le seul N* » mais comme « *le N qui vient d'être manipulé* » ou « *le N suivant dans la succession des tâches* ». Nous voyons en cela la saillance comme un indice important sur lequel baser l'identification de l'implicite, et donc comme un point de départ dans le processus de compréhension.

L'idée que nous défendons ici est que la saillance est un phénomène global, qui intègre les différentes facettes du contexte (perception visuelle, langage, tâche applicative, historique de l'interaction), et qu'une caractérisation de la saillance en vue de son calcul par un système de compréhension doit reposer sur des aspects plus généraux que les distinctions classiques entre thème et rhème ou entre topique et commentaire. Après une section dans laquelle nous précisons notre vision de la saillance et définissons les principes de primordialité et de singularité qui la caractérisent, nous proposons une caractérisation de la saillance linguistique par le biais d'une liste de facteurs de saillance. C'est sur la base d'une telle liste que peut se construire un modèle computationnel de la saillance. Nous décrivons alors les mécanismes des principales méthodes de calcul de la saillance. En adaptant la méthode de la moyenne des facteurs aux principes de primordialité et de singularité, nous montrons comment la méthode résultante peut être exploitée à la fois pour la saillance linguistique et pour la saillance visuelle. Les deux exemples que nous décrivons constituent ainsi un premier pas vers une modélisation de l'attention dans les systèmes de compréhension automatique.

2 Comment appréhender la saillance ?

Une première définition de ce qui est saillant est ce qui arrive en premier à l'esprit, ce qui capte l'attention. Ce point de vue correspond souvent à considérer comme saillant ce qui est naturel, simple, clair. Selon (Stevenson, 2002), les premiers travaux dans les années 1970 se sont focalisés sur la saillance visuelle avec cette idée de simplicité naturelle. (Osgood, Bock, 1977) parlent par exemple de simplicité naturelle (*naturalness*) et de clarté (*vividness*). Ainsi, l'ordre naturel des constituants dans une phrase reflète souvent l'ordre naturel des événements (agent–action–patient) et par conséquent une certaine hiérarchie de saillance. Du côté de la perception visuelle, ce point de vue rejoint celui de la Gestalt à propos de bonne forme : quelque chose de simple, d'immédiat, de percutant (Guillaume, 1979).

Deux autres définitions sont souvent avancées. En suivant la conception de la Gestalt selon laquelle la saillance dénote la force de résistance aux perturbations, est saillant ce qui est stable. D'un autre côté, est saillant ce qui est original ou nouvellement introduit dans la situation. Ce dernier facteur rejoint le critère de non-familiarité de (Loftus, Mackworth, 1978) : des objets non familiers tendent à être fixés plus longtemps et en deviennent saillants. Ce facteur rejoint également celui d'inattendu : est saillant l'élément perturbateur, inattendu, curieux, intrigant, énigmatique. Il s'agit en effet de l'élément sur lequel on s'interroge, ou sur lequel le regard va s'attarder pour résoudre le problème qu'il pose. Par exemple, tout élément visuel ou langagier pour lequel l'activité perceptive de reconnaissance s'avère difficile en devient saillant. Les deux définitions se contredisent ainsi totalement.

Un autre problème apparaît lorsque la saillance est définie par rapport à des facteurs cognitifs tels que l'attention, la mémoire ou la familiarité. Ces facteurs sont propres à l'utilisateur et montrent en quoi la saillance est subjective. Citons la familiarité culturelle (par exemple la présence d'un être humain ou d'un lion dans notre champ de vision est saillante et nous incite à nous tourner vers eux) et la familiarité individuelle (au cours de notre éducation et de notre vie passée, nous acquérons tous nos propres sensibilités). La saillance dépend de plus de l'attention de l'utilisateur au moment de l'interaction, de son intérêt, de son intention communicative : est saillant ce qui a de l'intérêt compte tenu de l'objectif de la communication. (Osgood, Bock, 1977) parlent ainsi de saillance liée à l'intérêt du locuteur (*motivation-of-speaker*). Nous pouvons déduire de ce facteur une dépendance de la saillance envers la tâche applicative : quand on invite des collègues à entrer dans un bureau, les chaises sont saillantes car ils vont vouloir s'asseoir. D'un autre côté, il n'y a pas équivalence entre saillance et intention communicative. La saillance peut au contraire prendre un rôle perturbateur : on peut se focaliser intentionnellement sur un objet précis tout en étant perturbé par un objet fortement saillant.

Tous ces problèmes se retrouvent lors du passage des définitions aux formalisations. Une illustration immédiate concerne la distinction entre « donné » (*given*) et « nouveau » (*new*) : le donné est saillant car il est stable et connu, mais le nouveau peut être tout autant voire plus saillant, étant justement nouveau et susceptible d'intriguer. Nous retrouvons la même difficulté à déterminer l'élément saillant dans la distinction entre « thème » et « rhème » : selon (Caron, 1989), la séquence la plus naturelle de deux phrases est celle où le rhème de la première phrase est repris comme thème de la seconde. Le rhème est repris et est donc saillant. Quant au thème de la nouvelle phrase, on peut aussi considérer qu'il est saillant. Caron conclut qu'il est impossible de savoir si c'est le thème ou le rhème qui est l'élément le plus saillant dans un énoncé. Même chose à propos de la distinction entre « présupposé » et « posé » : une information présupposée, donc implicite, peut s'avérer aussi saillante qu'une information explicite. Même chose à propos de la saillance de l'« agent » et de la saillance du « patient » : classiquement, par exemple dans la Théorie du Centrage (Grosz et al., 1995), l'agent est considéré comme plus saillant que le patient, lui-même considéré avant les autres rôles thématiques. Il n'y a cependant pas unanimité à propos de cette hiérarchie. Dans des travaux plus récents et basés sur des expérimentations, (Stevenson et al., 1994) montrent qu'entre agent et patient, la préférence est significativement pour le patient, du moins pour certaines constructions verbales. Ainsi, pour les phrases qui décrivent un événement, les conséquences de l'événement sont plus présentes dans les représentations mentales que les conditions initiales. Ces conséquences s'appliquant au patient, celui-ci en devient plus saillant que l'agent. Pour les phrases qui ne décrivent pas d'événement, tout dépend des composants de la phrase et rien ne peut être conclu. Pour déterminer quelle est l'entité saillante entre agent et patient, il faudrait détailler chaque type d'action, donc chaque type de verbe, en étudiant sa sémantique. Le problème s'avère complexe, du fait de

la multiplicité des critères qui entrent en jeu. (Stevenson et al., 1994) ainsi que (Pearson et al., 2001) détaillent par exemple le cas des verbes de transfert (donner quelque chose à quelqu'un) et montrent que le receveur est plus saillant que le donneur et que l'objet transféré. Bref, dans l'état actuel des recherches, il s'avère impossible de faire des liens définitifs entre la saillance et les distinctions classiques que nous avons citées. Les facteurs de saillance souvent avancés se recouvrent, se contredisent, et rien de précis ne peut être dégagé facilement.

Une dernière considération donnera le fil directeur de notre approche : dans le cadre d'une tâche consistant en la description téléphonique d'un itinéraire routier, (Edmonds, 1993) montre que la saillance dépend du contexte incluant le référent. Il donne l'exemple d'un immeuble *a priori* saillant par sa taille importante, et qui perd toute saillance lorsqu'il est entouré d'immeubles encore plus grands. Il considère également que certaines caractéristiques sont saillantes pour certains objets et pas pour d'autres, de même que certaines caractéristiques sont saillantes dans un but précis du dialogue et non dans un autre but. Par exemple, la caractéristique « taille » est saillante lorsque le but du dialogue est la désignation d'un immeuble, mais ne l'est pas lorsqu'il s'agit de la désignation d'une intersection de rues. Ainsi, plutôt que de chercher une définition propre à un objet ou à une structure en particulier, nous nous tournons vers l'émergence de saillance dans un ensemble contextuel. Les propriétés de l'énoncé et les propriétés physiques de l'ensemble des objets de la scène permettent de distinguer certains éléments, et de les considérer comme saillants. Plus précisément, la distinction ne peut se faire que selon deux principes que nous appelons « principe de primordialité » et « principe de singularité » :

- Principe de primordialité : l'entité saillante se distingue des autres entités du fait d'une importance particulière. Dans un énoncé décrivant un événement qui concerne un agent, un patient et un instrument, le patient est considéré comme le rôle thématique le plus important. Du fait de cette importance, l'entité correspondante devient l'entité du discours la plus saillante pour le rôle thématique.

- Principe de singularité : l'entité saillante se distingue des autres entités du fait d'une singularité. Dans une scène visuelle contenant des objets bleus et un objet rouge, l'objet rouge est le seul à être rouge, et, du fait de cette singularité, devient l'objet le plus saillant pour la couleur.

Selon la nature du facteur de saillance considéré, c'est ainsi le principe de primordialité ou le principe de singularité qui sera appliqué. La saillance ne dépendant pas d'un unique facteur, il nous reste à spécifier une liste de facteurs de saillance.

3 Une caractérisation de la saillance linguistique

Considérant qu'il est plus facile d'appréhender les facteurs de saillance visuelle que ceux de saillance linguistique, nous avons analysé les rôles des propriétés physiques des objets et des particularités visuelles d'une scène pour en dégager des facteurs applicables à la saillance en général. Nous retenons ainsi les facteurs suivants :

– facteurs intrinsèques aux unités (unités visuelles : objets ; unités linguistiques : mots) ;
– placement à un endroit stratégique (dans le cadre de l'image ; dans la structure de l'énoncé) ;
– isolement (singleton dans une partition en groupes perceptifs ; groupe de mots en apposition) ;
– rupture dans une continuité (par exemple dans le rythme de l'image ou de l'énoncé) ;

– répétition (dans la disposition des objets ; dans l'apparition des mots ou expressions) ;
– symétrie (dans la disposition des objets ; par une figure de style telle qu'un chiasme).

Ces facteurs mettent l'accent sur la structure de l'image ou de l'énoncé. D'autres facteurs relèvent du sens que prennent l'image et l'énoncé. Ce sont ces deux catégories (facteurs liés à la forme et facteurs liés au sens) que nous allons maintenant détailler dans le cas de la saillance linguistique.

Des facteurs liés à la forme de l'énoncé. C'est à la suite de (Stevenson, 2002) que nous distinguons les aspects formels, c'est-à-dire liés aux caractéristiques prosodiques et grammaticales de l'énoncé oral, des aspects sémantiques liés au contenu du message. Dans la majorité des travaux en linguistique computationnelle, par exemple dans (Alshawi, 1987), (Lappin, Leass, 1994), (Grosz et al., 1995), ou encore dans (Hajičová et al., 1995), ce sont essentiellement des facteurs formels qui définissent la saillance. Comme le soulignent (Krahmer, Theune, 2002), la récence est souvent mise en avant, les entités les plus saillantes étant définies comme étant les plus récemment mentionnées. A partir de ces travaux et de nos facteurs ci-dessus, nous proposons la classification suivante (cf. aussi (Landragin, 2004) pour une description détaillée) :

1. La saillance intrinsèque au mot : mots constitués de phonèmes sonores ; noms propres ; déictiques purs du fait de leur manque d'autonomie référentielle et de l'habitude qu'ils entraînent chez l'interlocuteur à faire attention aux conditions de leur énonciation.

2. La saillance par une mise en avant explicite lors de l'énonciation : prosodie particulière ; présence d'une pause avant et après la prononciation d'un mot ; erreur de prononciation.

3. La saillance par une construction syntaxique dédiée : détachements en tête (« *le triangle, le rouge, tu dois le mettre à côté du bleu* ») ; constructions clivées (« *c'est ... qui ...* »).

4. La saillance syntaxique liée à l'ordre d'apparition des mots : le début et la fin sont prédisposés pour rendre saillant le mot ou le groupe de mot qui y prend place ; les répétitions et les symétries.

5. La saillance grammaticale, c'est-à-dire liée aux fonctions grammaticales des mots : le sujet est souvent considéré comme le plus saillant (justifiant les constructions passives).

6. La saillance indirecte ou transfert grammatical de saillance : une entité du discours en lien grammatical direct avec l'entité focalisée en devient saillante, dans une moindre mesure.

Des critères liés au sens de l'énoncé. C'est quand nous abordons la sémantique que nous considérons les notions de thème, de focus ou encore de topique :

1. La saillance liée à la sémantique des mots : selon la sémantique du verbe, ce sera l'agent ou le patient qui sera considéré comme le rôle thématique primordial (cf. plus haut).

2. La saillance liée à la sémantique de l'énoncé : selon l'utilisation qui va être faite du calcul de la saillance, ce sera le thème ou le rhème qui sera considéré comme le plus important. Par exemple, un calcul destiné à exploiter une saillance préalable donnera une plus grande importance au rhème, et un calcul destiné à traduire une nouvelle mise en saillance donnera une plus grande importance au thème.

3. La saillance liée à la sémantique de la conversation (qui se construit au cours du dialogue) : compte tenu du lien fort qu'il entretient avec la saillance, le topique est considéré comme le plus important. S'il recouvre plusieurs entités de discours, celle qui est mentionnée le plus souvent sera considérée comme la plus saillante.

4. La saillance indirecte ou transfert sémantique de saillance : un référent très lié au topique est plus saillant qu'un référent qui ne lui est pas apparenté.

Maintenant que ces facteurs sont définis, il s'agit de déterminer les méthodes permettant de quantifier leur intervention dans un système de compréhension automatique.

4 Méthodes pour le calcul numérique de la saillance

Nous présentons ici rapidement les différentes méthodes de calcul utilisées pour confronter des facteurs de saillance et pour quantifier ainsi la saillance d'une entité de discours. L'ordre de présentation va de la méthode la plus facile à mettre en œuvre (celle que nous illustrerons ensuite de manière détaillée sur un exemple) à la plus difficile à mettre en œuvre. L'importance relative des facteurs est déterminée *a priori* dans les premières (méthodes statiques), et en cours de traitement dans les dernières (méthodes dynamiques).

La somme ou la moyenne des facteurs. A partir du moment où l'on dispose d'une liste de facteurs de saillance, la méthode la plus simple consiste à identifier pour chacune des entités du discours quels facteurs privilégient sa saillance, puis de compter ces facteurs, en divisant éventuellement ensuite le total par le nombre de facteurs. Il s'agit de la moyenne arithmétique classique, qui privilégie l'entité caractérisée par le plus grand nombre de facteurs jouant en sa faveur. Cette méthode ne nécessite qu'une liste non hiérarchisée de facteurs. En s'inspirant de la théorie de Tversky qu'ils décrivent, (Iwayama et al., 1990) puis (Pattabhiraman, 1993) utilisent pour leur part la moyenne géométrique, plus précisément la multiplication de deux scores, l'intérêt étant l'influence relative des deux termes du produit. D'une manière générale, la méthode de la moyenne présente plusieurs inconvénients : les facteurs ont tous la même importance or il se peut au contraire qu'un facteur ait beaucoup plus de poids qu'un autre ; en s'appliquant, un facteur peut en annuler un ou plusieurs autres (pénalisant d'autant l'entité auquel il s'applique) ; et il se peut que plusieurs facteurs soient fréquemment conjoints, et que leur prise en compte incrémentale favorise une entité plus qu'il ne l'est souhaitable. Certains de ces inconvénients seront illustrés dans la section suivante.

La prise en compte du facteur optimal. Il s'agit ici de classer *a priori* les facteurs de saillance par ordre d'importance, et de tester leur application sur chacune des entités du discours, en commençant par le facteur le plus important. Dès qu'un facteur s'applique, l'entité correspondante est considérée comme la plus saillante. Autrement dit, l'entité la plus saillante est celle qui satisfait le facteur le plus élevé (ou optimal). C'est une simplification du principe de la Théorie de l'Optimalité (Prince, Smolensky, 1993), qui, bien qu'initialement conçue pour la phonologie, constitue une métathéorie qui nous semble exploitable ici. Cette méthode nécessite d'être capable de fournir une hiérarchie des facteurs. C'est ce que font (Hajičová et al., 1995) lorsqu'ils privilégient par exemple les éléments focalisés dans l'énoncé à ceux désignés par un groupe nominal dans sa partie topique. Le problème avec une telle échelle, et d'une manière générale dès qu'on considère une hiérarchie, c'est qu'une entité peut satisfaire le seul facteur optimal alors qu'une autre entité peut satisfaire une multitude de facteurs secondaires et constituer ainsi un candidat théoriquement plus pertinent.

La moyenne pondérée des facteurs. La pondération de facteurs selon leur importance et la prise en compte de l'ensemble des facteurs par une moyenne s'avère une solution aux problèmes des deux méthodes précédentes. Déterminer des poids s'avère cependant délicat : l'intuition

ne suffit pas à justifier des chiffres tels que 0.8 ou 0.6, et une analyse de corpus peut aboutir à des résultats biaisés de par la nature du corpus ou les difficultés que pose l'identification par l'annotateur des causes de saillance. La pondération de (Alshawi, 1987) est l'une des premières et des plus intéressantes. Si les poids sont critiquables, avec par exemple une trop grande importance donnée à la récence, leur exploitation sur un large éventail de phénomènes linguistiques est en revanche appréciable. La méthode de la moyenne pondérée semble adéquate au calcul de la saillance mais nécessite un énorme travail de détermination des poids des facteurs. Pour être valide, ce travail devrait inclure le test systématique d'un facteur en inhibant tous les autres, puis le test de chaque paires de deux facteurs, etc.

Les méthodes procédurales. Pour résoudre des anaphores pronominales, (Mitkov, 1998) définit des heuristiques basées sur des scores intuitifs. Son approche caractérisée par le peu d'information exploitée ne nécessite même pas d'analyse syntaxique complète. Les facteurs de saillance en sont très réduits (ex : 0 pour un défini et -1 pour un indéfini). Cette approche illustre le recours à une méthode calculatoire extrême, avec son avantage d'être opérationnelle et parfois pertinente, et son inconvénient majeur : aucune plausibilité linguistique.

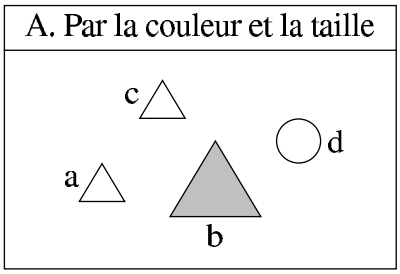
Les méthodes statistiques et les approches hybrides. (Pattabhiraman, 1993) utilise un réseau de relations statistiques entre concepts pour identifier la catégorie la plus saillante dans une situation donnée. Les résultats s'avèrent convaincants et montrent l'intérêt de certaines méthodes statistiques. C'est le cas de l'analyse factorielle des correspondances permettant (avec l'avis d'experts) de déterminer les influences relatives des divers facteurs, par exemple l'influence de la position initiale et de la fonction sujet sur le statut de thème. Cette idée reste cependant à l'état de perspective de recherche. Une autre perspective qui nous semble intéressante est l'exploration de méthodes hybrides, telle que la combinaison d'une méthode basée sur la moyenne pondérée de facteurs avec une méthode statistique, celle-ci remettant en question les différents poids compte tenu des interférences entre facteurs.

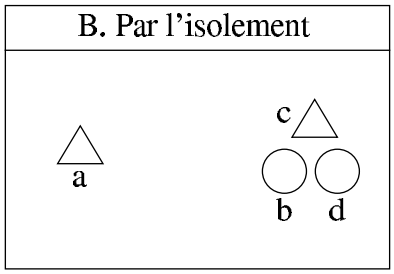
Les méthodes dynamiques. Figer une hiérarchie de facteurs peut sembler dangereux : rien ne dit que dans un contexte ou un autre, tel facteur prendra une importance toute particulière. Exemples : dans une suite de deux phrases où le thème de la seconde reprend le rhème de la première, le facteur lié à la distinction thème-rhème n'a pas le même poids dans les deux phrases ; lorsque le propos du discours est identifié et considéré comme saillant, ce calcul doit rester activé pour l'analyse de plusieurs phrases. Un autre exemple où l'importance des facteurs de saillance est gérée dynamiquement est celui de (Lappin, Leass, 1994), qui se basent sur (Alshawi, 1987) pour proposer un algorithme de résolution des anaphores pronominales. Un seuil initial varie au fur et à mesure du traitement des phrases. Certaines étapes consistent par exemple à diviser par 2 le poids de tel facteur. Même si cette approche présente quelques défauts (récence privilégiée, limitation à des facteurs purement formels) et reste à améliorer, elle constitue une première étape dans la gestion dynamique de scores qui nous semble théoriquement intéressante. Cette méthode est cependant très difficile à mettre en œuvre : il s'agit d'identifier l'ensemble des influences contextuelles sur chacun des facteurs de saillance. Face aux difficultés, les auteurs avouent eux-mêmes que les poids qu'ils proposent sont arbitraires. Le principal constat que nous pouvons faire à ce stade est ainsi le suivant : si les méthodologies mathématiques et statistiques semblent stabilisées, il en n'est pas de même de celles relatives au traitement de corpus pour déterminer les poids et les influences relatives des divers facteurs. C'est dans ce sens qu'il nous semble important de continuer les recherches. Et c'est aussi pour cette raison que nous ne présentons pas ici d'analyse de corpus.

5 Illustration de la moyenne des facteurs sur un exemple

Pour le moment, plusieurs méthodes peuvent être mises en œuvre sans trop de difficultés, et il nous semble utile de détailler l'une d'entre elles sur des exemples faisant intervenir le principe de primordialité et le principe de singularité. Nous retenons ainsi un exemple d'une scène visuelle dans laquelle les objets se distinguent par leur forme, leur couleur et leur taille. A ces propriétés physiques qui constituent autant de facteurs de saillance s'ajoute l'isolement que nous avons défini plus haut comme un autre facteur. Comme le montre la figure 1, des scores de 0 ou de 1 sont attribués à chacun des objets pour chacun des facteurs de saillance. C'est dans la manière d'attribuer ces scores que se traduisent les principes de primordialité et de singularité :

- Application du principe de primordialité : l'objet qui s'avère le plus important compte tenu du facteur de saillance considéré obtient un score de 1 pour ce facteur, et les autres objets un score de 0. C'est ce qui est fait avec le facteur taille dans la figure 1-A, l'objet b ayant la taille la plus importante.
- Application du principe de singularité : compte tenu d'un facteur de saillance et des valeurs qui peuvent être prises pour ce facteur, un objet ou une entité du discours qui partage sa valeur avec d'autres obtient un score de 0 pour ce facteur ; un objet qui est le seul à posséder sa valeur obtient un score de 1. C'est ce qui est fait avec le facteur forme, les valeurs instanciées étant « triangle » et « cercle ». Ainsi, l'unique cercle de la figure 1-A obtient le score de 1, alors que tous les objets de la figure 1-B restent à 0.





	a	b	c	d
forme :	0	0	0	1
couleur :	0	1	0	0
taille :	0	1	0	0
isolement :	0	0	0	0
saillance :	**0**	**0.5**	**0**	**0.25**

	a	b	c	d
forme :	0	0	0	0
couleur :	0	0	0	0
taille :	0	0	0	0
isolement :	1	0	0	0
saillance :	**0.25**	**0**	**0**	**0**

Figure 1: Scores numériques pour un premier calcul de la saillance visuelle.

Cette méthode nous semble opérationnelle, dans la mesure où repérer des singularités et des primordialités dans une liste de propriétés s'avère possible d'un point de vue computationnel, du moins à partir du moment où ces propriétés peuvent se représenter formellement.

Plus précisément à propos de la figure 2 : les deux cas présentés ont la même saillance alors que le second est pourtant intuitivement le plus saillant, du fait du présentatif. En effet, certains facteurs se contredisent, particulièrement la construction dédiée et la fonction grammaticale sujet : le cas B se voit attribuer un point de plus que le cas A pour la construction dédiée, mais du coup perd un point pour la fonction sujet. Or justement la construction dédiée prend le pas

A.	*« le triangle rouge se met à côté du bleu »*	
	« le triangle rouge »	*« le bleu »*
construction syntaxique dédiée :	0	0
ordre d'apparition des mots (début) :	1	0
fonction grammaticale (sujet) :	1	0
sémantique des mots (rôle thématique) :	1	0
sémantique de l'énoncé (thème) :	1	0
total de saillance entre 0 et 1 :	**0.8**	**0**

B.	*« c'est le triangle rouge que tu dois mettre à côté du bleu »*	
	« le triangle rouge »	*« le bleu »*
construction syntaxique dédiée :	1	0
ordre d'apparition des mots (début) :	1	0
fonction grammaticale (sujet) :	0	0
sémantique des mots (rôle thématique) :	1	0
sémantique de l'énoncé (thème) :	1	0
total de saillance entre 0 et 1 :	**0.8**	**0**

Figure 2: Scores numériques pour un premier calcul de la saillance linguistique.

sur la fonction sujet et celle-ci ne devrait plus être comptée, ou du moins devrait être comptée dans une proportion moindre. On arrive ainsi à la notion de poids : comme dans les approches d'Alshawi et de Lappin et Leass, les poids relatifs des différents facteurs sont d'importance. Reste que leur détermination est empirique et nécessite de coûteuses études de corpus.

6 Conclusion et perspectives

La saillance ne fait pas partie du message communiqué, mais tout le message se base sur elle, s'explique par elle, se structure en fonction d'elle. C'est pourquoi cette notion nous semble avoir une importance toute particulière dans les recherches en traitement automatique des langues. Nous avons voulu montrer ici que la notion de saillance linguistique pouvait être appréhendée d'une manière efficace en clarifiant ses rapports avec la structure informationnelle et en la rapprochant de la saillance visuelle. La liste des facteurs de saillance à laquelle nous aboutissons nous permet d'envisager comment la saillance linguistique aussi bien que la saillance visuelle peuvent être exploitées dans un système de compréhension automatique, qu'il s'agisse de dialogue homme-machine, de traduction ou de résumé automatique. Les principes de primordialité et de singularité que nous définissons et que nous appliquons à la méthode de la moyenne des facteurs constituent un premier pas vers une formalisation de la saillance. Cette formalisation reste à compléter avec l'apport des autres méthodes présentées dans cet article, avec par exemple des stratégies supplémentaires pour tenir compte des interactions entre facteurs de saillance. Nous aurons en particulier à spécifier des stratégies de confrontation de la saillance visuelle avec la saillance linguistique pour la résolution des références en contexte visuel.

Références

Alshawi H. (1987), *Memory and Context for Language Interpretation*, Cambridge University Press.

Caron J. (1989), *Précis de psycholinguistique*, Paris, PUF.

Edmonds P.G. (1993), *A Computational Model of Collaboration on Reference in Direction-Giving Dialogues*, Ms. Thesis, University of Toronto, Canada.

Grosz B.J., Joshi A.K., Weinstein S. (1995), Centering: A Framework for Modelling the Local Coherence of Discourse, *Computational Linguistics*, Vol. 21(2).

Guillaume P. (1979), *La psychologie de la forme*, Paris, Flammarion.

Hajičová E., Hoskovec T., Sgall P. (1995), Discourse Modelling Based on Hierarchy of Salience, *Prague Bulletin of Mathematical Linguistics*, Vol. 64.

Isard S. (1975), Changing the Context, In: Keenan E.L. (Ed.) *Formal Semantics of Natural Language*, London & New York, Cambridge University Press.

Iwayama M., Tokunaga T., Tanaka H. (1990), A Method for Calculating the Measure of Salience in Understanding Metaphors, In: *Proceedings of the National Conference on Artificial Intelligence*, Boston.

Kievit L., Piwek P., Beun R.J., Bunt H. (2001), Multimodal Cooperative Resolution of Referential Expressions in the DENK System, In: Bunt H., Beun R.J. (Eds.), *Cooperative Multimodal Communication*, Berlin & Heidelberg, Springer.

Krahmer E., Theune M. (2002), Efficient Context-Sensitive Generation of Referring Expressions, In: van Deemter K., Kibble R. (Eds.), *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*, Stanford, CSLI Publications.

Landragin F. (2004), Saillance physique et saillance cognitive, *Cognition, Représentation, Langage (CORELA)*, Vol. 2(2).

Lappin S., Leass H.J. (1994), A Syntactically Based Algorithm for Pronominal Anaphora Resolution, *Computational Linguistics*, Vol. 20(4).

Loftus G.R., Mackworth N.H. (1978), Cognitive Determinants of Fixation Location during Picture Viewing, *Journal of Experimental Psychology: Human Perception and performance*, Vol. 4.

Mitkov R. (1998), Robust Pronoun Resolution with Limited Knowledge, In: *Proceedings of the Eighteenth International Conference on Computational Linguistics*, Montréal.

Osgood C.E., Bock J.K. (1977), Salience and Sentencing: Some Production Principles, In: Rosenberg S. (Ed.), *Sentence Production: Developments in Research and Theory*, Hillsdale, Erlbaum.

Pattabhiraman T. (1993), Aspects of Salience in Natural Language Generation, Ph.D. Thesis, Simon Fraser University.

Pearson J., Poesio M., Stevenson R. (2001), The Effects of Animacy, Thematic Role and Surface Position on the Focusing of Entities in Discourse, In: *Proceedings of the First Workshop on Cognitively Plausible Models of Semantic Processing*.

Prince A., Smolensky P. (1993), Optimality Theory: Constraint Interaction in Generative Grammar, Technical Report, Rutgers University.

Stevenson R.J., Crawley R.A., Kleinman D. (1994), Thematic Roles, Focus and the Representation of Events, *Language and Cognitive Processes*, Vol. 9(4).

Stevenson R.J. (2002), The Role of Salience in the Production of Referring Expressions, In: van Deemter K., Kibble R. (Eds.), *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*, Stanford, CSLI Publications.

Topiques dialogiques

Anne Xuereb, Jean Caelen

Laboratoire CLIPS-IMAG – UJF/CNRS/INPG
Domaine universitaire, BP 53, 38041 Grenoble Cedex 9
{Anne.Xuereb, Jean.Caelen}@imag.fr

Mots-clés : Interprétation pragmatique, dialogue homme-machine
Keywords: Pragmatic analysis, man-machine dialogue

Résumé

Nous présentons dans cet article une extension de la SDRT (*Segmented Discourse Representation Theory*), pour un modèle d'interprétation pragmatique d'un système de dialogue homme-machine. Partant d'une discussion sur les présupposés et les implicatures conversationnelles, nous analysons l'approche de Ducrot en vue d'une intégration des topoï dans notre modèle. Nous y ajoutons la prise en compte des attentes dans le dialogue (effets projectifs des actes interlocutoires). Enfin nous proposons un mécanisme de résolution logique qui consiste à introduire plus systématiquement un nœud topique dans la SDRS (*Discourse Representation Structure*). Nous décrivons dans cet article les principes de traitement pragmatique mis en œuvre, et nous illustrons le processus d'analyse à l'aide d'un exemple.

Abstract

We present in this paper an extension of the SDRT model for the pragmatic interpretation in a man-machine dialogue system. After a discussion on presupposition and implicature phenomena we consider the Ducrot's approach based on *topos* concept. We consider also the point of view in which a speech act is an "expectation" of some result in the future, in a kind of projective effect. Then we describe a logical process including systematically a "topic node" in the SDRS (*Discourse Representation Structure*), subsuming the rhetoric relations having implications or hypothetic implications between utterances. Our paper focuses on the pragmatic processing principles for resolving implications in the dialogue. This is illustrated by an example.

1 Introduction

Le cadre de travail de cet article est le projet PVE, Portail Vocal d'Entreprise, supporté par le programme RNRT[1]. Il fait suite à nos travaux sur l'interprétation pragmatique en dialogue homme-machine présentés à TALN'04. Dans le présent article, notre propos est d'affiner la notion de topique et de l'intégrer plus complètement dans l'analyse et la représentation de dialogues dans la SDRT. La SDRT (*Segmented Discourse Representation Theory*) (Asher, Lascarides, 2003) est une formalisation de l'interprétation dynamique de l'énoncé en contexte. Elle est sujette à de nombreux travaux théoriques et de confrontation au terrain. On peut citer pour le dialogue des travaux comme (Muller et al., 2002), (Prévot, 2004).

La question du topique en SDRT classique est traitée en terme de relation rhétorique. Nous proposons de considérer ici le nœud topique non seulement comme élément structurant du dialogue mais également comme le réceptacle de représentations pragmatiques en cours de calcul. Nous décrivons dans cet article cette « extension » de la SDRT qui consiste à introduire plus systématiquement un nœud topique dans la SDRS (*Discourse Representation Structure*), en subsumant toute relation rhétorique pouvant contenir des présuppositions ou des implicatures conversationnelles ou des effets projectifs dans les énoncés liés.

2 Sémantique et pragmatique du dialogue

(Récanati, 2001) affirme que : « Une thèse centrale, et même fondatrice, de la sémantique contemporaine est que la signification d'une phrase détermine ses conditions de vérité. Cette détermination peut être plus ou moins directe. Elle est relative au contexte lorsque la phrase est indexicale : la signification est alors conçue comme une "fonction", appariant contextes et conditions de vérité. Ainsi la phrase "Je suis français", énoncée par Jean, est vraie si et ssi Jean est français. Dans les autres cas (non indexicaux), la signification de la phrase détermine directement ses conditions de vérité, en vertu de sa seule signification linguistique : la phrase "la neige est blanche" aurait ainsi la propriété d'être vraie si et ssi la neige est blanche. Searle soutient que la signification linguistique *sous-détermine* radicalement les conditions de vérité, même après que la valeur des expressions indexicales contenues dans la phrase ait été fixée. Etant donné une phrase quelconque (indexicale ou non), il n'est pas possible de spécifier un état de choses E tel que la phrase soit vraie si et ssi E est réalisé. Searle montre cela d'une façon tout à fait convaincante. Ses exemples établissent que l'on peut toujours imaginer un contexte où la phrase en question ne serait pas considérée comme vraie, quand bien même l'état de choses E serait réalisé, ils montrent aussi que la signification linguistique sous-détermine les conditions de vérité, *quelle que soit la phrase énoncée.*» Par exemple l'énoncé « Le bateau de Jean » n'indique pas le type de relation entre le bateau et Jean : possession, fabrication par Jean, rêve de Jean ? Il n'aura de conditions de vérité déterminées que si une relation particulière entre Jean et le bateau a été spécifiée, mais la spécification en question n'obéit à aucune règle ou procédure, elle apparaît au cours du dialogue de manière explicite ou peut même rester implicite entre les conversants qui en ont une connaissance commune. Ainsi au-delà du contexte et de l'arrière-plan, la situation dialogique participe également de la

[1] RNRT, Réseau National de Recherche en Télécommunication du ministère de la recherche français. Notre but dans le projet PVE est de concevoir et de réaliser une application complète de dialogue homme-machine jouant le rôle d'une assistante virtuelle. Les usagers dialoguent en langue naturelle avec un "agent conversationnel" pour faire réaliser les tâches courantes dans l'entreprise comme la prise de rendez-vous, l'organisation de réunions, la gestion d'agenda, etc. Il s'agit de dialogues finalisés, dans un univers limité.

négociation du sens (ou co-construction du sens). Par exemple, au cours du dialogue, le « bateau de Jean » peut prendre les connotations C1 ou C2 :

A : Le bateau de Jean est finalement resté à quai
B : Tu veux dire qu'il ne l'a jamais utilisé ?
C1 : Oui c'est resté un pur rêve / C2 : Oui, il ne l'a jamais terminé

2.1 Présuppositions et implicatures

A ces questions d'interprétation pragmatique sont aussi liés les problèmes de présupposition et d'implicature, bien connus.

(a) Les **présuppositions** sont généralement considérées comme des restrictions dans un domaine de définition donné (Corblin, 2003). Elles peuvent être marquées lexicalement, par exemple pour le verbe boire, *boire(x)* présuppose généralement *liquide(x),* ce qui restreint *x* à appartenir à l'ensemble de définition de la fonction *liquide(x)*, mais dans le cas des descriptions définies comme *le roi de France est chauve*, la contrainte porte sur l'existence du sujet $\exists x$: *roi_de_France(x)*. Les présuppositions correspondent à des engagements implicites des conversants qui partagent des connaissances communes, ce sont des pré-propositions. L'engagement du locuteur est marqué dans des expressions comme, *je regrette que p*, qui présuppose que p est vrai et sur lequel il prend parti. (Hambling, 1970) traite le problème de la présupposition à travers le tableau des engagements des locuteurs, qui est mis à jour au cours du dialogue. Il fait l'hypothèse que si un fait présupposé n'est pas (rapidement) contesté alors il est admis (en utilisant le principe de coopération de Grice). Dans ce cadre les présuppositions non marquées lexicalement restent difficiles à traiter. Par exemple dans *J'ai peu travaillé mon examen*, on peut présupposer que si le locuteur sait qu'il est le meilleur de la classe, il considère que cette épreuve est facile pour lui, tandis que s'il a des difficultés scolaires il considère cette épreuve comme insurmontable et qu'il est inutile dès lors de travailler. Ce type de présupposition fait intervenir les croyances des conversants sur la situation (le locuteur sait aussi que ses auditeurs savent), problème qui dépasse donc le cadre strict de la sémantique.

(b) Les **implicatures** sont des résultats d'inférence qu'un auditeur est susceptible de faire à partir d'un énoncé. Ce sont des post-propositions (contrairement aux présuppositions vues précédemment). On distingue les implicatures conscientes (ou intentionnelles), des implicatures inconscientes – appelées parfois *implicitures*. Les implicatures sont calculées à partir de ce qui est dit ou de ce qui est implicité conventionnellement. Pour (Grice, 1975) les implicatures - dites conversationnelles - proviennent du principe de coopérativité dans lequel ce qui est dit est pertinent (principe d'économie du dire). Cette notion d'implicature a depuis été généralisée (ou critiquée) par de nombreux auteurs, particulièrement autour des implicatures conventionnelles : certaines expressions contiennent en elles-mêmes des implications pragmatiques non détachables et non défaisables. Par exemple : « je suis garé derrière » contient dans la lexie *garé* l'implication d'un véhicule rangé dans un lieu adéquat. Il y a ici transfert métonymique entre *je* et *véhicule*. Dans ce cas les implications se construisent de manière montante et procèdent du niveau sémantique uniquement à partir de segments déclencheurs de l'énoncé (contrairement aux implicatures conversationnelles qui sont descendantes, contextuelles, globales et souvent conscientes). Pour résoudre ces

problèmes, (Recanati, 2003) propose deux mécanismes de résolution : un premier à l'aide d'un processus de projection sémantique ascendant (incluant le transfert) puis un second à l'aide d'un processus de saturation des indexicaux et de déduction pragmatique descendant et global. Ces deux processus n'opèrent pas dans un ordre chronologique ou hiérarchique mais opèrent de manière conjointe.

2.2 Les effets projectifs du dialogue et la SDRT

La SDRT utilise la notion de relation rhétorique pour structurer le discours. Cette notion se fonde quelque peu sur la notion de paire adjacente issue des théories de la conversation (Goffman, 1967) dans laquelle tout acte de langage tente de « fermer » une paire ouverte. Appliquée au dialogue, cette vision le limite à un système de résolution des attentes. Il nous semble cependant que dans la perspective d'un modèle projectif (Vernant, 1997) ou dans celle de la logique interlocutoire (Trognon, 1995), chaque acte est projeté vers le futur et prend sa signification dans un *interacte* construit de manière émergeante par les acteurs du dialogue. Il s'agit donc plutôt de « projeter » le dialogue en avant à chaque instant, chaque acteur prenant sa part dans l'action mais aussi en en déléguant une partie à autrui. La contrôle des effets de ces actions devient un moteur pour la poursuite du dialogue et la coordination mutuelle. Par exemple, dans la situation suivante où un homme A aborde une jeune fille B dans la rue à minuit,

A : Avez-vous l'heure ?
B : Non

il est évident que la réponse de B contient plus qu'une simple réponse à la question précédente, il contient aussi le projet « laissez-moi tranquille » de B. Les effets de ce « non » portent non seulement sur la fermeture de la question de A, mais il pose également un nouveau but potentiel (B espère faire partager ce but à A). Ce but sera peut-être repris par A dans le tour suivant, il deviendra alors un but conjoint. Cette potentialité n'est pas modélisée entièrement par la SDRT classique.

2.3 Topos chez Ducrot

Pour Ducrot, l'argumentation (qui, pour lui, structure le texte ou le discours) repose sur la synthèse de trois composants: le *topique*, le *logique* et l'*encyclopédique*. Ces trois éléments ne sont pas toujours facilement séparables. Pour Ducrot et Anscombre, le *topos* est « le garant qui autorise le passage de l'argument A à la conclusion C » (Ducrot et al. 1995: 85). C'est un principe général sous-jacent à un enchaînement argumentatif présenté dans un discours. Le *topique* est l'ensemble des topoï ou arguments qui structurent le discours. Les topoï sont des croyances communes qui induisent des conséquents mis sous forme de prédicats, ils contiennent les règles ou les principes d'inférences qui permettent, à partir d'un ou de plusieurs faits singuliers et d'une hypothèse générique sur la réalité, de conclure à l'existence d'un autre fait singulier. Ducrot distingue ainsi les implications contextuelles qui prennent essentiellement deux statuts : celui de *conclusion impliquée* ou celui d'*hypothèse anticipatoire*. Il y a donc deux plans d'inférence dans le composant *logique* : le plan *conclusif* et le plan *constructif*. Le composant *encyclopédique* quant à lui est indissociable du topique et du logique. L'encyclopédique spécifie la connaissance du monde, le savoir référentiel, culturel, partagé par le locuteur et son allocutaire. En quelque sorte le topos chez Ducrot généralise l'implicature (cela devient l'ensemble des implications que l'on peut faire à partir des croyances mutuelles) mais aussi le présupposé (cela devient l'ensemble des hypothèses

que l'on peut poser *a priori*). Ainsi, par exemple, dire : *Pierre a travaillé toute la journée*, c'est produire le topos $\exists x : Pierre(x) \wedge fatigué(x)$. Le sens du verbe travailler produit un faisceau de topoï auquel se tissent les arguments et se construit le discours.

Cette approche nous paraît intéressante pour compléter la notion d'interacte au plan sémantique en ce sens qu'elle offre une voie pour traiter les hypothèses anticipatoires dans le dialogue et donne un cadre précis à la prise en compte de la dimension pragmatique de l'encyclopédie que nous restreindrons à l'ontologie de l'application dans la suite (à cause du cadre restreint des mondes dans le dialogue homme-machine).

2.4 Discussion et position du problème

La SDRT est un cadre théorique fécond, qu'il y a lieu de compléter pour le dialogue, par certaines représentations ou mécanismes de résolution pour tenir compte de tous les phénomènes discutés ci-dessus, à un niveau adéquat d'articulation entre sémantique et pragmatique. Le *moment* du dialogue est notamment un facteur spécifique à prendre en compte en plus du contexte et de l'arrière-plan des connaissances. Par exemple :

E1 : (Allô) / Je suis Paul Dupont

E1 dans ce cas se présente spontanément pour se faire identifier. Il a l'intention de demander un service. On est en ouverture de dialogue. Les présupposés sont ici que celui qui parle est bien une personne et celui qui se nomme, les implicatures sont qu'il se présente et qu'il est connu de son interlocuteur et le topos est qu'il est membre de droit d'un certain service qu'il va certainement demander (effet projectif).

E2 : Je suis Paul Dupont / (vous savez bien)

E2 dans ce cas se présente pour affirmer ou confirmer son identité ou lever un doute. On est dans une phase de négociation et non plus dans l'ouverture, l'énoncé prend ici valeur d'argument. Les présupposés sont donc que les droits du locuteur sont peut-être mis en doute par son interlocuteur, les implicatures sont que la personne s'affirme sincère et le topos est un argument pour revendiquer ces droits, l'effet projectif est de conclure la négociation.

Ces deux exemples montrent à l'évidence que $F^S_U(nom(U))$ = *je suis Paul Dupont* prend un sens différent selon le contexte dialogique. De $F^S_U(nom(U))$/Ouverture on peut déduire que U se présente et de $F^S_U(nom(U))$/Négociation que U prouve son identité. Cela montre que la représentation du dialogue doit contenir d'autres informations pragmatiques que celles venant strictement des énoncés, de leurs présupposés et des implicatures, ou des topoï de Ducrot, mais aussi du contexte dialogique lui-même. C'est ce que nous nous proposons de modéliser à travers le concept de topique dans la SDRT.

3 Topique et thème

Dans ce paragraphe nous décrivons la mise en œuvre à travers la SDRT des requis énoncés ci-avant.

3.1 La représentation logico-sémantique

Le composant de compréhension sémantique de notre système de dialogue homme-machine fournit une représentation de l'énoncé sous forme logique : une DRS (Xuereb et al., 2004). Un

énoncé est modélisé par une conjonction d'actes de langage, chaque acte étant de la forme Fp, où F est la force illocutoire et p le contenu propositionnel (Vanderveken, 1985).

- La force illocutoire F est de type F^A faire une action sur le monde (acte déclaratif) ; F^S faire savoir une information (asserter) ; F^{FS} faire faire savoir une information (poser une question) ; F^F faire faire une action (ordonner) ; F^D faire devoir (donner une obligation) ; F^P faire pouvoir (proposer un choix).
- Le contenu propositionnel est représenté sous forme logique : il comprend une liste de marqueurs de référence (des variables typées sémantiquement), des prédicats et des équations mettant en jeu ces marqueurs de référence.

3.2 Le rôle structurant du topique

Le topique peut être vu sous trois aspects :

(a) Une relation structurelle et un constituant discursif qui ont pour rôle de rassembler l'information sous-jacente :

- Pour les relations subordonnantes, le constituant subordonné est le topique de la relation. Par exemple les Elaborations (question constructive $Elab_q$, précision, clarification, explication) mettent en jeu une relation partie/tout entre les constituants principaux de chaque segment. Ces relations introduisent un topique subordonné, qui une fois résolu monte dans le topique dominant,
- Les relations coordonnantes (comme C = Continuation par exemple) introduisent un topique subsumant les constituants reliés : c'est un nouveau constituant composé.

(b) Un nœud de résolution de la structure SDRS : Le constituant topique est le siège des résolutions des anaphores pronominales et associatives (ellipses). Les résolutions sont effectuées après inférence des relations de discours. Lors de la phase de mise à jour de la structure, l'ensemble des référents et prédicats établis dans la sous-structure sous-jacente (après résolution des anaphores, et prise en compte des présuppositions) remonte dans le topique. Au cours du dialogue, la SDRS globale se constitue ainsi par établissement progressif de topiques de niveau de plus en plus élevé (union des éléments coordonnés, ou remontée des éléments subordonnés), jusqu'au topique dominant, constitué de l'ensemble de l'information établie par les participants. Les présuppositions sont intégrées sous la forme de relations de discours ajoutées au contexte. C'est dans le nœud topique que se font les corrections éventuelles (remises en causes, corrections, effacements, etc.)

(c) Une unité logique de contenu du savoir partagé élaboré au cours du dialogue : L'avancement du dialogue se modélise ainsi par la construction d'une structure arborescente de topiques. Chaque topique subsumant les topiques sous-jacents. A la fin, l'ensemble du savoir est rassemblé dans le topique sommet : c'est l'ensemble des référents et prédicats établis. Sur cette structure logique des topiques on peut venir lier la structure logique de la tâche donnée par l'arbre des thèmes (le niveau logique de Ducrot). Le niveau encyclopédique (ici réduit au monde de l'application) est représenté par une ontologie qui vient couvrir le modèle de la tâche. Grâce à l'ontologie, on résout les présuppositions et les implicatures et on pose les anticipations : ce sont des connaissances extérieures au dialogue qui sont alors

intégrées dans le nœud topique (ajout de prédicats et relations). Poser des « attentes » à ce niveau permettra ensuite, au tour suivant, de résoudre les attachements ambigus.

3.3 Topique et arbre de thèmes

L'introduction d'un nouveau nœud topique est déclenchée par la détection d'un changement de thème. Les thèmes sont ceux du domaine, ils sont organisés sous forme d'arbre. Ils sont eux-mêmes liés au modèle de tâche via l'ontologie dont nous ne parlerons pas dans cet article.

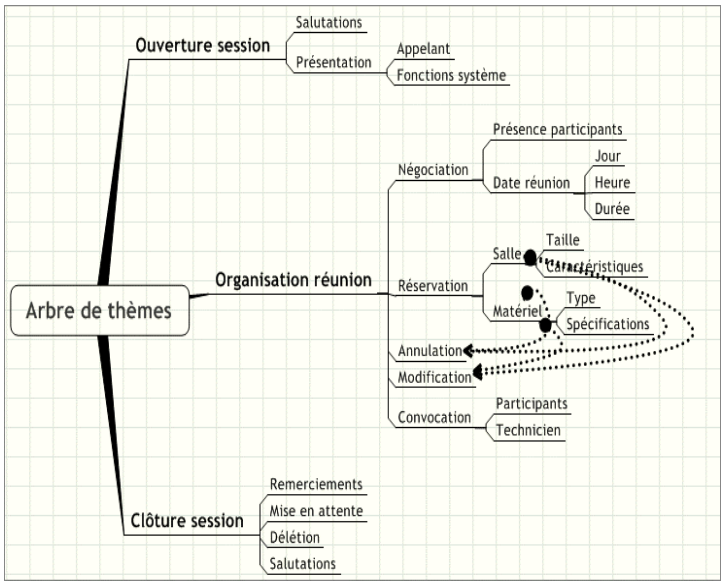


Figure 1: Arbre de thèmes pour la tâche « organisation de réunion » dans l'application PVE.

3.4 Les résolutions déclenchées par la structure logique du dialogue

La modélisation de la logique du dialogue par la SDRT met en évidence les contraintes d'accessibilité pour la résolution des sous-spécifications : les SDRS π_1, π_n étant reliées par une relation $R(\pi_1, \pi_n)$, π_n et ses sous-DRS accèdent aux référents (DRS-accessibles) de π_1. Par exemple, dans les paires Question-Réponse complètes (QAP), indirectes (IQAP) et partielles (PQAP), les anaphores propositionnelles des segments Réponse sont résolues en accédant au segment Question. Le résultat de l'application du segment Réponse sur le segment Question est stocké dans le nœud topique immédiatement dominant.
Les résolutions des anaphores et ellipses se font par unification avec un référent défini et accessible : on accède prioritairement au segment immédiatement dominant, en cas d'échec on accède au nœud supérieur. D'autre part, chaque relation rhétorique porte une sémantique spécifique qui enrichit la sémantique des segments eux-mêmes : le contexte est mise à jour en tenant compte à la fois du contenu propositionnel des segments reliés, et de la sémantique propre de la relation rhétorique.

4 Résultats : exemple d'analyse

Dans l'extrait de dialogue homme-machine réel présenté ci-après, U désigne l'utilisateur, et M l'agent conversationnel (machine). Nous illustrons succinctement quelques mécanismes d'interprétation pragmatique représentatifs.

U : Luc Blanc à l'appareil.	π_1
Est-ce que la salle Lafayette est disponible demain ?	π_2

M : Non. Elle est disponible jeudi	π_3, π_4
U : Bon eh bien réservez-la moi	π_5
M : voulez-vous un technicien pour le rétro-projecteur ?	π_6

$\pi_{1\,U}$: [F^S ; a1: personne ; Identité+annonce(a1) ; a1.NomComplet = "Luc Blanc"]/Ouverture

1. L'annonce de l'identité en ouverture de dialogue active la relation arrière-plan qui permet de résoudre le référent a1 de type *personne* qui est présupposé être U,
2. Présupposition : Luc Blanc = Prénom + Nom, a1 = U
3. Implicature : a1: membre_connu(a1) est un « acteur » connu du système
4. Effet projectif : le locuteur en se présentant se met en attente d'un service. S: service ; demander_service(a1, S) dans le topos ayant _droit(a1, S)

$\pi_{2\,U}$: [F^{FS} ; s2: salle, d2: date, e2: booléen_dispo ; Agenda+demande (s2, d2, e2); s2 =Lafayette ; d2 = jour + 1 ; e2 = 0]

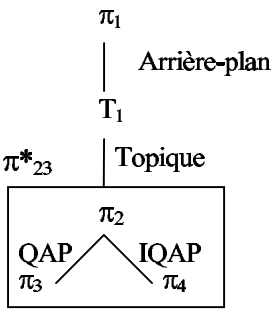
1. U pose une question sur la disponibilité de la salle qui active le prédicat Agenda+demande et positionne e2 (e2=0 signifie indisponible)
2. Présupposition : existence de la salle Lafayette.
3. Effet projectif : le service probablement demandé est la réservation de salle et ses dérivés (réserver du matériel par exemple et un technicien en début de réunion)).

π_{3M} : [neg(x) x = ?, prop(x)] se résout par x = π_2 ;

π_{4M} : [F^S ; v: indéfini, d3: date, e3: booléen_dispo ; Agenda+annonce(v, d3, e3) ; v = ? e3 = 0, d3 = *plus proche* jeudi]

1. Présupposition : jeudi = jeudi prochain.
2. Implicature : salle Lafayette non disponible entre demain et jeudi prochain ; topos : être disponible est une condition préalable à une réservation.
3. Effet projectif : réserver à partir de jeudi, Agenda+reserver (s2, d2, e2) ; s2 = Lafayette ; d2 ≥ jeudi ; e2 = 1

On obtient après π_4 le graphe suivant :



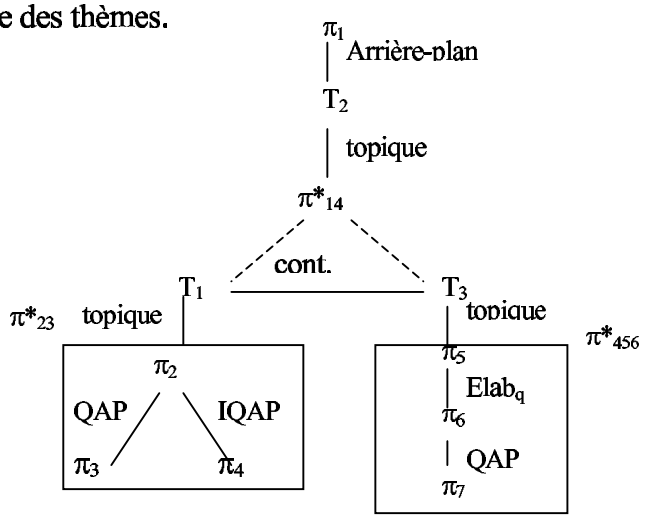
1. Le topique T1 contient les prédicats et référents associés aux résolutions (QAP(π_2, π_3) + IQAP (π_2, π_4) , ainsi que les présuppositions, implicatures, et effet projectif.
2. Soit [s: salle ; s.nom = Lafayette ; p: personne ; personne.NomPrenom = LucBlanc ; ayant-droit-service(p) ; non-dispo(s, demain) ; non-dispo(s, demain –jeudi-1) ; dispo(s, jeudi prochain)]
3. Topos + effet projectif : attente de demande de réservation. [S : service = réserver_salle ; demander_service(p, S)]. Tous les actes à potentiel ouvrant (F^{FS}) sont saturés.

4. IQAP(π2 , π4) : La séquence des deux actes F^{FS}+F^{S} oriente vers une paire question+réponse. Le principe de maximisation de la cohérence déclenche l'inférence que *jeudi* n'est pas *demain* mais une date ultérieure : c'est une relation IQAP avec une implicature que « demain elle n'est pas disponible » mais qu'elle le sera tous les jours à partir de jeudi prochain inclus. L'effet projectif est que le service demandé sera une réservation de salle car le topos porte sur la réservation potentielle.

$\pi_{5\,U}$: [F^{FS} ; v :indéfini, d5 :date ; réservation+demande (v, d5) v = ? , d5= ?] ;

1. π_5 est un acte à potentiel ouvrant (F^{F}) et il pose un nouveau thème (*réservation*). π_4 se lie au topique T1 par Continuation ; l'insertion de cette relation déclenche alors l'introduction du topique T_2 dominant le constituant complexe π^*_{14} formé par C(T_1, π_4). Le processus de résolution des anaphores, par accès à π_4 et application des contraintes sur les types sémantiques fournit v : salle ; v = s.
2. Par la prise en compte du topos présent dans T_1 (attente de réservation) et de l'hypothèse de cohérence (si demande de réservation alors elle est liée logiquement à la disponibilité préalablement énoncée), on déduit d = date de disponibilité de s ; d$\in T_1$ d'où d = jeudi prochain (il est à noter que sans la prise en compte du topos le système aurait décelé l'absence du paramètre date-reservation et aurait posé la question « pour quelle date ? »).

π_{6M} : s'explique par le topos lié à réservation de salle qui entraîne celui de gestion de matériel associé au prédicat réserver_salle. Ce qui illustre ici le lien entre la structure pragmatique (SDRT) et l'arbre des thèmes.



Le graphe final contient un topique principal T_2 constitué par les coordinations des topiques T_1, T_3. Il contient la réunion des référents et prédicats établis dans les deux topiques qu'il domine, T_1 et T_3. Tous les référents sont entièrement définis (salle Lafayette, Luc Blanc) et les prédicats les mettant en jeu sont agenda+annonce, reservation+demande, technicien+demande). Les implicites ont pris le statut de référent accessible (réserver jeudi prochain). Ce nœud constitue le contexte structuré à cet instant du dialogue. Si l'acte suivant vient remettre en cause un des éléments (par exemple « non, je parlais de jeudi en quinze ») alors on insère une relation Correction, et le référent (date de réservation) est mis à jour dans T2, sans modifier la structure sous-jacente. L'élaboration incrémentale de la structure de topiques (un topique dominant subsume les topiques sous-jacents) a permis la prise en compte sur le plan logique de certains implicites du dialogue.

5 Conclusion

Nous proposons un modèle étendu de la SRDT dans le cadre spécifique du dialogue homme-machine finalisé en introduisant systématiquement un nœud topique dans la SDRS globale qui prend en compte le contexte commun aux deux interlocuteurs évoluant au cours du dialogue (ce qui a été dit, mais aussi ce qui est projeté par anticipation). Pour cela nous avons considéré un cadre plus large que celui des présuppositions et des implicatures, en introduisant les effets projectifs des actes de dialogue. Nous nous sommes inspirés également de la notion de topos et nous avons validé manuellement ce modèle sur l'ontologie[2] que nous avons constituée dans le cadre d'un service de portail vocal d'entreprise (projet PVE, RNRT). Nous avons aussi spécifié un prototype informatisé : le moteur de l'interpréteur utilise un raisonnement hypothétique. Pour chaque tour de parole, les sites d'attachement disponibles de la DRS courante sont calculés ainsi qu'une hypothèse de relation pour chaque nœud encore non étiqueté. Les inférences sont déclenchées sur la base des hypothèses pour tenter une résolution. Une hypothèse est acceptée ou refusée suivant le succès ou l'échec de cette résolution. Une hypothèse acceptée ne sera alors plus réévaluée au tour suivant.

Références

ASHER N., LASCARIDES A. (2003), *Logics of Conversation*. Cambridge University Press.

CORBLIN F. (2003), Presuppositions and commitment stores. in Proceedings *Diabruck, 7th Workshop on the Semantics and the Pragmatics of Dialogue, Wallerfangen.*

DUCROT, O., 1984, *Le Dire et le Dit*. Paris, Minuit.

GOFFMAN E. (1967), *Interaction Ritual : Essays on face-to-face Behavior*. Anchor Books, NY.

GRICE, H.P. (1975), Logic and conversation. P. Cole and J. Morgan, eds., *Syntax and Semantics*, vol. 3, Academic Press, pp. 41-58.

HAMBLING C.L. (1970), The effect of when it's said. *Theoria*. 3 : 249-263.

MULLER P., Prévot L. (2002), Conversation sous les topiques, du contenu propositionnel à la structure du dialogue. *Information - Interaction – Intelligence,* Hors série 2002, pp.179-196.

PREVOT L. (2004), Structures sémantiques et pragmatiques pour la modélisation de la cohérence dans les dialogues finalisés. Thèse de doctorat de l'université Paul Sabatier, Toulouse.

RECANATI F. (2001), Déstabiliser le sens. *Revue Internationale de Philosophie* 2/2001(217).

RECANATI F. (2003), Embedded Implicatures. *Philosophical Perspectives*.

TROGNON A. (1995), Structures interlocutoires. Cahiers de Linguistique Française, (17):79-98, 1995.

VANDERVEKEN D. (1985), *Logique illocutoire*, Bruxelles, Mardaga éd.

VERNANT D. (1997), *Du discours à l'action*. Presses Universitaires de France, Paris.

XUEREB A., CAELEN J. (2004) Un modèle d'interprétation pragmatique en dialogue homme-machine basé sur la SDRT, Actes de *TALN'04, XIème Conférence sur le Traitement Automatique du Langage Naturel*, ISBN 2-9518235-5-5, pp. 505-514.

XUEREB A., CAELEN J. (2005) Actes de langage et relations rhétoriques en dialogue homme-machine. Presses universitaires de Nancy (à paraître). Présenté au séminaire Dialogue et logique, Nancy, 2004.

[2] Cette ontologie couvre un domaine restreint de réservation de salle.

Détection automatique d'actes de dialogues par l'utilisation d'indices multi-niveaux

Sophie Rosset, Delphine Tribout
LIMSI - CNRS
F-91403 Orsay Cedex
{rosset, tribout}@limsi.fr

Mots-clefs : actes de dialogue, dialogue homme homme, détection automatique, indices multiniveaux

Keywords: dialog acts, human human dialog, automatic detection of dialog acts, mulilevel information

Résumé Ces dernières années, il y a eu de nombreux travaux portant sur l'utilisation d'actes de dialogue pour caractériser les dialogues homme-homme ou homme-machine. Cet article fait état de nos travaux sur la détection automatique d'actes de dialogue dans des corpus réels de dialogue homme-homme. Notre travail est fondé essentiellement sur deux hypothèses . (i) la position des mots et la classe sémantique du mot sont plus importants que les mots eux-mêmes pour identifier l'acte de dialogue et (ii) il y a une forte prédictivité dans la succession des actes de dialogues portés sur un même segment dialogique. Une approche de type *Memory Based Learning* a été utilisée pour la détection automatique des actes de dialogue. Le premier modèle n'utilise pas d'autres informations que celles contenus dans le tour de parole. Dans lex expériences suivantes, des historiques dialogiques de taille variables sont utilisés. Le taux d'erreur de détection d'actes de dialogue est d'environ 16% avec le premier modèle est descend avec une utilisation plus large de l'historique du dialogue à environ 14%.

Abstract Recently there has been growing interest in using dialog acts to characterize human-human and human-machine dialogs. This paper reports on our experience in the annotation and the automatic detection of dialog acts in human-human spoken dialog corpora. Our work is based on two hypotheses: first, word position is more important than the exact word in identifying the dialog act; and second, there is a strong grammar constraining the sequence of dialog acts. A memory based learning approach has been used to detect dialog acts. In a first set of experiments only the information contained in each turn is used and in a second set, different histories of the dialogue are used. A dialog act error rate of about 16 % is obtained for the simplest model. Using other informations, such as history of the dialog, the results grow up to 14%.

1 Introduction

Afin de saisir la complexité de dialogues homme-homme collectés dans des services d'appels, il semble intéressant d'explorer et de corréler différents types d'information, disponibles sous

la forme d'annotations faites à différents niveaux : lexical, sémantique et fonctionnel. D'autre part, il peut être utile de modéliser la structure du dialogue afin d'utiliser cette information dans des systèmes de dialogue. Une analyse souvent effectuée sur les dialogues concerne les actes de dialogue. Les actes de dialogues sont en quelque sorte des unités fonctionnelles abstraites qui décrivent les actions des locuteurs tout en généralisant les variations de forme et de contenu des énoncés. Par exemple, on considère que les assertions (*assert*), les demandes d'informations (*information-request*), les marques d'accord (*acknowledgement*) sont des actes de dialogue qui permettent d'approcher ce que les locuteurs souhaitent accomplir par leur parole. À la base de ce courant, on trouve l'idée selon laquelle *dire c'est faire*[1], c'est-à-dire selon laquelle tout acte d'énonciation serait la réalisation d'un acte social. Cette conception de la parole vient de philosophes du langage ((Austin J. L., 1962),(Searle J. R., 1969)) qui considèrent le dialogue comme un lieu d'interaction sociale et la parole comme un moyen d'(inter)action. Austin considère ainsi qu'une énonciation, outre un contenu explicite, permet également d'accomplir un acte et a donc à ce titre une fonction pragmatique. L'idée d'Austin est en outre que ces fonctions des énoncés peuvent être étudiées indépendamment de leur structure syntaxique mais selon un certain contexte. Plusieurs travaux récents sont fondés sur l'idée que les actes de dialogue sont une bonne façon de caractériser les dialogues, tant dans les interactions homme-homme que homme-machine. Pour exemple, nous pouvons citer les travaux de (Cattoni R. et al., 2001), de (Di Eugenio B. et al., 1998) ou encore ceux de (Isard A., Carletta J. C., 1995). De nombreuses taxonomies d'actes de dialogue ont donc été établies (Traum D., 2000). Pour ce qui concerne les systèmes de dialogue et les annotations de corpus de dialogue homme-homme et homme-machine, une taxonomie fréquemment utilisée et largement répandue est celle de DAMSL [2]. Quant aux approches pour l'annotation automatique en actes dialogiques, il en existe plusieurs qui diffèrent légèrement. Par exemple, ayant observé dans plusieurs corpus que les différents actes de dialogue sont fortement corrélés à des suite de mots précis (appelés *cue-phrases*), (Hirschberg J. et Litman D. J., 1993) se fondent sur ces indices pour les détecter. Le problème de cette approche est toutefois que ces suites de mots sont fortement dépendantes de la tâche et du domaine. Afin de pallier ce problème, (Reithinger N. et Klesen M., 1997) proposent l'utilisation de n-gramme de mots. (Samuel K. et al., 1998), quant à eux, se situent à l'intersection de ces deux approches et utilisent les suites de mots et un sous-ensemble d'indices dialogiques modélisés par des n-gramme de mots. Il est toutefois difficile de ne pas constater que, en règle générale, la relation entre les actes de dialogue et les mots n'est pas univoque. Par exemple, un simple mot comme *oui* peut correspondre à différents actes de dialogue comme une réponse à une question, la confirmation d'une information, un backchannel... D'un autre côté, un acte de dialogue comme une assertion peut correspondre à plusieurs mots ou suites de mots comme *ma date de naissance est le 31/08/70* ou *68 euros 50*... Afin de réduire autant que possible la dépendance à la tâche tout en gérant ces correspondances multiples, nous avons cherché à élaborer une méthode de détection des actes de dialogue sans utilisation explicite du lexique, notre hyothèse étant que cette information n'est pas strictement indispensable.

Dans cet article nous présentons donc notre méthodologie pour la détection automatique des actes de dialogue. Puis nous présentons les différentes expériences que nous avons menées et qui nous ont permis d'améliorer la détection automatique des actes de dialogue grâce à la prise en compte et l'intégration d'indices dialogiques supplémentaires.

[1]Ceci fait référence au titre de la version française de (Austin J. L., 1962).

[2]Cette taxonomie a été utilisée et adaptée dans nombre de projets. Le projet européen et américain AMITIÉS (Automated Multilingual Interaction with Information and Services) par exemple s'est fondée sur elle pour proposer une méthode d'annotation des dialogues sur différents niveaux

nombre de dialogues	134
nombre de tours	4273
nombre moyen de tours/dialogue	32
nombre de segments dialogiques	5623
nombre moyen de segments dialogiques/dialogue	42
nombre moyen de segments dialogiques/tour	1.3
nombre de mots distincts	1976
nombre total de mots	40494

Table 1: Descriptif du corpus GE_fr

2 Corpus et méthodologie

2.1 Corpus utilisé

Dans ce travail, le corpus utilisé (cf. tableau 1) consiste en une série de dialogues homme-homme en français, enregistrés dans un centre d'appel d'un service de prêts bancaires (GE_fr). Ces dialogues couvrent une grande variété de thèmes comme la demande d'informations (limites de crédits possibles, informations sur le disponible), le passage d'ordres (modification des limites de crédits, changement des mensualites...), la gestion de compte (ouverture et fermeture, modification des informations personnelles)... Le corpus au complet est constitué de 134 dialogues. Ces dialogues sont transcrits avec l'outil d'alignement transcription/signal Transcriber. Ce corpus a été divisé en trois partie pour l'apprentissage (94 dialogues, 2923 tours de parole et 3912 segments dialogiques), le développement (22 dialogues, 687 tours de parole et 884 segments dialogiques) et le test (18 dialogues, 663 tours de parole et 827 segments dialogiques). Ce corpus a par ailleurs été tagué en entités spécifiques : entités nommées (personne, lieu, date etc.), entités dépendantes de la tâche (i.e. entités nommées faisant appel à une connaissance spécifique du domaine ; par exemple numéro de compte, adresse, montant disponible sur un compte etc.).

2.2 Principes d'annotation

Ce corpus a été annoté avec le schéma d'annotation dialogique proposé dans le cadre du projet AMITIÉS et fondé sur la taxonomie de DAMSL (Hardy H. et al., 2002). Les annotations devant permettre de décrire et résumer l'intention du locuteur, huit classes ont été établies afin d'obtenir une annotation fine et sur différents niveaux. La taxonomie obtenue est la suivante :

- **Classe 1** *Information Level* : permet d'annoter un énoncé dans son rapport à la réalisation de la tâche. Les tags possibles sont : *Communication-mgt, Out-of-topic, Task, Task-management-Completion, Task-management-Order, Task-management-Summary, Task-management-System-Capabilities.*
- **Classe 2** *Statement* : permet d'annoter les énoncés déclaratifs ayant un contenu informatif explicite. Les tags possibles sont : *Assert, Commit, Explanation, Expression, ReExplanation, Reassert.*
- **Classe 3** *Conventional* : permet de noter les aspects conventionnels d'un dialogue homme-homme, comme les ouvertures et fermetures. Les tags possibles sont : *Closing, Opening.*

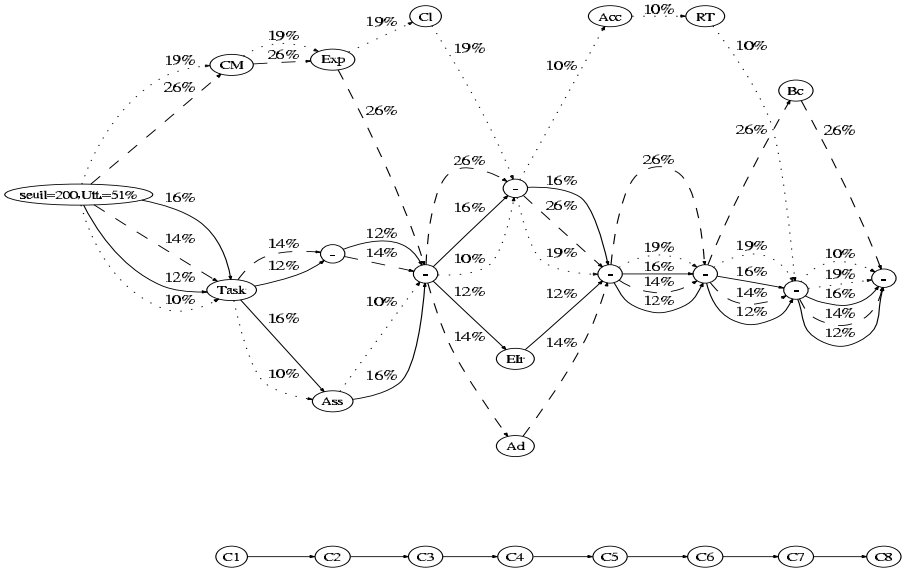


Figure 1: Les successions les plus fréquentes d'ADs

- **Classe 4** *Influence on Listener* : permet de rendre compte de l'intention du locuteur en tant qu'agissant sur le dfoulement du dialogue. Les tags possibles sont : *Action-directive, Explicit-Confirm-request, Explicit-Info-request, Implicit-Confirm-request, Implicit-Info-request, Offer, Open-Option, Re-Action-directive, Re-Confirm-request, Re-Info-request, Re-Offer.*

- **Classe 5** *Agreement* : permet de spécifier l'accord ou le désaccord du locuteur avec ce qui précède. Les tags possibles sont : *Accept, Accept-part, Maybe, Reject, Reject-part.*

- **Classe 6** *Answer* : permet de préciser si l'énonce en question constitue une réponse à un énoncé précédent. Le tag utilisé est alors : *True.*

- **Classe 7** *Understanding* : permet de noter les degrés de compréhension du locuteur. Les tags possibles sont : *Backchannel, Completion, Correction, Non-understanding, Repeat-rephrase.*

- **Classe 8** *Communicative Status* : permet d'annoter les apartés les interruptions, les changement de sujet... Les tags possibles sont : *AbandStyle, AbandTrans, AbandChangeMind, AbandlossIdeas, Interrupted, Self-talk.*

Les actes de dialogue couvrant différents aspects conversationnels, un même segment dialogique peut contenir plusieurs actes de dialogue. Par conséquent plusieurs de ces classes peuvent être sélectionnées pour décrire un même segment dialogique. Chaque segment dialogique peut en effet être catégorisé selon son niveau informationnel ainsi que selon son aspect conventionnel, son influence sur la suite du dialogue... Ceci implique qu'un segment dialogique peut potentiellement recevoir une étiquette de chacune de ces classes. Par exemple, le segment dialogique *A for Alpha* est annoté avec l'étiquette *Explicit-Confirm-request* de la classe *Influence on Listener* et *Non-understanding* de la classe *Understanding*.

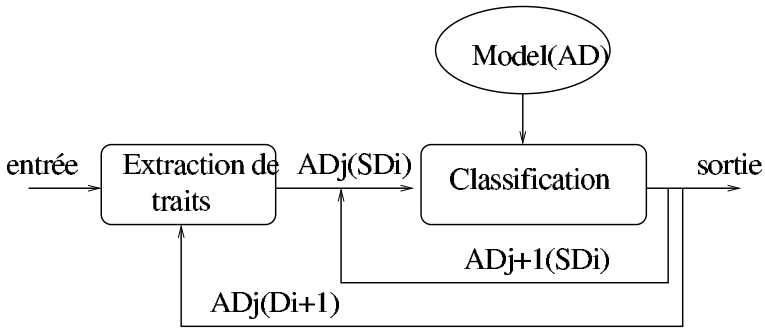


Figure 2: Raisonnement fondé sur la similarité et utilisation des hypothèses précédentes

2.3 Méthodologie pour l'annotation automatique

Pour effectuer l'annotation automatique nous avons représenté chaque enoncé du corpus d'apprentistisage et du corpus de developpement par un vecteur puis effectué des comparaisons. Pour cela nous avons choisi une approche de type *Memory Based Learning* car elle fonctionne bien sur de petites quantités de données. En outre, différentes études (van den Bosch A. et al., 2001; Daelemans W. et al., 1999) ont montré qu'elle était particulièrement bien adapté au traitement automatique de la langue. Nous avons utilisé l'implémentation IB1-IG du logiciel Timbl (Daelemans W. et al. 2003) avec une distance de manhatan. Dans cette métrique, la distance entre deux objets est simplement la somme de la différence entre les différents traits de ces objets. Le principe de cette approche est relativement simple : il s'agit de comparer le vecteur entrant à l'ensemble des vecteurs du modèle et d'assigner à celui-ci la classe du vecteur du modèle dont il est le plus proche. Le corpus d'apprentissage a servi de modèle de vecteurs et le corpus de developpement a constitué l'ensemble des vecteurs à classer. Les traits choisis pour la construction des vecteurs sont l'identité du locuteur (Client ou Agent), le nombre de segments dialogiques dans le tour considéré, les premiers mots du segment annoté et les tags des huit classes définies. Pour ce qui est de l'utilisation des mots comme traits, notre hypothèse étant que les premiers mots sont plus importants que l'ensemble des mots, seuls les premiers mots de chaque segment dialogique ont été utilisés. Quant aux tags, si aucune étiquette d'une classe considérée n'est pertinente pour le segment alors la classe est représentée par le tag "NA" (not applicable) de manière à avoir toujours le même nombre de traits dans un vecteur. Ainsi, les énoncés du corpus d'apprentissage constituant le modèle de vecteurs ont été représentés de la façon suivante (pour un nombre de mots égal à 4) :

pour un énoncé tel que
Agent: *donnez moi votre numéro de compte*

ayant pour annotation les étiquettes :

DAs: *information-level=Task ; influence-on-listener=Action-directive*

le vecteur correspondant est :

Agent 1 donnez moi votre numéro Task NA NA Action-directive NA NA NA NA

De la même façon, l'enoncé

Client : alors [numerique] [numerique] [numerique] [numerique] [numerique] [numerique]

ayant pour annotation les étiquettes :

DAs: *information-level=Task ; statement=Assert ; agreement=Accept ; answer=true*

est représenté par le vecteur :

Client 1 alors [numerique] [numerique] [numerique] Task Assert NA NA Accept true NA NA

Alors qu'*a priori* la combinatoire des tags est relativement importante, seulement 197 combinaisons différentes sont retrouvées dans le corpus d'entraînement. Il y a donc un facteur de prédictivité relativement important dans la succession des classes et des étiquettes sélectionnées. Ceci est illustré par la figure 1 qui peut être vue comme une grammaire de succession d'étiquettes dialogiques. Sur cette figure, les six successions les plus fréquentes d'étiquettes dialogiques sont représentées. Elles couvrent 51% des séquences du corpus d'entraînement. Ainsi, si Task est sélectionné pour la classe 1 (52%) alors la classe 2 recevra soit NA (26%) soit Assert (26%) et la classe 3 NA. Ceci montre qu'il semble y avoir un facteur de prédictivité relativement important dans la succession des classes et des étiquettes sélectionnées. C'est pourquoi l'annotation des segments dialogiques du corpus à annoter est réalisée en huit étapes : une pour chacune des huit classes d'annotation. A chaque étape un tag est affecté à la classe correspondante et celui-ci est ajouté au vecteur lors de l'étape suivante. En outre, l'annotation dialogique est effectuée en tenant compte de l'ensemble du tour de parole et dans l'ordre des segments composant un tour. C'est-à-dire que pour chaque tour de parole, le premier segment dialogique est d'abord annoté en huit étapes, puis le second segment est ajouté au premier et annoté à son tour en huit étapes. La méthode consiste donc pour le système à extraire du tour de parole donné en entrée les traits retenus (identité du locuteur, nombre de segments dialogiques, N premiers mots) et à les placer dans un vecteur ($SD_1(AD_i)$=[SpkrId, #Utt., w_1,w_2,AD_{i-1}]). L'assignation d'une étiquette dialogique à ce vecteur (la classification) est faite en comparant ce vecteur à l'ensemble des vecteurs du modèle, et ceci à huit reprises (une pour chacune des classes d'annotation) en ajoutant à chaque étape de la classification l'étiquette déterminée à l'étape précédente. Dès que le premier segment dialogique de l'énoncé en cours de traitement est annoté, si cet énoncé comporte un ou plusieurs autres segments dialogiques, les N premiers mots du segment suivant sont ajoutés au vecteur initial et la classification s'effectue de la même façon. Ainsi le nouveau vecteur est composé de : $SD_i(SD_{i-1} + AD_i$+w_{1SD_i},w_{2SD_i},). Cette méthode utilisée pour annoter automatiquement les énoncés est représentée par la figure 2.

3 Expériences et résultats

3.1 Première expérience : modèle de base

La première expérience, effectuée sur le corpus d'apprentissage, a consisté à faire varier le nombre de mots donnés en entrée : (1) les 4 premiers mots du premier segment dialogique, (2) les 4 premiers mots du premier segment dialogique et les 2 premiers mots du segment pour les segments dialogiques suivants, (3) les 2 premiers mots du premier segment dialogique et les 2 premiers mots du segment pour les segments dialogiques suivants. Les résultats sont présentés dans le tableau 2. Par ailleurs, un examen des poids attribués par TiMBL à chaque trait des vecteurs montre que les mots sont dotés d'un faible poids, ce qui indique qu'ils ne constituent pas le critère le plus pertinent pour catégoriser les vecteurs entrant. Cette observation va bien dans le sens de notre hypothèse lexicale. Toutefois, un examen plus attentif des résultats montrent que certains actes de dialogues sont plus dépendants du lexique que d'autres. Le tableau 3 permet de mettre cette constatation en évidence.

Data	# dial	#seg. dial.	#tour	%erreur	exp.
GE_fr dev	22	884	687	14.0	4words
				13.2	4+2words
				13.0	2+2words
GE_fr Test	18	827	663	16.5	4words
				16.2	4+2words
				17.2	2+2words

Table 2: Taux d'erreur de détection d'actes de dialogues sur corpus developpement et test.

DA	GE_fr dev
Response-To	52% (148)
Backchannel	12.5% (161)
Accept	46.9% (147)
Assert	25.9% (243)
Expression	7.9% (378)
Comm-mgt	8.3% (420)
Task	12.4% (427)

Table 3: Taux d'erreur sur les 7 tags les plus fréquents.

3.2 Réduction de la variation lexicale

Un prétraitement des énoncés a ensuite été effectué pour tenter de réduire la variation de certains énoncés jugés équivalents au niveau sémantique et fonctionnel. L'hypothèse sous-jacente à cette tentative est que l'annotation automatique gênée par certaines variations formelles d'énoncés pourtant équivalents du point de vue du sens pourrait être améliorée si ces variations étaient neutralisées. Ainsi pour la conjugaison des verbes la variation en temps gêne souvent l'annotation. Les formes "je voudrais", "je voulais", "j'aurais voulu", par exemple sont à peu près équivalentes dans le contexte de demande d'information. Nous avons donc décidé de réduire ces variations à une même forme neutralisée "*vouloir", que nous distingons de l'infinitif avec le signe "*". Les neutralisations concernent également certains morceaux d'énoncés récurrents que nous avons jugés équivalents comme "je vous écoute", "c' est à quel propos", "c 'est pourquoi", "c' est à quel sujet"... qui sont tous une façon pour l'agent d'inviter le client à exposer le sujet de son appel, et que nous avons donc neutralisés en *invite. Ces énoncés neutralisés ont été déterminés après l'étude des dialogues du corpus d'apprentissage et concernent notamment les différentes façons de remercier ("je vous remercie", "c' est moi qui vous remercie", "merci beaucoup", "merci bien"...), de demander ("vous pouvez me donner", "vous pouvez me rappeler", "donnez-moi", "rappelez-moi"...) d'ouvrir la conversation ("bonjour", "bonsoir", "allo")... La réduction de ces variations lexicales a été effectuée automatiquement sur l'ensemble du corpus en utilisant l'outil de détection en entités nommées qui a été étendu. Cette étape a permis de faire passer la taille du lexique de 1976 mots à 1649 mots. Les mêmes expériences que précédemment ont ensuite été menées sur ce corpus et les résultats obtenus dans ces conditions montrent que la réduction de la variation lexicale améliore l'annotation automatique. Le tableau 4 donne les résultats obtenus avec ce second modèle.

Exp.	%erreur Dev	%erreur Test.
4words	14.0	16.4
4+2words	12.9	16.2
2+2words	12.7	16.7

Table 4: Taux d'erreur de détection d'actes de dialogues sur corpus développement et test après réduction de la variation

# Seg. Dial	%erreur Dev.	%erreur Test
1 SD	13.2	16.8
2 SD	9.2	16.0
3 SD	22.4	19.2

Table 5: Taux d'erreur de détection d'actes de dialogues sur corpus développement et test par segment dialogique

3.3 Utilisation des historiques

Des expériences supplémentaires ont ensuite été menées en jouant sur l'historique du dialogue afin de voir si cela améliorait la détection automatique des actes de dialogue. Partant des résultats précédemment obtenus, deux hypothèses ont servi de base à ces expériences. La première est que s'il existe plusieurs segments dialogiques au sein d'un tour, ceux-ci entretiennent des relations entre eux et sont organisés selon certains principes. La seconde est que le dialogue constituant une alternance de tours de parole se répondant les uns aux autres, les actes de dialogue d'un tour sont en partie conditionnés par les actes de dialogue du tour précédent. Les expériences suivantes ont donc été réalisées afin de voir si ces deux hypothèses se vérifiaient.
En ce qui concerne la première hypothèse, nous avons constaté à partir des expériences précédentes qu'au sein d'un tour de parole les résultats de l'annotation automatique étaient systématiquement meilleurs pour le deuxième segment dialogique que pour le premier, et tout aussi systématiquement, nettement moins bons pour le troisième segment[3] ainsi que le montre le tableau 5. Le premier segment dialogique, qui est ajouté au vecteur lorsqu'on annote le suivant, semble donc aider à déterminer le second. Celui-ci serait donc étroitement lié au premier et constituerait une sorte de prolongement ou de complément du premier. En revanche, pour le troisième segment les segments précédents qui sont eux aussi ajoutés au vecteur ne semblent pas aider à sa détection. De cette constatation nous avons émis l'hypothèse que si le second segment dialogique d'un tour de parole est dans la continuité du premier, le troisième semble au contraire être en rupture avec eux. Nous avons donc essayé d'annoter automatiquement les segments dialogiques sans tenir compte des deux premiers segments pour annoter le troisième, c'est-à-dire sans mettre les deux premiers segments dans le vecteur du troisième. Avec cette méthode les résultats du troisième segment sont nettement meilleurs comme le montre le tableau 6 (exp. 1). Ceci semble donc confirmer notre hypothèse de départ selon laquelle les segments dialogiques d'un tour de parole entretiennent entre eux des relations qui ne sont pas toutes de la même nature.
En ce qui concerne la seconde hypothèse, les tours de parole se répondant les uns les autres, il nous a semblé intéressant de prendre en compte les informations contenues dans le tour précédent pour déterminer les segments dialogiques d'un tour donné. Toutefois, compte tenu de ce qui vient

[3]les tours de parole comprenant plus de trois segments dialogiques n'étant pas assez nombreux pour pouvoir émettre une hypothèse à leur sujet l'étude s'est bornée aux trois premiers segments.

Exp.	Dev. %erreur	Test %erreur	Seg. Dial
Exp. 1	13.2	16.8	1
	9.2	16.5	2
	19.7	12.5	3
Exp. 2	10.4	13.7	1
	10.4	15.6	2
	25	14.2	3
Exp. 3	10.4	13.7	1
	10.4	15.6	2
	15.8	13.3	3

Table 6: Taux d'erreur de détection d'actes de dialogues sur corpus developpement et test avec variation sur les historiques

d'être exposé sur les relations que semblent entretenir les segments dialogiques au sein d'un même tour de parole, il nous a semblé que toutes les informations du tour précédent n'étaient sans doute pas pertinentes et que seules les informations relatives au dernier segment dialogique devait avoir une influence sur les segments dialogiques du tour suivant. Nous avons donc ajouté au vecteur les annotations du dernier segment du tour précédent. Les résultats concernant le premier et le deuxième segments ont ainsi été améliorés, mais pas ceux du troisième qui se sont plutôt dégradés (cf. tableau 6 (exp. 2)).
Compte tenu de ces résultats et de ceux obtenus par l'expérience précédente, nous avons essayé de mêler les deux méthodes afin d'améliorer encore l'annotation automatique. Ainsi, pour le premier et le deuxième segments d'un tour de parole les annotations du dernier segment dialogique du tour précédent ont été ajoutées au vecteur, tandis que pour le troisième segment ni les deux segments précédents ni le dernier segment du tour précédent n'ont été pris en compte. Avec cette méthode "mixte", les résultats ont ainsi été améliorés pour tous les segments (cf. tableau 6 (exp. 3)).

4 Conclusion et Perspectives

Ces travaux réalisés dans le cadre de l'annotation dialogique automatique ont permis de mettre en avant le fait que les tours de parole ne sont pas des suites d'énoncés anarchiques mais sont au contraire structurés selon certains principes. La succession des segments dialogiques qui les composent semble en effet organisée : le premier segment dialogique d'un tour de parole semble être lié au dernier segment du tour précédent, le deuxième segment d'un tour semble lié au premier, tandis que le troisième semble indépendant des précédents. En outre, l'étude des dialogues dans leur entier a également fait ressortir une certaine structure du dialogue.

En ce qui concerne l'annotation automatique d'actes de dialogue, les différentes expériences menées ont contribué à son amélioration. Toutefois, celle-ci peut encore être améliorée davantage et d'autres méthodes sont à envisager. Nous envisageons notamment d'utiliser des informations sémantiques, disponibles actuellement sous la forme d'annotations appliquées aux mêmes données, afin de voir si elles ne permettraient pas une meilleure détection des actes de dialogue. Une autre méthode que nous envisageons est d'effectuer une classification des énoncés du corpus d'apprentissage et de faire pour chacune classes dégagées des modèles propres. Ceux-ci seraient

ainsi plus spécificques et les nouveaux énoncés pourraient alors être annotés en fonction du modèle qui leur correspond le mieux. Enfin, les résultats présentés font tous l'assomption que le nombre de segments dialogiques par tour de parole est connu ainsi que les frontières elles-mêmes. Il est bien entendu que dans le cadre d'un système automatique, ces informations doivent être estimées. Une étude précédente (Rosset S. et Lamel L., 2004) avait montré qu'il était possible de prédire de manière raisonnable le nombre de segment dialogique dans un énoncé. Des taux d'érreurs sur cette tâche de l'ordre de 12% ont été rapportés. Environ 5 à 7% des tour de parole présentaient une insertion ou une suppression de frontière. Ces modèles pour la détection de frontières de segments dialogiques doivent donc également être intégrés au système de détection d'actes de dialogue pour avoir un système entièrement automatique.

Références

J. L. Austin (1962), How to do thing with words, Oxford: Clarendon Press.

R. Cattoni, M. Danieli, A. Panizza, V. Sandrini, C. Soria (2001), Building a corpus of annotated dialogues: the ADAM experience, *C*orpus Linguistics 2001.

W. Daelemans, J. Zavrel, K. van der Sloot, A. van den Bosch (2003), TiMBL: Tilburg Memory Based Learner, v5.0, Reference Guide, *I*LK Technical Report ILK-03-10, (http://ilk.kub.nl/software.html#timbl)

W. Daelemans, A. van den Bosch, J. Zavrel (1999), Forgetting exceptions is harmful in language learning, *M*achine Learning Vol 34:11-43.

B. Di Eugenio, P. W. Jordan, J. D. Moore, R. H. Thomason (1998) An empirical investigation of collaborative dialogues, Actes de *A*CL, COLING.

H. Hardy, K. Baker, H. Bonneau-Maynard, L. Devillers, S. Rosset, T. Strzalkowski (2002), Semantic and Dialogic annotation for Automated Multilingual Customer Service, Actes de *I*SLE workshop.

J. Hirschberg, D.J. Litman (1993), Empirical Studies on the Disambiguation of Cue Phrases, *C*omputational Linguistics Vol. 19(3):501-530.

A. Isard, J.C. Carletta (1995), Replicability of transaction and action coding in the map task corpus, Actes de *A*AAI Spring Symposium: Empirical Methods in Discourse Interpretation and Generation.

N. Reithinger, M. Klesen. (1997), Dialogue act classification using language models, Actes de *E*urospeech'97

S. Rosset et L. Lamel (2004), Automatic Detection of Dialog Acts Based on Multi-level Information, Actes de *I*CSLP'04

K. Samuel, S. Carberry, K. Vijay-Shanker (1998), Dialogue act tagging with transformation-based learning, Actes de *C*OLING-ACL

J. R. Searle (1969), Speech acts, Cambridge University Press.

D. Traum (2000), 20 Questions on Dialog Act taxonomies, *J*ournal of Semantics, Vol. 17(1):7-30.

A. van den Bosch, E. Krahmer, M. Swerts (2001), Detecting problematic turns in human-machine interactions: Rule-induction versus memory-based learning approaches, Actes de *A*CL'01

Comment mesurer la couverture d'une ressource terminologique pour un corpus ?

Goritsa Ninova (1), Adeline Nazarenko (2),
Thierry Hamon (2), Sylvie Szulman (2)

LIPN UMR 7030
Université Paris 13 & CNRS
99, av. J.-B. Clément
93430 Villetaneuse
(1) cylvago@yahoo.fr
(2){prénom.nom}@lipn.univ-paris13.fr

Mots-clés : couverture lexicale, terminologie, statistique lexicale
Keywords : lexical coverage, terminology, lexical statistics

Résumé Cet article propose une définition formelle de la notion de couverture lexicale. Celle-ci repose sur un ensemble de quatre métriques qui donnent une vue globale de l'adéquation d'une ressource lexicale à un corpus et permettent ainsi de guider le choix d'une ressource en fonction d'un corpus donné. Les métriques proposées sont testées dans le contexte de l'analyse de corpus spécialisés en génomique : 5 terminologies différentes sont confrontées à 4 corpus. La combinaison des valeurs obtenues permet de discerner différents types de relations entre ressources et corpus.

Abstract This paper proposes a formal definition of the notion of lexical coverage. This definition is based on four metrics that give a global view over a lexical resource to corpus relationship, thus helping the choice of a relevant resource with respect to a given corpus. These metrics have been experimented in the context of specialised corpus analysis in genomics. 5 terminologies have been confronted to 4 different corpora. The combination of resulting figures reflects various types of corpus *vs* . resource relationships.

1 Introduction

On parle couramment de « couverture lexicale » sans définir clairement ce qu'on entend par là. Différents auteurs mettent sous ce terme différentes notions et mesures. Le problème est d'autant plus complexe que les ressources utilisées comportent souvent des expressions polylexicales dont la projection en corpus peut se faire de différentes manières. Le présent article propose de définir un ensemble de métriques pour cerner cette notion de couverture dans le cas général d'une ressource constituée d'une liste de termes mono- et polylexicaux. Ces mesures sont testées

pour différents couples ressource/corpus. Les premiers résultats obtenus sont encourageants. Ils montrent qu'on peut en effet documenter le comportement d'une ressource pour un corpus donné en préalable à tout traitement, et ainsi guider le choix de la ressource.

Après avoir souligné les enjeux de cette problématique et les questions qu'elle soulève (section 2), nous présentons dans la section 3 un ensemble de métriques. Celles-ci sont exploitées dans la perspective du traitement automatique de corpus de génomique. Les résultats de ces expériences sont présentés et discutés dans la section 4 de cet article.

2 Problématique

2.1 Enjeux

Le traitement de corpus spécialisé fait appel à des ressources sémantiques qu'on appelle généralement spécialisées parce qu'elles décrivent un domaine particulier d'activité. Ces ressources peuvent être de différents types selon les traitements envisagés, mais elles doivent comporter une dimension lexicale dès lors qu'elles sont destinées à l'analyse et l'interprétation de données textuelles.

Les ontologies du web sémantique doivent ainsi être ancrées lexicalement (avec des items lexicaux associés aux noeuds de l'ontologie) si elles doivent servir à indexer des textes. Les techniques d'accès au contenu des documents textuels sont diverses (extraction d'information, question-réponse, outils de navigation ou de résumé) mais elles reposent toutes sur une analyse sémantique partielle des documents, qui implique la reconnaissance de certains éléments du discours (entités nommées et termes du domaine, notamment), leur typage sémantique et leur mise en relation (Nazarenko, 2005). De ce fait, ces techniques reposent également sur des lexiques, terminologies ou thesaurus spécialisés pour identifier le vocabulaire de spécialité. Les catégories sémantiques et les relations lexicales sont utilisées (quand elles existent) pour désambiguïser les textes et en guider l'interprétation.

Dès lors que les applications de traitement automatique des langues (TAL), y compris au niveau sémantique, sont de plus en plus guidées par le lexique, la question du choix des ressources à exploiter prend de l'importance. La situation relève souvent à la fois de la pléthore et de la pénurie. D'un côté, il existe de nombreuses ressources terminologiques, surtout dans des domaines comme la biologie ou la médecine où l'effort d'organisation des connaissances est ancien[1]. Mais d'un autre côté, les « bonnes ressources » sont rares : le degré de spécialisation ou le point de vue représenté par la ressource est généralement différent de celui du texte que l'on cherche à analyser. Ce constat a été fait par (Charlet *et al.*, 1996), toujours dans le domaine de la médecine pourtant reconnu pour la richesse de ses bases de connaissances. Dans la pratique, comme on ne peut ni se passer de ressource, ni en reconstruire de nouvelles pour chaque nouvelle application, on fait souvent avec ce qu'on a. Dans certains cas, on peut spécialiser la ressource et l'adapter en fonction du domaine et de la tâche visés (problématique de l'adaptation lexicale ou « lexical tuning » (Basili *et al.*, 1998)) mais cela suppose néanmoins une ressource initiale.

Une question se pose alors : parmi l'ensemble des ressources qui paraissent recouvrir en partie et *a priori* le domaine du corpus à traiter, laquelle ou lesquelles choisir et sur quels critères ? Cette question est d'autant plus importante qu'on doit souvent limiter le nombre de ces ressources pour réduire l'inévitable travail de préparation des données et pour éviter les problèmes d'incohérence. Il est en général trop coûteux d'exploiter en parallèle différentes ressources pour les tester en les comparant au regard de l'application visée. Les experts du domaine ne sont pas toujours d'un grand secours non plus. Même s'ils sont capables de décrire

[1] Voir par exemple, UMLS (Unified Medical Language System, http://www.nlm.nih.gov/research/umls/).

le sous-domaine couvert par une ressource et le point de vue qui y est représenté, ils sont d'ordinaire peu à même de mesurer son adéquation proprement lexicale.

Ce problème du choix des ressources est souvent résolu de manière très empirique, ce qui ne permet pas de capitaliser d'une expérience à l'autre. Il est donc important de se doter de critères formels permettant de décrire le comportement d'une ressource par rapport à un corpus donné et d'en guider le choix. C'est l'objet de ce travail : nous proposons un premier ensemble de métriques pour apprécier la couverture d'un corpus par une ressource terminologique.

2.2 Difficultés

Pour des dictionnaires traditionnels, on exprime l'adéquation à un corpus en termes de couverture et on l'apprécie à partir du nombre d'occurrences de mots du corpus qui se rattachent à des entrées du dictionnaire. La couverture est plus difficile à définir pour des ressources terminologiques.

La première difficulté tient à la diversité des ressources terminologiques qui rend problématique leur comparaison. La *nature* de l'information diffère d'une ressource à l'autre. Au-delà des listes de termes, les termes eux-mêmes peuvent être typés et les types peuvent être organisés en hiérarchie (thesaurus). Dans les ressources les plus riches, les termes sont de surcroît liés entre eux par des liens sémantiques. Les ressources ont par ailleurs des *degrés de spécialisation* divers. Il est difficile de comparer un lexique de 10 000 unités qui comporterait de nombreuses unités également présentes dans des dictionnaires généralistes et un lexique de 500 unités dont très peu figurent dans des dictionnaires classiques. Les ressources s'opposent enfin par leur *degré de lexicalisation* : certaines se contentent de lister des étiquettes de concepts ; d'autres considèrent ces étiquettes dans leur dimension lexicale et linguistique. Ces dernières rendent compte des différentes formes sous lesquelles ces unités sémantiques peuvent se réaliser en corpus, jusqu'à associer des règles de désambiguïsation contextuelles aux unités polysémiques (Nédellec, Nazarenko, 2005).

La deuxième difficulté tient au fait qu'on cherche à confronter deux objets qui ne sont pas de même nature. La ressource et le corpus s'opposent comme la langue s'oppose au discours : il faut comparer un ensemble d'éléments de lexique (la ressource) avec un ensemble d'occurrences (le corpus). Par voie de conséquence, il faut aussi comparer des unités potentiellement polylexicales avec des occurrences observées en corpus, nécessairement monolexicales. Comme il s'agit d'apprécier *a priori* l'adéquation des ressources aux corpus, nous ne présupposons en effet aucune étape de reconnaissance terminologique préalable.

Dans ce premier travail, nous focalisons l'étude sur les ressources terminologiques considérées comme des listes de termes, sans exploiter les éventuelles informations qu'elles contiennent concernant leurs règles de variation, leur typage sémantique ou les relations sémantiques qu'ils entretiennent. Les premiers éléments étant posés, il est évidemment nécessaire de poursuivre, par exemple, en prenant en compte la désambiguïsation des termes polysémiques, les liens de variations entre termes et la structure sémantique. Ces points ne sont pas abordés ici.

2.3 État de l'art

La question de la sélection des ontologies pour une application prend de l'importance avec l'augmentation du nombre des ontologies disponibles et la standardisation de leurs formats. Cette préoccupation est au coeur de la problématique du web sémantique. (Buitelaar *et al.*, 2004) montre que la création d'une bibliothèque d'ontologies (OntoSelect) suppose de définir des critères permettant de sélectionner une ontologie particulière. Trois critères sont proposés : les degrés de structuration et de connectivité sont des mesures proprement ontologiques, mais le critère de couverture est établi relativement à une collection de documents. Ce dernier critère est

cependant défini de manière assez fruste[2] : il ne prend qu'imparfaitement en compte la dimension proprement linguistique des « étiquettes de concepts ».

Brewster *et al.* (2004) vont plus loin. Ils proposent d'évaluer les ontologies relativement à un corpus donné. La notion de couverture qu'ils proposent est plus riche que la précédente. Elle repose sur le nombre de termes en corpus qui correspondent à des concepts de l'ontologie, une fois effectués un calcul de variation pour reconnaître des formes de termes non canoniques et une expansion sémantique pour autoriser une adéquation à différents niveaux de généralité. À partir de là, une ontologie est évaluée en fonction du nombre de concepts qui trouvent leur contrepartie en corpus. Ce deuxième travail prend davantage en compte la nature linguistique des réalisations lexicales des concepts en corpus (notion de variation, quasisynonymie entre un hyperonyme et son hyponyme) mais il est centré sur l'évaluation et la cohérence interne d'une ontologie alors que notre objectif est plutôt de guider le choix d'une ressource pour un corpus donné, ce qui confère un autre rôle à la notion de couverture et impose de la définir plus précisément.

Sur le plan lexical, la question de la couverture n'a guère été étudiée[3]. De manière intuitive, on tend à préférer des ressources de grande taille (en nombre d'entrées), avec l'idée qu'elles sont soit plus complètes sur un domaine restreint soit plus génériques et moins liées à un domaine particulier. (Nirenburg *et al.*, 1996) critique ce présupposé en soulignant que la taille de la ressource donne une vue très partielle de sa couverture. Dans ce travail, les auteurs cherchent cependant à apprécier la qualité intrinsèque de la ressource alors que nous défendons l'idée qu'une ressource n'a pas de valeur propre et qu'elle n'a de valeur que par les utilisations qui peuvent en être faites. Au total, la question du choix de la ressource étant donné un corpus a moins retenu l'attention que la question ultérieure : une fois cette ressource choisie, comment l'adapter à ce corpus (Basili *et al.*, 1998) ?

La statistique lexicale a souligné depuis ses débuts (Muller, 1977 ; Manning, Schütze, 1999) qu'il existe une relation fonctionnelle entre une ressource (un vocabulaire) et un corpus mais elle n'a pas abordé le problème des unités polylexicales que contiennent les terminologies.

3 Proposition de métriques

Afin d'apprécier l'adéquation d'une ressource à un corpus, nous proposons différentes mesures. Il s'agit de caractériser la couverture de la ressource ainsi que son degré de spécialisation.

3.1 Remarques terminologiques

Nous posons les définitions suivantes :

- Le *texte T* du corpus est un ensemble ordonné de mots[4]. Les mots sont définis par leur forme graphique et repérés par leur position dans le texte.

---

[2] «Coverage is measured by the number of labels for classes and properties that can be matched in the document».

[3] La notion de couverture lexicale n'est pas définie dans les ouvrages de statistique linguistique (Oakes, 1998). Quand la question est abordée (Manning, Schütze, 1999, p. 130), c'est uniquement pour apprécier le nombre de mots inconnus dans un texte.

[4] La notion de « mot » est difficile à définir. Nous considérons ici comme mots les unités résultant d'une segmentation du texte, étant donné un algorithme de segmentation clairement défini. Dans les exemples présentés ici, tous les caractères d'espacement et de ponctuation sont considérés comme séparateurs de mots.

- Le *vocabulaire V* du corpus est l'ensemble des vocables, *i.e.* l'ensemble des mots différents du corpus. Les vocables sont des unités monolexicales.

- Le *lexique L* de la ressource est l'ensemble des lexies ou entrées lexicales de la ressource[5], qu'elles soient composées de un ou plusieurs mots, spécialisées ou non.

Les vocables du corpus et les lexies étant de natures différentes, on ne peut pas comparer directement le vocabulaire et le lexique. Pour établir cette comparaison, nous considérons la « partie utile » de la ressource et sa « décomposition », ainsi que leurs complémentaires, définis de la manière suivante (fig. 1) :

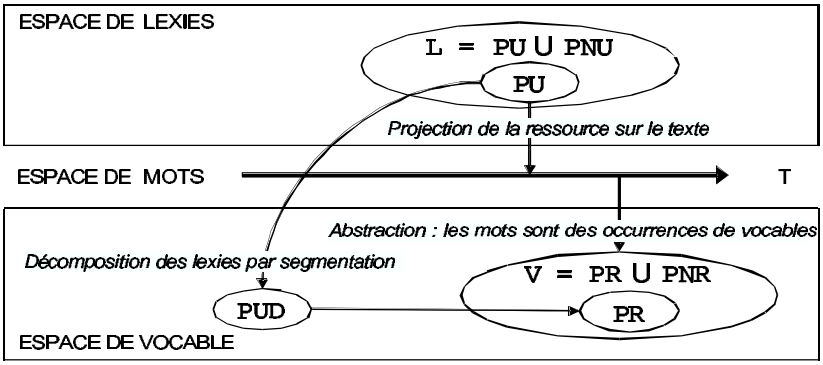


Figure 1 : Construction des ensembles de référence

- La *partie utile de la ressource PU* est l'ensemble des lexies de la ressource qui apparaissent dans le corpus. C'est un sous-ensemble de *L*.

- La *partie utile décomposée PUD* est l'ensemble de tous les vocables des lexies de *PU*. Elle est obtenue par décomposition en vocables élémentaires des lexies de *PU*. En supposant que cette décomposition est faite selon les mêmes règles qui ont permis de segmenter le corpus, cet ensemble de vocables *PUD* correspond aussi à la partie du vocabulaire du corpus qui est reconnue (*PR*) par la ressource. On a donc *PUD=PR*, où *PR* est un sous-ensemble de *V*.

- La *partie inutile de la ressource PNU* est l'ensemble des lexies qui n'ont pas d'occurrence dans le corpus. C'est le complémentaire de *PU* par rapport à *L*.

- La *partie inconnue du vocabulaire du corpus PNR* est l'ensemble des vocables de *V* non reconnus par la ressource. C'est le complémentaire de *PN* par rapport à *V*.

3.2 Mesures

Les métriques que nous proposons pour apprécier l'adéquation d'une ressource terminologique à un corpus sont définies comme des rapports entre les différents ensembles définis ci-dessus. On peut distinguer les mesures qui portent sur les formes et celles qui portent sur les occurrences.

La première mesure permet d'apprécier le degré de spécialité de la ressource par rapport à un corpus. La *contribution* (*Contr*) est la proportion de lexies du lexique qui figurent en corpus. Elle est définie par la formule ci-dessous. Nous désignons par *surplus* (*Surpl*) la proportion de lexies « inutiles ». On retrouve ici la notion d'excès de ressource introduite par (Brewster *et al.*, 2004). La contribution est forte si beaucoup des lexies de la ressource se retrouvent en corpus et donc si le domaine de spécialité de la ressource correspond bien à celui du corpus. À l'inverse,

[5] Comme nous l'avons souligné plus haut, nous ne considérons pas à ce stade les autres informations sémantiques apportées par la ressource.

un surplus élevé indique que la ressource est relativement générique et donc potentiellement utile pour des corpus variés. Ces mesures étant indépendantes de la taille de la ressource, on peut comparer les contributions de ressources très différentes.

$Contr = |PU| / |L|$ $Surpl = 1 - Contr = |PNU| / |L|$ où $|X|$ représente le cardinal de X

Une autre mesure permet d'apprécier dans quelle mesure la ressource « couvre » le vocabulaire du corpus. Pour avoir des ensembles comparables, il faut comparer la partie reconnue du vocabulaire et le vocabulaire dans son ensemble. Les deux mesures duales de la *reconnaissance (Rec)* et de l'*ignorance (Ign)* sont définies ci-dessous. La reconnaissance est la proportion des lexies décomposées reconnues en corpus par rapport au nombre total de vocables du corpus. La reconnaissance augmente 1) si on trouve dans le lexique les termes spécialisés employés dans le corpus mais aussi 2) quand la ressource comporte beaucoup de mots de la langue générale comme par exemple les mots grammaticaux. Seule la confrontation des différentes mesures permet de se faire une idée plus précise du comportement d'une ressource. Dans le cas 2, la forte reconnaissance tend à être associée à un surplus important. Une forte reconnaissance combinée à une contribution élevée indique une ressource spécifique bien adaptée au corpus considéré.

$$Re\,c = |PR| / |V| = |PUD| / |V| \qquad Ign = 1 - Re\,c = |PNR| / |V|$$

Parler de « couverture » évoque l'idée d'un corpus tout ou partiellement « couvert » par la ressource. La couverture est donc calculée relativement au corpus plutôt qu'à son vocabulaire. Nous définissons la *couverture* (*Couv*) comme la proportion d'occurrences de mots correspondant à des vocables entrant dans les lexies de la partie utile de la ressource. Dans la formule ci-dessous, $freq_i$ représente le nombre d'occurrences d'une lexie i de *PU* non incluses dans une occurrence d'une autre lexie plus large et $longueur_i$ est la longueur de la lexie en nombre de mots. Dans le cas de termes enchâssés (p. ex. *système et système de fichiers*), seule l'occurrence du terme le plus large entre dans la mesure de fréquence. Cette mesure de couverture est indépendante de la taille du corpus, ce qui rend les mesures de couverture d'une ressource comparables même sur des corpus de taille différente.

La dernière mesure complète la mesure de couverture. C'est la *densité* (*Dens*), définie par la formule ci-dessous, où f_{PUD} est la fréquence moyenne des lexies de *PU* dans le corpus et f_V est la fréquence moyenne des vocables dans le corpus. C'est une mesure normalisée de la fréquence des lexies utiles en corpus. Pour avoir une mesure indépendante de la taille du corpus, la fréquence moyenne des lexies de *PU* est pondérée par la fréquence moyenne des vocables dans le corpus.

$$Couv = \sum_{i=1}^{PU} freq_i\, x\, longueur_i / |T| \qquad Dens = f_{PUD}/f_V$$

3.3 Exemple

À titre d'exemple, considérons la ressource et le texte suivants :

- L={*système*, *système de fichiers*}
- T=« Il a réparé le système de fichiers »

On a $|L|=2$ et $|T|=7$. Dans ce cas particulier, on a $|V|=|T|=7$. Toutes les unités du lexique se retrouvant en corpus, on a par ailleurs PU=L et PUD=PR={*système, de, fichiers*}.

On obtient donc les mesures suivantes : *Contr*=1, *Rec*=3/7, *Couv*=3/7. Notons que l'occurrence de la lexie *système* qui entre dans l'occurrence plus large de la lexie *système de fichier* n'est pas comptabilisée en tant que telle dans la couverture.

4 Résultats

4.1 Protocole expérimental

Nous avons testé ces métriques dans le cadre de projets de recherche et d'extraction d'information dans le domaine de la génomique. Ce type d'application spécialisée requiert en effet d'exploiter des ressources et le choix de/des ressource(s) à exploiter s'avère souvent délicat. Nous avons considéré différents corpus de génomique et différentes ressources terminologiques *a priori* assez bien adaptées au domaine d'application (Hamon, 2005). À des fins d'évaluation, nous avons complété ces données expérimentales par un autre corpus qui porte sur les plantes carnivores et qui relève d'un domaine un peu différent. Il faudra élargir cette expérimentation en prenant en compte un autre corpus extérieur au champ de la biologie et une ressource dite de « langue générale ».

Nous avons travaillé sur quatre corpus, tous du domaine de la biologie, mais différant les uns des autres par leur style et leurs caractéristiques lexicographiques (tableau 1). Le premier corpus (*Transcript*) est constitué de 2 209 résumés d'articles scientifiques issus de la base Medline[6] à partir de la requête « Bacillus subtilis transcription ». Le second corpus (*Transcript-932* ou *932*) a été construit à partir du premier, en sélectionnant 932 phrases dans lesquelles apparaissent deux noms de gènes. Le troisième corpus (*Drosophile-1199* ou *1199-droso*) (Pillet, 2000) est similaire au second. Il s'agit de 1 199 phrases extraites des résumés de Flybase[7], qui contiennent deux noms de gènes. Le quatrième corpus regroupe différents documents issus du web se rapportant aux plantes carnivores (*Carnivore*).

	Vocabulaire V	Taille T	Fréquence moyenne
Transcript	18 720	405 423	21,66
Transcript-932	3 305	29 848	9,03
Drosophile-1199	3 232	22 691	7,02
Carnivore	27 201	273 605	10,06

Le tableau 1. Caractéristiques lexicographiques des corpus

Pour étudier la couverture des ressources terminologiques, nous avons sélectionné cinq ressources spécialisées publiquement disponibles[8] : 1) les mots clés SwissProt (keywlist) utilisés pour indexer la base de séquençage des protéines, 2) Gene Ontology (GO) qui porte sur les différents types d'organismes vivants, 3) le MeSH qui est dédié à l'indexation de la base de données Medline et rassemble des termes très divers utilisés dans le domaine médical. Nous avons également retenu deux glossaires proposant une grande variété de termes : 4) le glossaire de biochimie et de biologie moléculaire (GlossBioch) qui comporte des termes courants et 5) le glossaire de terminologie de biologie moléculaire (GoMBT).

---

[6] www.ncbi.nlm.nih.gov

[7] Flybase est une base de données structurées et bibliographiques sur la drosophile : http://flybase.bio.indiana.edu/

[8] Ces ressources sont disponibles aux adresses suivantes :
keylist : ftp://ftp.expasy.org/databases/swiss-prot/release/keywlist.txt
GO : http://www.geneontology.org/, version téléchargée en septembre 2002
MeSH : http://www.nlm.nih.gov/mesh/meshhome.html (Medical subject headings, Library of Medicine)
GlossBioch : http://www.portlandpress.com/pcs/books/prod_det.dfm?product=1855780887
GoMBT : http://www.asheducationbook.org/cgi/content/full/2002/1/490

Ressources	MeSH	GO	keywlist	GlossBioch	GoBMT
Taille en nombre de lexies	89 949	16 736	2934	836	263

Tableau 2. Tailles comparées des différentes ressources

4.2 Analyse des résultats

Les calculs des différentes métriques pour les 5 ressources et les 4 corpus ci-dessus sont synthétisés dans les graphiques des figures 2 et 3.

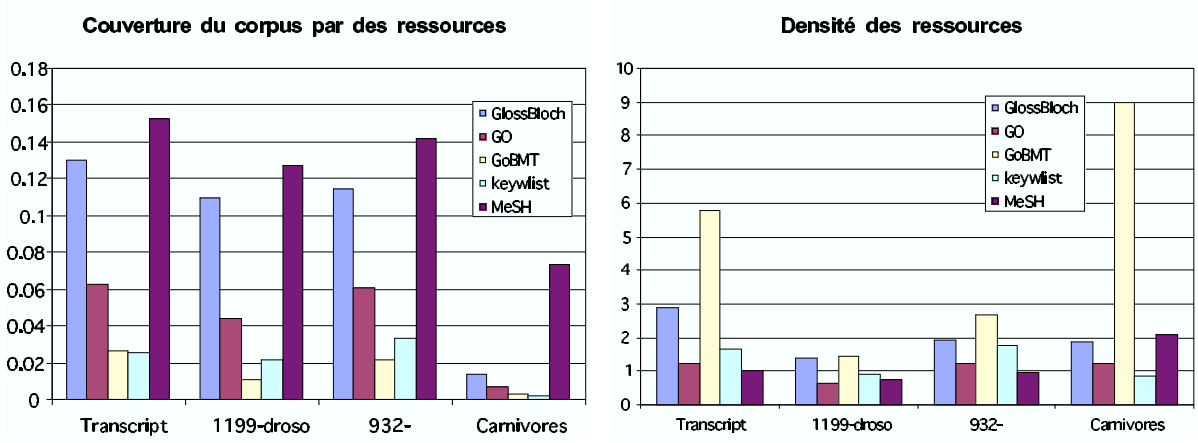


Figure 2. Mesures d'adéquation de différentes ressources à différents corpus : couverture et densité

Les mesures gomment l'effet de taille aussi bien sur les corpus que sur les ressources. Le glossaire GlossBioch a une couverture similaire à celle de MeSH qui comporte pourtant 50 fois plus de termes (fig. 2). Le comportement des ressources est comparable pour le corpus *Transcript* et son sous-corpus *Transcript*-932 (fig. 3). On peut donc envisager de sélectionner une ressource à partir d'un sous-corpus sans chercher à projeter la ressource sur l'intégralité du corpus, ce qui facilite les expérimentations.

La contribution fait exception cependant. Elle est à la fois sensible à la taille de la ressource et à celle du corpus : on remarque qu'elle est moindre pour un petit corpus (GloBioch pour *transcript-932*) et pour les ressources volumineuses (MESH pour *Transcript*). Malgré cette sensibilité aux effets de taille, c'est une mesure intéressante : une forte contribution pour une petite ressource est un bon indicateur de pertinence (*cf.* GoBMT et keywlist pour *Transcript*).

La troisième remarque concerne les deux mesures de reconnaissance et de couverture qui paraissent assez bien corrélées. On note une grande stabilité dans le sens et l'ampleur de leur écart : un corpus est d'autant mieux couvert que son vocabulaire est reconnu. C'est donc l'absence de corrélation qui est significative. Nos expériences montrent par exemple que le glossaire GlossBioch a une couverture nettement supérieure à celle de GO sur *Transcript*, pour une reconnaissance similaire. C'est le signe que GlossBioch reflète mieux la langue de spécialité du corpus *Transcript*, en dépit de sa taille modeste (fig. 3), et la preuve que la taille des ressources n'est pas un critère suffisant. Dans ce cas particulier, les mesures font apparaître un comportement des ressources contraire aux intuitions initiales des biologistes qui recommandaient à tort d'utiliser GO.

Le dernier point porte sur la densité. Elle permet d'apprécier la fréquence des lexies en corpus. De manière surprenante, la plus forte densité s'observe pour une petite ressource très spécialisée (glossaire GoBMT, fig. 2) et pour le corpus le plus différent thématiquement (*Carnivore*). Seule l'analyse détaillée des lexies de la partie utile du glossaire permet de comprendre ce résultat contre-intuitif. Moins de 10% des lexies figurent dans le corpus mais ces lexies ont de fortes fréquences. On trouve notamment *can* (676 occ.), *fish* (121 occ.) *tel* (8 occ), tous les trois décrits dans la ressource comme des noms de gènes. Ce sont des mots ambigus reconnus à tort comme

noms de gènes dans le corpus *Carnivore*. Une forte densité peut ainsi aussi bien refléter une bonne adéquation de la ressource en termes de spécialisation que des phénomènes d'ambiguïté. Une simple mesure de fréquence pondérée n'apparaît donc pas suffisamment éclairante. Il faudrait sans doute considérer le profil lexical des lexies de la partie utile de la ressource par rapport à l'ensemble des vocables du corpus pour pouvoir prédire la nature sémantique de la couverture. Ce profil devrait permettre d'apprécier la dispersion des fréquences et donc de mieux repérer des correspondances artificielles entre certains termes spécialisés et des occurrences de mots courants.

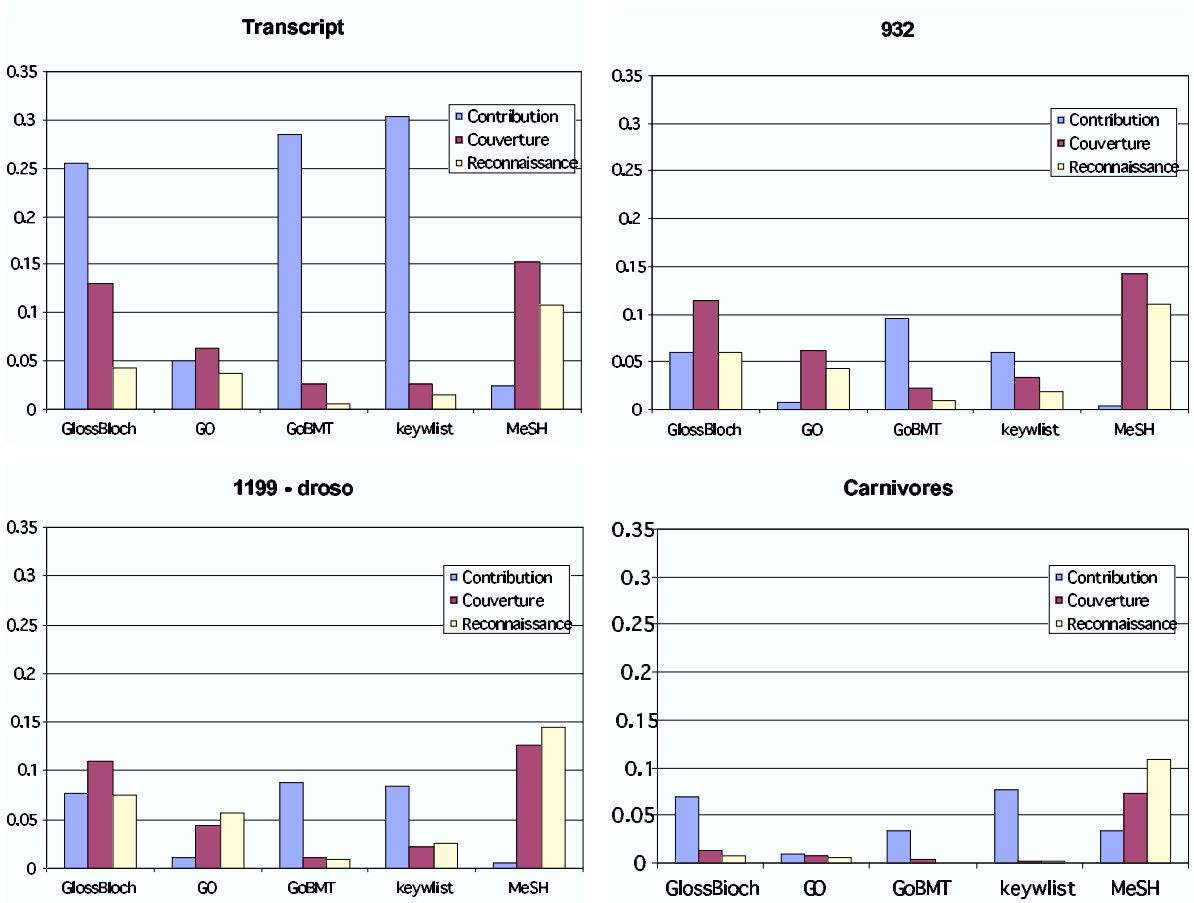


Figure 3. Mesures de contribution, couverture et reconnaissance de 5 ressources sur 4 corpus : *Transcript, Transcript-932 (932), Drosophile (1199-droso)* et *Carnivore*

5 Conclusion et perspectives

Pour permettre de caractériser avec une certaine fiabilité et une certaine reproductibilité le comportement d'une ressource lexicale pour un corpus donné, nous avons défini et testé un ensemble de métriques qui donne une idée de la « couverture », notion vague mais très couramment utilisée qui prend de l'importance avec l'augmentation du nombre de ressources disponibles. Ces métriques ne peuvent prétendre suppléer une analyse précise de l'apport d'une ressource : elles visent à éclairer le choix des ressources et des traitements à mettre en œuvre. Les expériences que nous avons menées montrent l'intérêt de ce type de métriques mais nous avons également souligné les limites des mesures proposées. Il faudrait définir une mesure de densité plus riche que nous ne l'avons fait et, pour compléter l'image globale de couverture que nous cherchons à construire, tenir compte de la répartition des occurrences des lexies de la ressource. La notion de couverture lexicale telle qu'elle est définie ici doit par ailleurs être étendue pour prendre en compte les variantes de lexies, leurs types sémantiques et même leurs relations sémantiques.

Références

BUITELAAR P., EIGNER T., DECLERCK T. (2004), OntoSelect: A Dynamic Ontology Library with Support for Ontology Selection, In *Proc. of the Demo Session at the Int. Semantic Web Conf.*, Hiroshima, Japan, Nov. 2004.

BREWSTER, C., Alani, H., DASMAHAPATRA, S. and WILKS, Y. (2004), Data Driven Ontology Evaluation. In *Proc. Of the Int. Conf. on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal.

BASILI R., PAZIENZA M.T., STVENSON M., VELARDI P., VINDIGNI M., WILKS Y. (1998), An Empirical Approach to lexical Tuning, In *Proc. of the Workshop on Adaptating Lexical and Corpus Ressources to Sublanguages and Applications (First Int. Conf. on Language Resources and Evaluation LREC 1998)*, P. VELARDI (ed.), May, Grenada.

CHARLET J., BACHIMONT B., BOUAUD J., ZWEIGENBAUM P. (1996), Ontologie et réutilisabilité : expérience et discussion, in *Acquisition et Ingénierie des Connaissance*, N. Aussenac , P. Laublet and C. Reynaud (éd.), pp. 69-87, Cépaduès-Editions, Toulouse.

HAMON H. (2005), Indexing specialized documents : are terminological resources sufficient ?, in *Actes des 6èmes journées Terminologie et Intelligence Artificielle (TIA 2005)*, pp. 71-82, Rouen.

HOVY E. (2001), Comparing sets of semantic relations in ontologies. In *Semantics of Relationships*, R. GREEN, C.A. BEAN and S.H. MYAENG (eds.), chapter 6, Kluwer , Dordrecht, NL.

PILLET V. (2000), *Méthodologie d'extraction automatique d'information à partir de la littérature en science en vue d'alimenter un nouveau système d'information. Application à la génétique moléculaire pour l'extraction de données sur les interactions*. Thèse doctorat, Aix-Marseille III.

MANNING C., SCHÜTZE H. (1999). *Foundations of Statistical Natural Language Processing*, The MIT Press.

MULLER C. (1977), *Principes et méthodes de statistique lexicale*, Hachette Université, Paris.

NAZARENKO A. (2005). Sur quelle sémantique reposent les méthodes automatiquesd'accès au contenu textuel ? *Sémantique et corpus*, A. CONDAMINES (coord.), ch. 6, pp. 211-244, Hermès/Lavoisier.

NEDELLEC C., NAZARENKO A. (2005), *Ontology and Information Extraction : a necessary symbiosis*, in Ontology Learning and Population, P. Buitelaar, P. Cimiano, B. Magnini (eds), IOS (to appear).

NIRENBURG S., MAHESH K. and BEALE S. (1996), Measuring semantic coverage, in *Proc. of the 16th Conf. on Computational Linguistics (COLING'96)*, Copenhagen Denmark, ACL, pp. 83-88.

Construction automatique de classes de sélection distributionnelle

Guillaume Jacquet, Fabienne Venant

LaTTICe – CNRS UMR 8094
Langues, Textes, Traitements Informatiques et Cognition
Ecole Normale Supérieure
1 rue Maurice Arnoux
F-92120 Montrouge
guillaume.jacquet@ens.fr
fabienne.venant@ens.fr

Mots-clés : classes de sélection distributionnelle, espace distributionnel, désambiguïsation, corpus, contexte
Keywords : semantic classes, distributional space, disambiguation, corpus, context

Résumé

Cette étude se place dans le cadre général de la désambiguïsation automatique du sens d'un verbe dans un énoncé donné. Notre méthode de désambiguïsation prend en compte la construction du verbe, c'est-à-dire l'influence des éléments lexicaux et syntaxiques présents dans l'énoncé (cotexte). Nous cherchons maintenant à finaliser cette méthode en tenant compte des caractéristiques sémantiques du cotexte. Pour ce faire nous associons au corpus un espace distributionnel continu dans lequel nous construisons et visualisons des classes distributionnelles. La singularité de ces classes est qu'elles sont calculées à la volée. Elles dépendent donc non seulement du corpus mais aussi du contexte étudié. Nous présentons ici notre méthode de calcul de classes ainsi que les premiers résultats obtenus.

Abstract

This study is placed within the general framework of the automatic verb sense disambiguation. To assign a meaning to a verb, we take into account the construction of the verb, i.e. the other lexical and syntactic units within the utterance (co-text). We now seek to finalize our method by taking into account the semantic features of this co-text. We associate with the corpus a continuous distributional space in which we build and visualize distributional classes. The singularity of these classes is that they are computed "on line" for disambiguate a given context in the given corpus. They thus depend not only on the corpus but also on the studied context. We present here our method of computation of classes and first results obtained.

1 Introduction

L'objet de cette étude est la désambiguïsation automatique du sens d'un mot en contexte. Elle s'inscrit dans le cadre théorique de la construction dynamique du sens, proposé par Victorri et Fuchs (1996) : on associe à chaque unité polysémique un espace sémantique. Le sens de l'unité dans un énoncé donné correspond à une région plus ou moins étendue de cet espace, déterminée par l'interaction dynamique de toutes les unités présentes dans l'énoncé. Ploux et Victorri ont développé un logiciel, Visusyn, permettant de construire automatiquement l'espace sémantique associé à une unité lexicale à partir d'un dictionnaire de synonymes (Ploux, Victorri, 1998). Ce logiciel a ensuite été étendu de façon à pouvoir prendre en compte des données distributionnelles issues d'un corpus dans une tâche de désambiguïsation automatique (François et al, 2003 ; Venant, 2004). Une réflexion sur le rôle de la syntaxe dans la construction du sens, s'appuyant notamment sur les travaux de Goldberg (1995) et Kay (2000), a permis d'enrichir le logiciel d'un nouveau module. Considérant que les constructions syntaxiques sont porteuses d'un sens intrinsèque et qu'à ce titre elles font partie intégrante du cotexte, ce module permet d'associer à chaque construction syntaxique une zone dans l'espace sémantique de l'unité étudiée (Jacquet, 2004). Nous cherchons maintenant à finaliser notre méthode de désambiguïsation en tenant compte des caractéristiques sémantiques du cotexte lexical. Il s'agit d'associer une zone de l'espace sémantique non plus à chaque unité rencontrée en contexte mais à des classes d'unités. Ceci permettra de traiter des noms propres ou des cooccurrences rares dans le corpus. Par exemple, le sens de *jouer du luth* sera associé à *jouer de* (*le luth, la guitare, le piano,* ....). On pourra alors, dans l'espace sémantique du verbe *jouer*, définir une zone correspondant à la classe (*luth, guitare, piano,...*), distincte de celle associée à une autre classe comme (*charme, prestance, influence,* ...) pour *jouer de son charme*. Ces classes sont déterminées automatiquement à partir d'un corpus. Leur singularité est qu'elles sont dépendantes de l'énoncé étudié. Nous voulons construire non pas une ontologie de la langue française mais des classes distributionnelles avec pertinence d'emploi. Pour ce faire nous associons au corpus un espace distributionnel continu dans lequel nous construisons et visualisons les classes de sélection distributionnelle associées au contexte étudié.

2 Des classes de sélection distributionnelle

La technique que nous utilisons s'inscrit dans le cadre bien connu de l'analyse distributionnelle « à la Harris ». Elle est exploitée depuis longtemps dans la communauté du TAL pour la construction de bases de connaissances ou de ressources terminologiques à partir de textes (Frérot, 2003 ; Habert et Nazarenko, 1996 ; Fleury, 1998 ; Aussenac-Gilles et al, 2000, Pantel et Lin, 2001 ; ...). Notre méthode est entièrement automatique. Elle ne fait appel à aucune modélisation préalable de connaissances sémantiques sur le corpus et utilise des rapprochements de mots sur la base de contextes syntaxiques partagés. En tout cela elle se rapproche des travaux de Greffenstette (1994). Les contextes nous sont fournis par l'analyseur Syntex (Bourigault et Fabre, 2000). Comme le précise D. Bourigault : « Là où G. Greffenstette se contente volontairement d'une analyse syntaxique relativement rudimentaire, réalisée par l'analyseur Sextan, nous avons fait le choix d'une analyse, certes encore partielle, mais plus large et plus précise, réalisée par Syntex. De ce fait, les procédures statistiques d'analyse distributionnelle de Greffenstette ne concernent que des mots simples, alors que nous pouvons prendre en compte des entités complexes (contextes ou termes) », cela nous permet de prendre en compte des distinctions plus fines, de créer des classes plus riches en information sémantique et donc plus efficaces dans leur apport à la désambiguïsation automatique. Notre travail est à rapprocher de celui de Habert et al (2004). Nous travaillons nous aussi à partir des rapports de dépendance syntaxique élémentaire entre un contexte et les

mots pleins qu'il régit ou qui le régissent et nous considérons les mots comme des points dans l'espace à n dimensions des contextes (que nous appelons l'espace distributionnel). Nous poursuivons cependant des objectifs différents. Nous ne cherchons pas à créer des classes de mots ayant le même sens mais des classes de mots dont le comportement sémantique influence de la même façon un contexte donné. Autrement dit si nous voulons trouver la classe de *luth (guitare, piano,...)* ce n'est pas pour caractériser le sens de *luth* mais pour désambiguïser *jouer* dans *jouer du luth*. Nous ne cherchons pas non plus à « faire parler le corpus dans sa globalité » comme le font Aussenac-Gilles et al (2000). Les classes qu'ils construisent se constituent en navigant autour d'éléments saillants ou prototypiques et leur permettent d'obtenir une image sémantique du corpus. Nous nous intéressons au contraire à des mots relativement peu fréquents, et qui ne représentent donc pas une ligne de force du corpus, pour rechercher dans leur classe sémantique des mots plus fréquents et dont l'apport à la désambiguïsation automatique sera plus pertinent. Certes les classes obtenues rendent compte de l'information sémantique présente dans le corpus mais de façon mouvante (Habert et al, 1999). Chaque interrogation concerne un contexte et un mot différents et donne lieu à des regroupements différents au sein de l'espace distributionnel. Nos classes s'apparentent plutôt aux classes d'objet décrites par Gaston Gross (2004) : « tout changement de sens d'un prédicat est corrélé à un changement de son schéma d'arguments. Soit la phrase *Vous suivrez ce chemin*. Si on remplace l'objet *chemin* par des substantifs comme *route, rue, voie, sentier* le verbe *suivre* garde le même sens. On regroupera ces mots sous le terme générique de <voies>. Si en revanche, on remplace le mot *chemin* par *cours*, alors on a affaire à un autre emploi et le substantif *cours* peut être remplacé par *séminaire, stage, formation, cycle d'étude*, etc., qu'on rangera sous le classifieur <enseignement> ». Nous partageons avec Gross l'idée que « la mise au point du sens exige que l'on soit à même de préciser la nature sémantique des arguments que prend un emploi donné de prédicat » mais la différence entre les classes que nous cherchons et les classes d'objets de Gross c'est que nous ne cherchons pas à établir des classes en langue. Nos classes dépendent du contexte et surtout du corpus étudié. Gross cherche à créer des classes pouvant figurer dans un dictionnaire, c'est à dire calculées une fois pour toutes sur le lexique et indépendantes du corpus étudié. Nous proposons quelque chose de plus souple. Nos classes sont calculées en ligne pour désambiguïser un contexte dans un corpus donné. Elles ne sont valables que pour ce contexte même s'il peut y avoir des recouvrements. Elles ne sont pas nécessairement générales ni référentiables par un classifieur conceptuel comme <enseignement>. Elles caractérisent un comportement sémantique au sein d'un corpus donné plutôt qu'une notion et ne sont absolument pas hiérarchisables. L'intérêt de travailler à partir d'un contexte particulier est de limiter le nombre d'éléments à classer. Lorsqu'on étudie par exemple les compléments d'un verbe donné, on ne cherche pas à classer tous les noms de la langue française mais seulement les noms pertinents dans le contexte de ce verbe. Les classes sont obtenues plus facilement et sont plus significatives que des classes construites sur la globalité du lexique.

3 Des classes de sélection distributionnelle : pourquoi ?

L'algorithme utilisé dans Visusyn repose sur l'analyse d'un graphe de synonymie. Dans ce cadre on considère qu'un sens possible pour un mot est donné par une clique de synonymes de ce mot, c'est à dire un ensemble de synonymes du mot, tous synonymes entre eux, le plus grand possible. Au stade actuel de son développement, Visusyn est capable, étant donné une construction verbale, de déterminer le sens le plus probablement pris par le verbe dans cette construction. Considérons par exemple les énoncés suivants :

1) *jouer la fille sérieuse*
2) *jouer avec sa fille*

En considérant les têtes nominales de compléments (*fille*) d'une part, et la construction syntaxique (V+SN / V+SP (avec +SN)) d'autre part, Visusyn calcule que le sens le plus probable de *jouer* est *interpréter, incarner* dans l'énoncé (1) et *s'amuser, plaisanter* dans (2). Ce modèle est opérationnel et en cours d'évaluation (Jacquet, à paraître) mais on sait d'ores et déjà qu'il échoue sur des énoncés du type :

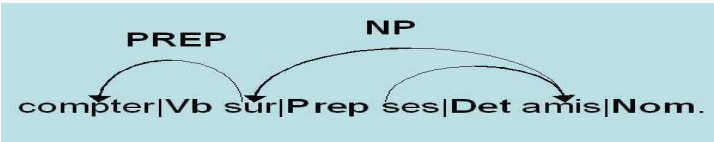
3) *Jouer du luth*
4) *Jouer à Wimbledon*

On se heurte ici à un double problème. Nous avons à faire à des compléments très peu représentés dans le corpus de référence. Or notre calcul repose sur l'utilisation des fréquences de cooccurrence de *luth* ou *Wimbledon* avec chacun des synonymes de *jouer*. Si ces fréquences sont trop faibles, le résultat du calcul est peu fiable. La première idée a été de remplacer *luth* par l'ensemble de ses synonymes, et de calculer leurs fréquences de cooccurrence avec chaque synonyme de *jouer*. Le problème est que les synonymes de *luth* sont trop peu nombreux et trop peu fréquents dans le corpus pour que le calcul soit efficace. Quand à *Wimbledon*, comme la plupart des noms propres, il ne possède aucun synonyme. Nous avons donc cherché à pallier ce manque d'informations quantitatives en fournissant à notre système des informations sémantiques sur les mots en question. Si nous pouvons associer à *luth* un ensemble de mots représentatifs des instruments de musique (*luth, guitare, piano, violon, etc*), nous retombons alors sur des énoncés interprétables par Visusyn et nous sortons de l'impasse. L'idée n'est certes pas nouvelle mais l'originalité de notre travail réside dans le fait que les classes que nous voulons construire vont dépendre du contexte dans lequel le mot considéré est inséré. Par exemple pour le mot *luth*, dans le contexte « jouer du », la classe qu'on cherche à construire est celle des instruments de musique mais si on s'intéresse à l'énoncé *poser un luth*, la classe construite pour *luth* correspondra plutôt à une classe générale d'objets matériels. Le sens de *poser* dans *poser un luth* est en effet celui de *poser* dans *poser un objet* plutôt que celui de *poser* dans *poser ses congés* ou dans *poser une question*. Notre objectif, à terme, est de remplacer dans notre système de désambiguïsation les noms propres ou rares par leurs classes contextuelles et de retrouver ainsi le sens des verbes.

4 Des classes de sélection distributionnelle : comment ?

4.1 Données initiales

Nous travaillons sur un corpus constitué de tous les articles du journal Le Monde sur dix ans, soit 200 millions de mots[1]. L'analyseur syntaxique Syntex (Bourigault et Fabre, 2000) est utilisé pour extraire de ce corpus une liste de mots ou syntagmes[2], structurée par des relations de dépendance syntaxique. Par exemple l'énoncé « compter sur ses amis », sera analysé par Syntex de la manière suivante :



[1] Nous remercions Benoit Habert de nous avoir autorisés à travailler à partir de son corpus
[2] Syntex permet de considérer certains syntagmes tels que *chef d'état*, *groupe financier* ou *Parc des Princes* comme des unités à part entière.

A partir de cette analyse, Syntex nous fournit la liste des mots lemmatisés contenus dans le corpus avec leur fréquence ainsi que la liste des *triplets* {recteur ; relation ; régi} du corpus, avec leur fréquence. On a par exemple le triplet {compter(V) ; PREP_SUR ; ami(N)}[3] dont la fréquence est 13. Il y a 20 millions de triplets différents (20 125 540 très exactement). Nous appellerons contexte lexico-syntaxique (C.L.S.) le couple formé par un des mots du triplet et la relation syntaxique. Chacun des triplets va être séparé en un C.L.S. régi, un C.L.S. régissant, et deux mots. Le triplet {compter(V) ; PREP_SUR ; ami(N)} donnera ainsi deux C.L.S., « compter(V).PREP_SUR » présent 8860 fois dans le corpus et « PREP_SUR.ami(N) » présent 88 fois et deux mots *compter(V)* et *ami(N)* de fréquences respectives 81485 et 38856. Nous obtenons ainsi une liste de mots (ou syntagmes) et une liste de contextes lexico-syntaxiques munis de leurs fréquences respectives. Ces listes constituent nos données de départ.

4.2 Données filtrées :

Les listes ainsi obtenues constituent une base de données colossale difficilement exploitable en l'état. Pour des raisons de taille et surtout de fiabilité nous avons dû filtrer les informations qu'elle contient. Nous avons appliqué successivement les critères suivants : chaque mot et chaque C.L.S. doivent être présents au moins 100 fois dans le corpus et chaque triplet doit être présent au moins 10 fois dans le corpus.

Après filtrage le corpus contient 31417 mots et 61202 contextes. A partir de ces données, nous construisons l'espace multidimensionnel engendré par les C.L.S.. C'est ce que nous appelons l'espace distributionnel associé au corpus. Chaque mot y est représenté par un point. La coordonnée d'un mot M sur l'axe engendré par un contexte C est la fréquence relative du triplet formé par M et C. Cet espace est muni de la distance du Chi2 : soit n le nombre de mots, p le nombre de contextes, M_i et M_k des mots de coordonnées (x_i^j) et (x_k^j) alors

$$d(M_i,M_k)^2 = \sum_{j=1}^{p} \frac{1}{x_\bullet^j}\left(\frac{x_i^j}{x_i^\bullet} - \frac{x_k^j}{x_k^\bullet}\right)^2 \text{ où } x_\bullet^j = \sum_{i=1}^{n} x_i^j \text{ et } x_i^\bullet = \sum_{j=1}^{p} x_i^j$$

4.3 Etude d'un mot dans un contexte lexico-syntaxique donné

Soient les énoncés *descendre le Mont-blanc* et *descendre la Seine*. Imaginons que nous cherchons à désambiguïser le verbe *descendre*. Une des forces de notre méthode est qu'elle permet d'étudier des cooccurrences non présentes dans le corpus. Par exemple, on ne rencontre aucune occurrence de *Mont-blanc* ni *Seine* qui soit objet de *descendre*. Il est cependant possible d'étudier les mots *Mont-blanc* et *Seine* dans le C.L.S. « descendre.OBJ ». Nous allons d'abord chercher dans l'espace distributionnel tous les mots qui ont une coordonnée non nulle selon cette dimension, c'est à dire tous les mots du corpus filtré employés dans ce contexte. On ajoute à la liste des mots obtenus les mots *Mont-blanc* et *Seine*. Notons que l'on fait une recherche toutes catégories confondues et que l'ensemble recherché peut contenir aussi bien des adjectifs, des noms communs, des noms propres ou même des entités plus complexes.

Si cet ensemble contient plus de 100 mots, on ne prend que les 100 mots les plus proches (au sens du Chi2) de *Mont-blanc* dans l'espace distributionnel. Notons MOTS l'ensemble formé. On va ensuite recenser tous les contextes pour lesquels au moins un des éléments de MOTS a

[3] Pour les relations prépositionnelles, les deux triplets {compter(V) ; __ ; sur(Prep)} et {sur(Prep) ; __ ; ami(N)} sont fusionnés en un seul {compter(V) ; PREP_SUR ; ami(N)}

une coordonnée non nulle. Notons CONT l'union de tous ces contextes. Dans le cas de *Mont-blanc*, MOTS contient 24 mots et CONT contient 5762 contextes. Nous pouvons dans un premier temps visualiser l'ensemble MOTS grâce à une analyse factorielle des correspondances (AFC) qui nous fournit 10 axes de visualisation synthétisant le mieux l'information des 5762 contextes de CONT.

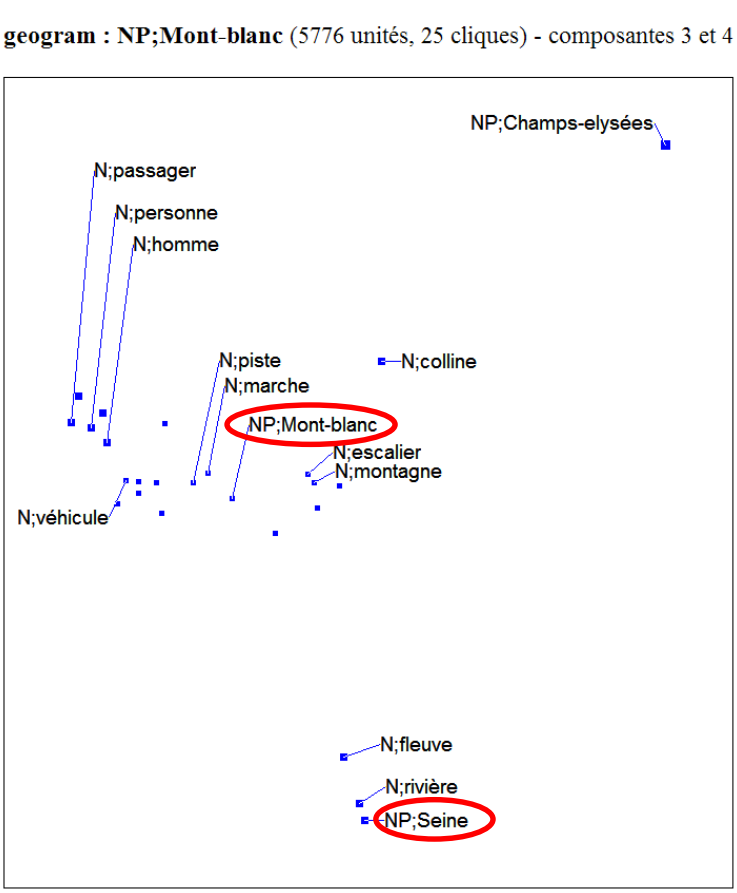


Figure 1 : distribution des mots compléments d'objet de *descendre* auxquels on a ajouté *Seine* et *Mont-blanc*

La Figure 1 fait clairement apparaître trois axes sémantiques organisant les compléments d'objet du verbe *descendre* : les personnes, les monts ou surfaces inclinées, et les cours d'eau. Notons que *descendre un avion* se trouve entre *descendre une personne* et *descendre les marches*. Il est remarquable que *Seine* et *Mont-blanc* qui ne sont pas des compléments d'objet de *descendre* dans le corpus étudié trouvent automatiquement leur place le long de l'axe qui leur correspond le mieux. *Seine* est placé à coté de *rivière* et *fleuve*, alors que *Mont-blanc* est entouré de *piste*, *montagne* et *escalier*. La visualisation proposée correspond aux composantes 3 et 4 de l'AFC. Autrement dit, l'information est contenue dans l'ensemble des composantes de l'AFC, et obtenir une visualisation lisible nécessite de parcourir les différentes composantes. C'est pourquoi la construction des classes distributionnelles va tenir compte des dix premières dimensions de l'AFC. On pondère lors de la clusterisation chaque axe par le coefficient $1-((x-1)^2/100)$ (où x est le numéro de l'axe). Nous avons choisi la clusterisation K-mean de Matlab (Seber, 1985). K-mean emploie un algorithme itératif en deux phases dont le but est de minimiser la somme des distances entre points et centre de gravité sur le nombre k de clusters. Pour un mot M et un contexte lexico-syntaxique C donnés, notre modèle propose un ensemble de classes de sélection distributionnelle employées avec le contexte C. Nous avons la possibilité d'ordonner ces classes en fonction de leur proximité avec le mot M, la première classe sémantique étant celle qui contient le mot M. Il arrive parfois que M soit l'unique mot de la classe, dans ce cas nous considérons que la classe la plus proche de M est la seconde. Ainsi dans le contexte « descendre.OBJ » la classe la plus proche de *Seine* est «N ;fleuve, N ;rivière, NP ;Seine » et la classe la plus proche de *Mont-blanc* est « N;montagne, N;piste ».

5 Evaluation des résultats

Nous proposons maintenant une première évaluation de notre système. Elle porte sur quatre contextes particulièrement ambigus en fonction des caractéristiques sémantiques de leurs

arguments : « descendre.OBJ », « jouer.PREP_à », « regarder.OBJ », « décider.SUJ ». Pour chaque contexte, nous calculons la classe sémantique la plus proche de quinze mots vedettes différents. Voici la liste des 60 cooccurences étudiées. Nous avons marqué d'un [P] celles qui sont présentes dans le corpus :

« **descendre.OBJ**» **:** NP;Seine, NP;Rhône, NP;Gange, NP;Danube, NP;Mississippi, NP;Chirac, NP;Jospin, NP;Pdg, NP;Kennedy, NP;Mont Blanc, NP ;Everest, NP;Pyrénées, NP;Alpes, NP; Broadway

« **jouer.PREP_à** » NP;Monopoly[P] , N;tarot, N;domino[P] , N;lego, NP;Paris[P] , NP;Washington, P;Wimbledon[P] P;Lyon, NP;Broadway, NP;New York[P], NP;Londres[P], NP;Marseille[P], NP;Lille, NP;Parc des princes[P]

« **décider.SUJ** » :NP;Paris[P], NP;France[P], NP;Washington[P], NP;Wimbledon, NP;Londres[P], NP;Clinton[P], NP;président, NP;Jospin[P], NP; Kennedy, NP;Onu[P], NP;Cgt[P], NP;Otan[P], NP;Rpr[P], NP ;PS[P], NP;Vivendi, NP;Renault[P]

« **regarder.OBJ** » **:** NP;Chirac, NP;Picasso, NP;Seine, NP;Arte, NP; Kennedy, NP;Internet, NP;Alpes, NP;Jospin, NP;Paris, NP;Lyon, NP;Lelouch, NP;Kubrick, NP;Etats-Unis, NP;Tf[P], NP;Tintin, N;Videocassette

L'évaluation, inspirée des travaux de Lin et Pantel (2001), consiste à juger si la classe proposée par le modèle est acceptable ou non. Huit juges, dont nous nous sommes naturellement exclus, vont donner une note de un à quatre, de la manière suivante. La classe est très mauvaise : 1 ; La classe est assez mauvaise : 2 ; La classe est assez bonne : 3 ; La classe est très bonne : 4.

Contexte « descendre.OBJ »		
Mots vedettes	**Cluster le plus proche du mot vedette**	**Ev.**
NP;Seine	N;fleuve, N;rivière, N;Seine	4
NP;Gange	N;fleuve, N;rivière	3,9
NP;Chirac	N;homme, N;personne	3,8
NP;Pdg	N;homme, N;personne	3,1
NP;Mont Blanc	N;montagne, N;piste	3,5
NP ;Pyrénées	N;rivière, N;fleuve	1,6
Contexte « jouer.PREP_à »		
Mots vedettes	**Cluster le plus proche du mot vedette**	**Ev.**
NP;Monopoly	N;billard, N;bridge, N;cache-cache, N;domino, N;ping-pong, N;poker, N;pétanque, N;souris, N;yo-yo, NP;Monopoly	3,4
NP;Lego	N;billard, N;bridge, N;cache-cache, N;domino, NP;Lego, N;ping-pong, N;poker, N;pétanque, N;yo-yo, NP;Monopoly	3,5
NP;Paris	NP ;New York, NP;Avignon, NP;Londres, NP;Marseille, NP;Paris	3,5
NP;Washington	NP ;New York, NP;Londres, NP;Paris, NP;Washington	3,6
NP;Wimbledon	N;basket, N;basket-ball, N;football, N;loterie, N;rugby, N;tennis, S;jeu vidéo	1,9
NP;Broadway	N;base-ball, N;cricket, N;foot, N;golf, N;loto, N;volley-ball, NP;Broadway	1,4
Contexte « décider.SUJ»		
Mots vedettes	**Cluster le plus proche du mot vedette**	**Ev.**
NP;Washington	NP;Etats-Unis, NP;Washington	3.7
NP;Wimbledon	NP;Europe, NP;France, NP;Italie	2
NP;président	NP;Pdg, S;directeur général, S;président de le conseil, S;secrétaire de état	3.7
NP;Chirac	NP;Chirac, NP;Clinton, NP;Eltsine, NP;Jospin, NP;Mitterrand, S;chef de le état, S;premier ministre, S;président de le république	3.4
NP;Otan	N;armée, N;force, N;police	3.4
Contexte « regarder.OBJ »		
Mot vedettes	**Cluster le plus proche du mot vedette**	**Ev.**
NP;Chirac	N;enfant, N;femme, N;gens, N;homme, N;personne, N;public, NP;Chirac	3.1
NP;Seine	N;mer	2.8
NP;Arte	NP;Arte, NP;Tf, S;chaîne de télévision	3.8
NP;Alpes	N;montagne	3.8

Figure 2 : mots évalués dans le contexte « descendre.OBJ »

Nous avons détaillé quelques notes pour chaque contexte dans la Figure 2. Dans le contexte « descendre.OBJ », on peut constater que les noms propres *Gange* et *Seine* ont dans leur première classe *fleuve* et *rivière*. Les noms propres *Chirac* et *PDG* ont dans leur première classe *homme* et *personne*. Dans le contexte « jouer.PREP_à », des noms propres comme *Lego* ou *Monopoly* sont rattachés à des classes de jeux alors que *Paris* et *Washington* sont rattachés à des noms de villes.

On peut noter des erreurs telles que « descendre.OBJ » avec *Pyrénées* qui donne des classes contenant *fleuve* et *rivière*. Ou encore « jouer.PREP_à » avec *Broadway*, qui donne une classe contenant des noms de sport.

	jouer.à	descendre.Obj	regarder.OBJ	décider.Suj	**TOTAL**
% 3 et 4	**81,63**	**80,61**	**60,71**	**85,71**	**76,90**
4	48,98	57,14	51,02	45,55	**50,00**
3	32,65	23,47	20,95	18,71	**26,90**
2	5,10	10,20	9,11	8,13	**14,52**
1	8,16	8,16	7,29	6,51	**7,14**
Non réponse	5,10	1,02	0,91	0,81	**1,43**
Moyenne des notes	3.3	3.2	2.9	3.4	**3.2**
Moyennes >3 (en %)	85.71	57.14	31.25	68.75	**60**

Figure 3 : Evaluation des classes obtenues pour 4 contextes, 60 cooccurrences, 8 juges.

La Figure 3 propose une synthèse des résultats sur les quatre contextes. Les résultats de l'évaluation sont tout à fait satisfaisants. 76.9 % des notes sont supérieures ou égales à 3 et correspondent donc à des jugements de classe assez ou très bonnes. On peut noter que dans la moitié des cas la note mise est un 4. La moyenne des notes sur les 4 contextes est de 3,2 ce qui veut dire que les classes proposées sont globalement attestées par les juges. Enfin 60% des classes ont sur l'ensemble des juges une note moyenne strictement supérieur à 3. On peut donc dire que 60% des classes proposées par notre système sont bonnes ou assez bonnes. 5 d'entre elles ont été jugées très bonnes par tous les juges. Ce sont les classes correspondant aux énoncés « descendre la Seine/ le Rhône/ le Danube » et « le PS/ la CGT décide ».

Il est par ailleurs intéressant d'étudier le comportement d'un même nom propre dans des contextes différents. Par exemple, *Wimbledon* peut être employé dans des énoncés tels que *revenir de Wimbledon*, *Wimbledon décide* ou *jouer à Wimbledon*. La Figure 4 présente pour chacun des contextes correspondants, les deux classes les plus proches de *Wimbledon*. On observe que les différentes facettes de *Wimbledon* sont mises en évidence en fonction du contexte. Le contexte «décide.SUJ » met en valeur *Wimbledon* en tant que zone géographique de décision. Le contexte « jouer.PREP_à» insiste sur la singularité de *Wimbledon* en tant que compétition de tennis. Les classes obtenues pour le contexte « revenir.PREP_de » peuvent être interprétées comme des lieux d'activité.

	Wimbledon dans le contexte « Wimbledon décide»	**Wimbledon** dans le contexte « jouer à Wimbledon »	**Wimbledon** dans le contexte « revenir de Wimbledon »
1er cluster	NP;Europe, NP;France, NP;Italie	N;basket, N;basket-ball, N;football, N;loterie, N;rugby, N;tennis, S;jeu vidéo	NP;Allemagne, NP;Etats-unis
2ème cluster	N;monde, N;pays, N;région, N;ville	N;base-ball, N;cricket, N;foot, N;golf, N;loto, N;volley-ball	N;guerre, N;mission, N;travail

Figure 4 : différentes classes de *Wimbledon* en fonction du contexte

6 Conclusions et perspectives

Nous avons présenté ici une méthode de construction de classes de sélection distributionnelle en contexte. Cette méthode, au vu des résultats préliminaires présentés ici, nous semble très prometteuse. Des expérimentations plus poussées sont cependant encore nécessaires pour finaliser notamment la méthode de clusterisation (détermination du nombre de clusters optimal, pondération des axes de l'AFC) ou le mode de filtrage du corpus. Il nous faudrait valider sur un plus grand nombre de contextes, sur différentes catégories de mots et en faisant appel à un plus grand nombre de juges. Nous devons aussi étudier la variation des classes obtenues lorsqu'on change le corpus de travail. Le type d'évaluation que nous avons commencé à mettre en place n'est qu'une étape. Le but à terme est d'utiliser les classes obtenues, dans notre système de désambiguïsation. Il nous faut donc mettre au point le module de Visusyn correspondant. Nous devrons déterminer quelle est la façon la plus pertinente de prendre en compte les classes de sélection distributionnelle dans Visusyn. On peut envisager d'injecter l'ensemble d'une classe, de lui associer une région dans l'espace sémantique du mot à désambiguïser, mais on peut aussi imaginer de travailler avec un représentant de la classe, une sorte de prototype. Se pose alors la question du statut du prototype : est ce le mot le plus proche du centre de gravité ou bien le centre de gravité lui-même ? Dans le dernier cas, le prototype ne serait pas un item lexical mais une sorte d'abstraction sans dénomination linguistique qu'on ne pourrait appréhender que par son profil d'utilisation (il se construit à 26% avec descendre.OBJ, à 10% avec revenir.DE, …) Ce sont les résultats de ce module, après évaluation de leur fiabilité, qui constitueront la véritable validation des classes obtenues. Les jugements des locuteurs humains se réfèrent en effet à une utilisation quotidienne du français. Ainsi certaines classes qui nous semblaient pertinentes dans une tâche de désambiguïsation sont rejetées massivement. C'est le cas par exemple de la classe formée pour *Wimbledon* dans le contexte « jouer.PREP_à ». Il nous semblait que le fait d'obtenir des noms de sport était satisfaisant puisque cela permet de donner à *jouer* le sens de « *pratiquer un sport* ». Or la moyenne des notes de cette classe n'est que de 1,9.

Une autre perspective de travail est de quitter le champ de la désambiguïsation pour celui de la catégorisation. Un de nos résultats importants est de montrer que la classe d'un même mot varie avec le contexte. Le fait par exemple que *Wimbledon* soit catégorisé parfois comme une zone géographique de décision et d'autre fois assimilé à un sport nous donne à penser que la langue n'est pas organisée selon un système hiérarchique de classes fixes, même si on accepte les recouvrements de classes. La catégorisation d'un mot se ferait plutôt « à la volée », de façon dynamique et en contexte.

Remerciements

Nous remercions les personnes qui ont accepté d'évaluer nos résultats, Sophie Prévost, Laure Sarda et Bernard Victorri pour leurs commentaires avisés, Benoît Habert pour la mise à disposition de son corpus, et plus particulièrement Didier Bourigault qui nous a fourni nos données de départ ainsi que de précieux conseils.

Références

AUSSENAC-GILLES N., BIÉBOW B., SZULMAN N. (2000), Revisiting Ontology Design: a method based on corpus analysis, Actes de 12th International Conference on Knowledge Engineering and Knowledge Management. Juan-Les-Pins

BOURIGAULT D. (2002), Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus, Actes de *TALN 2002*, Nancy, pp. 75-84

BOURIGAULT D., FABRE C. (2000), Approche linguistique pour l'analyse syntaxique de corpus, *Cahiers de Grammaires,* n° 25, Université Toulouse - Le Mirail, pp. 131-151.

FRANCOIS F., MANGUIN J.L., VICTORRI B. (2003), *La réduction de la polysémie adjectivale en cotexte nominal : une méthode de sémantique calculatoire*, Cahier du Crisco n°14

FLEURY S. (1998), Gaspar, un dispositif de TALN basé sur la programmation à Prototypes, Actes de TALN'98, Paris.

FREROT C., BOURIGAULT D, FABRE C. (2003), Marier procédures d'apprentissage endogènes et ressources exogènes dans un analyseur syntaxique de corpus – Le cas du rattachement verbal à distance de la préposition « de » , *T.A.L*, 44-3.

GOLDBERG A. (1995), *Constructions : a construction grammar approach to argument structure*, Chicago and London, University of Chicago Press.

GREFENSTETTE G (1994)., *Explorations in Automatic Thesaurus Discovery*, London, Kluwer Academic Publishers.

GROSS G (2004), Réflexions sur le traitement automatique des langues, Actes de *JADT 2004,* Vol. 1 545-556 .

HABERT B., ILLOUZ G., FOLCH H. (2004), Dégrouper les sens: pourquoi, comment?, Actes de *JADT 2004,* Vol. 1 565-576 .

HABERT B., FOLCH H. ET ILLOUZ G. (1999). Sortir des sens uniques : repérer les mots « mouvants » dans le domaine social. *Sémiotiques*, vol. (17). *Dépasser les sens iniques dans l'accès automatisé aux textes*, Habert B. (resp.) : 121-151.

HABERT B., NAZARENKO A. (1996)., La syntaxe comme marche-pied de l'acquisition des connaissances : bilan critique d'une expérience, *Journées sur l'acquisition des connaissances*, AFIA, Sète

JACQUET G. (2004), Using the construction grammar model to disambiguate polysemic verbs in French, Actes de *ICCG3 (International Conference on Construction Grammar),* Marseille.

JACQUET G. (à paraître), A model of disambiguation of polysemic verbs in French, *Constuctions, http://www.constructions-online.de/*

KAY P. (2000), Argument Structure Constructions and the Argument-Adjunct Distinction, Actes de *ICCG1*, Berkeley, p 30.

LIN D., PANTEL P. (2001), Induction of Semantic Classes from Natural Language Text, Actes de *ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2001*.

PLOUX S., VICTORRI B. (1998), Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes, *Traitement automatique des langues*, Vol. 39, n°1, pp.161-182.

SEBER G.A.F. (1984), *Multivariate Observations*, Wiley, New York. pp. 317-322.

VENANT F. (2004), Polysémie et calcul du sens, Actes de *JADT 2004,* Vol 2. 1146-1157.

VICTORRI B., FUCHS C. (1996), *La polysémie, construction dynamique du sens*, Paris, Hermès.

Sentiment Analysis for Issues Monitoring Using Linguistic Resources

Ecaterina Rascu (1), Kai Schirmer (2) and Johann Haller (1)

(1)(1) Institut für Angewandte Informationsforschung
Martin-Luther-Str. 14, D-66111 Saarbrücken
{kati;hans}@iai.uni-sb.de
(2) Schirmer Media Research
Lietzenburger Str. 90, D-10719 Berlin
kaischirmer@web.de

Mots-clés : Etude d'opinion, outil de veille économique, classification des opinions
Keywords: Sentiment analysis, issues monitoring system, fine-grained sentiment classification

Résumé L'identification et l'évaluation des avis, opinions ou jugements exprimés sur un sujet, une entreprise, ou un produit sont des tâches essentielles dans le domaine de l'analyse des médias. L'étude d'opinion est employée pour repérer de nouvelles tendances, mesurer le degré de satisfaction des clients ou pour alerter quand des tendances négatives risquent d'être défavorable à l'image de marque de l'entreprise. Dans cet article nous présentons un outil de veille économique qui permet de classifier très finement des documents publiés en ligne ainsi que d'identifier et d'évaluer les opinions exprimées dans des articles en ligne et des forums de discussions. Après la présentation des diverses composantes du système et des ressources linguistiques utilisées, nous décrivons en détail **SentA**, la composante d'étude d'opinions, et évaluons sa performance.

Abstract Sentiment analysis dealing with the identification and evaluation of opinions towards a topic, a company, or a product is an essential task within media analysis. It is used to study trends, determine the level of customer satisfaction, or warn immediately when unfavourable trends risk damaging the image of a company. In this paper we present an issues monitoring system which, besides text categorization, also performs an extensive sentiment analysis of online news and newsgroup postings. Input texts undergo a morpho-syntactic

analysis, are indexed using a thesaurus and are categorized into user-specific classes. During sentiment analysis, sentiment expressions are identified and subsequently associated with the established topics. After presenting the various components of the system and the linguistic resources used, we describe in detail **SentA**[1], its sentiment analysis component, and evaluate its performance.

1 Introduction

In recent years the tendency to exploit the huge amount of information available on the internet, especially in form of news articles or newsgroup postings for marketing and corporate communication purposes, has increased considerably. Text mining techniques have been developed to find relevant texts, classify them into meaningful clusters and also to extract specific information from them. One of the more sophisticated information extraction tasks, which has proved to be extremely valuable for companies, is the identification of sentiments towards a topic, a company, or a product in online news and newsgroup postings. This is essential for evaluating trends in public opinion, determining the level of customer satisfaction and taking preventive measures in case the level of dissatisfaction risks damaging the image of the company.

Various approaches to automatic sentiment evaluation have been proposed in order to efficiently deal with the large amount of available data as well as to reduce the high costs associated with the manual evaluation of such information. On the one hand, statistical text mining approaches, especially machine learning methods such as support vector machines for finding minimum cuts in graphs, have been used in order to classify texts as either positive or negative (Pang, Lee, 2004; Mullen, Collier, 2004). The approaches are known to be effective but also of limited use in areas where a high precision of results is needed. On the other hand, more elaborate approaches use linguistic resources in order to identify (1) expressions indicating opinions, (2) the polarity of the detected expressions as well as (3) the entity or topic to which the opinion refers. Typical examples of such systems are the ones developed by the CELI group (Dini, Mazzini, 2002) which analyzes customer opinions about mobile phones and identifies the polarity of opinions about specific parts or functions as well as the **Sentiment Analyzer** of IBM and the University of Texas (Yi et al., 2003) which extracts opinions about a given subject from online documents. However, obtaining a more fine-grained classification of opinions in terms of both granularity and specificity would be of much interest for many companies or agencies. For instance, within the **DeepThought** project experiments were carried out in which the CELI approach was extended by integrating deeper processing steps involving HPSG (Beermann et al., 2004). In this way more specific information concerning topics or the causes that brought about a particular judgment could be identified.

In this paper we present an issues monitoring system which performs a fine-grained analysis and classification of sentiments related to user-specific topics or aspects of them. The aim of our approach is to establish not only the polarity of the sentiments towards a specific topic identified in the text - as done in the above mentioned systems - but also to identify the

[1] **SentA**: Sentiment Analysis

degree of positive or negative orientation as well as quantify the degree of emotional implication carried by sentiment expressions. To this end, the system makes use of various linguistic resources. Thesauri, lexicons and specific patterns are used to identify topics in German online texts, detect sentiment expressions, establish sentiment orientation and strength of emotional involvement, as well as associate the detected sentiment expressions to the established topics.

In the following section we describe the issues monitoring system: we present its architecture, the resources used during the monitoring process, and illustrate its way of functioning by means of an example. Then, in Section 3, we focus on the various steps involved in the sentiment analysis with **SentA**, whereas in Section 4 we present an experiment carried out to evaluate the sentiment analysis process and discuss the results. Finally, we present our conclusions and point to future work.

2 The Issues Monitoring System

2.1 Architecture and Resources

In order to perform a fine-grained analysis and classification of sentiments related to user-specific topics, we have adapted an automatic indexing tool described in (Ripplinger, Schmidt, 2001). The modified architecture is shown in Figure 1.

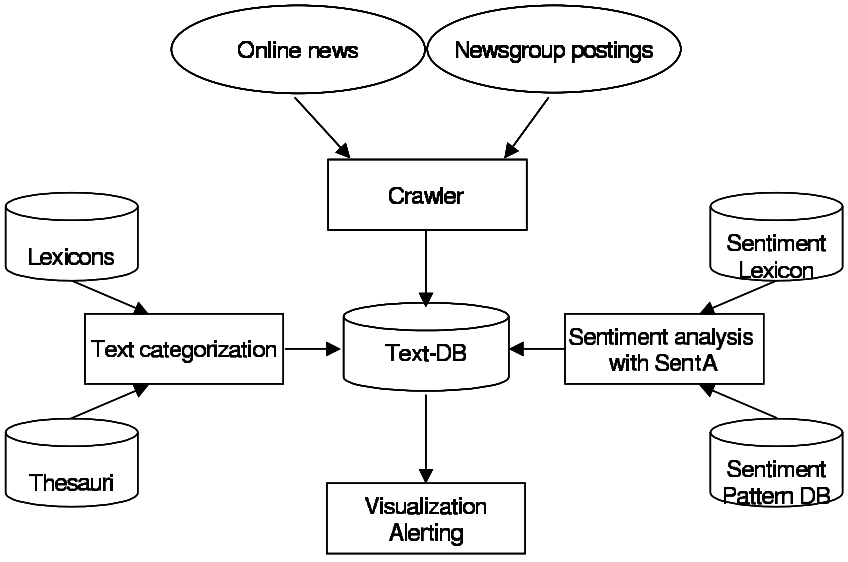


Figure 1 : Architecture of the issues monitoring system

The crawler downloads hourly news articles and newsgroup postings from specific websites. Then, in order to classify the retrieved texts, the input undergoes linguistic processing. The first step consists in the morphological analysis of the input with the package MPRO (Maas, 1996). It involves lemmatization, part-of-speech tagging and homograph analysis. Monolingual lexicons are used in order to assign information concerning word class, semantic

features, as well as derivation or decomposition in case of compound words. Then, shallow parsing is carried out to disambiguate the input and identify multiword terms and their respective variants. Eventually, a list of descriptors defining the topic of the text or aspects of it is generated by statistically evaluating the semantic load of the text and by weighing this information against a thesaurus. At this stage the text is classified using either a classification scheme provided by the user or the predefined codes assigned to terms in the lexicon (Ripplinger, Schmidt, 2001).

The next processing step is the analysis of the sentiments expressed in the input texts with **SentA**. Two types of resources are involved in this process. The first one is a sentiment lexicon in which expressions are encoded together with a manually assigned sentiment orientation. In some entries word stems are fully specified whereas in others regular expressions are used in order to ensure a broader coverage. The second resource is a sentiment pattern database including an ordered set of typical patterns involving sentiment expressions. Such patterns are used to establish if the sentiment orientation of an expression has been changed by certain features revealed during the analysis or by the immediate context. As will be shown in Section 3.2, sentiment orientation might be reinforced, attenuated, or even reversed depending on the context. Moreover, specific patterns are used to associate sentiment expressions either to the main topic or to aspects of it.

The last step is visualizing the analysis results. The sentiment values are cumulated for each user-specific issue. On the basis of such information, the issues monitoring system is able to show the development of attitudes or opinions towards a specific subject for a given period of time. The charts in Figure 2 show the results of the categorization and sentiment analysis processes. The chart on the left hand side (LHS) presents the three most negative issues over an interval of seven days (*7 Tage*) in online news. In the case shown here the three issues are *Sicherheit* (*safety*), *Pünktlichkeit* (*punctuality*), and *Mehdorn*, who is the chairman of the executive board at the German railway company *Deutsche Bahn*. The y-axis shows the sentiment value from neutral (*neut*) to negative (*neg*) on the LHS whereas on the right hand side (RHS) the same values are visualized through the changing colours of the column from yellow (neutral) to red (negative). A dot at the end of a line in the chart represents the position of an issue during the current week while the other end indicates the position of the same issue the week before. The x-axis shows the *media presence value* (*Medienpräsenz*) of an issue. This value is calculated by taking into account the length of the articles relevant for that issue and specific website frequency scores. The more an issue is located at the RHS the more spread it is among the public. The chart on the RHS illustrates the development of an issue, i.e. *railway* (*Bahn*) over a period of 12 months. The y-axis shows media presence values on the LHS as well as negative sentiment values with colours changing from yellow (neutral) to red (negative) on the RHS. The x-axis shows the 12-month time period[2].

[2] Since the system has been in use only since October, the RHS chart does not contain any information for the previous months.

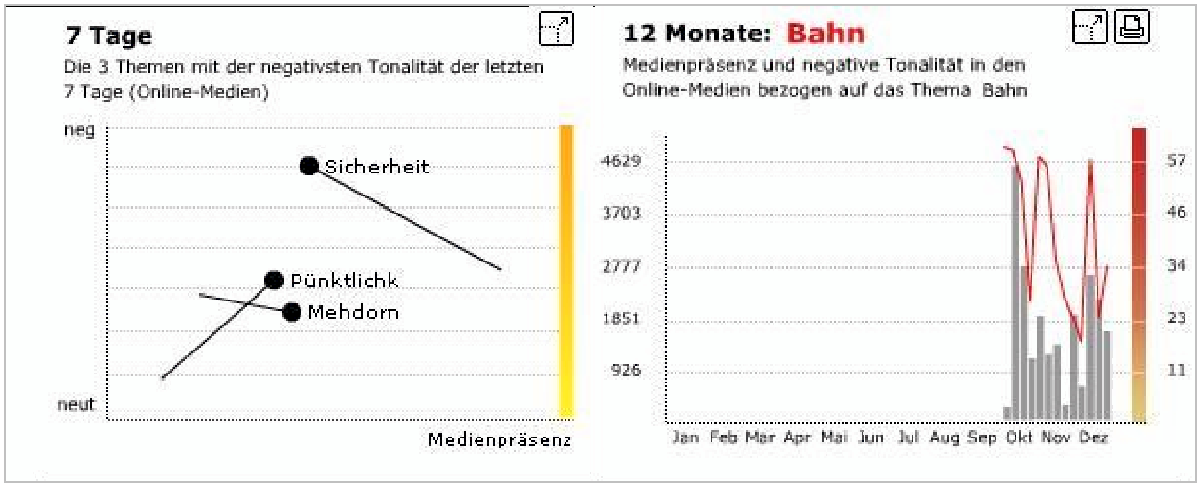


Figure 2 : Visualizing the results of the categorization and sentiment analysis processes

2.2 Example

Example (1)[3] illustrates the text categorization and the sentiment analysis processes for a given sentence. As described in Section 2.1 the descriptors *Fernverkehr* (*long distance traffic), Nahverkehr* (*short distance traffic)*, and *Preis* (*price*) are computed as being relevant to the main topic of the text. In the field *Oberbegriff* (*Hyperonym*), the hyperonyms of these descriptors are listed. Hyperonym information as well as the data concerning general categorization schemes listed in the *Special Field* are encoded in the thesaurus.

1. Die Preise im Fernverkehr sind seit Sonntag um 3,1 Prozent, im Nahverkehr sogar um durchschnittlich 3,6 Prozent teurer.
   (*Ticket prices in long distance traffic increased by 3.1 percent, in short distance traffic actually by an average of 3.6 percent.*)

Descriptors:	Fernverkehr[100]; Nahverkehr[100]; Preis[50];
Special Field:	n6021 (Personenbeförderung)[100]; n6000 (Landverkehr)[100];
Oberbegriff:	Verkehr[100]; Preis[50];
Opinions:	{ori=teurer,opinion=S-1+2,desc=fernverkehr;nahverkehr}

The field *Opinions* contains the results of the sentiment analysis process. The sentiment expression *teurer* (*more expensive*) is assigned a semantic orientation value (*opinion=S-1+2*) and is associated with the descriptors *Fernverkehr* and *Nahverkehr* (*desc=fernverkehr; nahverkehr*). Section 3 provides a more detailed description of the sentiment analysis process.

[3]In order to keep explanations simple, only those aspects of the output are presented that are relevant for our paper. In all examples the results of the sentiment analysis are given in form of attribute-value pairs grouped in a feature bundle. The attributes used for illustration have the following meaning: *ori* – surface form of the input word, *opinion* – sentiment orientation, *desc* – descriptor.

3 Sentiment Analysis with SentA

In this section we describe the various steps involved in sentiment analysis and illustrate them with examples. The first step of the analysis is the matching of the sentiment lexicon against an input text (cf. Section 3.1). Secondly, the patterns concerning change in sentiment degree are applied in order to identify contexts reinforcing or attenuating sentiments (cf. Section 3.2). In a third step negation patterns, which are used to discover contexts that change the polarity of sentiment orientation, are considered. Eventually those patterns are applied that try to associate the discovered sentiment expressions to the main topic or to one of the aspects of a topic identified in the analyzed text.

3.1 Lexical Matching

When matching the sentiment lexicon against the input text, the sentiment expressions detected in the text are associated with the corresponding sentiment orientation values found in the lexicon. In Example (2) the verb *hofft* (*hopes*) is identified as sentiment expression and associated with the sentiment orientation value *S+1+2* (*opinion=S+1+2*). The sentiment orientation value involves two dimensions: the first dimension is the degree of positive or negative orientation measured on a scale from -6 to +6; the second one quantifies the degree of emotional implication on a scale from 0 to +6. The latter dimension was introduced in order to quantify how strong the preference or aversion towards a specific topic is. In our example these two dimensions are: +1, indicating a slightly positive orientation, respectively +2, marking a somewhat increased emotional involvement.

2. Jetzt hofft die Bahn AG auf Verkehrszuwachs.
   (*The Bahn AG hopes that traffic will increase*)
   {expression=hofft, opinion=S+1+2,desc=bahn}

3.2 Pattern Matching

Pattern matching in **SentA** serves various purposes. After identifying sentiment expressions in input texts, **SentA** examines if the marked items are actually relevant for sentiment analysis. For example specific patterns are used to identify and filter out items that appear in questions or in the vicinity of particular structures. Then, typical patterns are applied in order to determine if the orientation of the detected sentiment expressions has been altered. We consider two categories of patterns to detect such cases: degree patterns and negation patterns. A third category of patterns called topic-relevant sentiment patterns is used in order to determine possible associations of a sentiment expression with one or several descriptors. The patterns are implemented in KURD[4], a flat pattern matching formalism (Carl & Schmidt-Wigger, 1998).

Degree Patterns

[4] KURD is an acronym representing the first letters of the basic actions of the formalism : Kill, Unify, Replace, Delete.

Degree patterns are used to establish if the sentiment orientation value of an expression has been reinforced or attenuated. In certain cases the meaning of a sentiment expression is modified through specific features such as the use of the comparative or the superlative degree in case of adjectives or adverbs. Context may also have such an influence on meaning. In (3) for example, the negative meaning of the noun *Protest* is reinforced through the use of the modifying adjective *heftig* (*fierce*).

3. Dies führte zu heftigen Protesten.
   (*This lead to fierce protests.*)
   {ori=Protesten,opinion=S-4+3}.

Specific patterns relying both on morphological and contextual information have been implemented in **SentA** in order to recognize such contexts and adjust the value of the sentiment orientation feature accordingly. In (3) the original sentiment orientation of the word *Protest*, *opinion=S-2+2*, is augmented to *opinion=S-4+3*. After detecting contexts of sentiment reinforcement or attenuation, the next step in the sentiment analysis process is applied.

Negation Patterns

Negation within an utterance changes the semantic orientation of sentiment expressions. Therefore, negation patterns are applied to the modified input, in order to detect negation markers and adjust the semantic orientation of the corresponding sentiment expressions accordingly. In (4) the semantic orientation of the adjective *zufrieden* (*satisfied*) is changed from *S+1+1* to *S-1+0* under the influence of the detected negative marker *nicht*.

4. "Wir werden uns damit nicht zufrieden geben", so Kohl.
   (*"We won't be satisfied with that", said Kohl*)
   {ori=zufrieden,opinion=S-1+0}

Topic-Relevant Sentiment Patterns

Once the sentiment orientation value of an expression has been established, **SentA** tries to associate the identified sentiment expression to the corresponding topic or issue. In Section 2 we mentioned that during the text categorization process, so-called descriptors, indicators of the topic or of its aspects, are computed. The topic–relevant sentiment patterns try to associate these descriptors to the sentiment expressions marked in the text. Some of these patterns are linguistically motivated. However, when no linguistically motivated pattern applies and the utterance contains both descriptors and sentiment expressions, other patterns relying on information concerning general sentence structure or vicinity try to associate them. As shown in examples (2) and (5) the base form of the descriptors detected in the sentence are assigned as values to the attribute *desc*.

5. Die Preise im Fernverkehr sind seit Sonntag um 3,1 Prozent, im Nahverkehr sogar um durchschnittlich 3,6 Prozent teurer.
   (*Ticket prices in long distance traffic increased by 3.1 Percent, in short distance traffic actually by an average of 3.6 Percent.*)
   {ori=teurer,opinion=S-2+3,desc=fernverkehr;nahverkehr}

In case of example (2) the association of the descriptor *Bahn* to the sentiment expression *hofft* was motivated by a linguistic pattern, i.e. a verb expressing sentiments is likely to refer to the grammatical subject of the sentence. On the other hand, in (5), patterns relying on information concerning coordination and sentence structure in general are used to associate the predicatively used adjective *teuer* to the descriptors *Fernverkehr* and *Nahverkehr*.

In the next section we evaluate the overall performance of **SentA**. Besides, the reliability of the various resources involved in the sentiment analysis process is tested.

4 Evaluation of SentA

For the evaluation of the system we used two classes of texts dealing with the German Railway Company *Deutsche Bahn*. The first category, henceforth *Bahn-News*, consists of 25 news texts automatically retrieved by the issues monitoring system from the internet. The second category of texts, *Bahn-Group*, includes 25 texts retrieved from various newsgroups on the net.

We established gold standards by manually annotating the text segments expressing opinions in all texts. Besides identifying the sentiment expressions (SE) in the test texts, the manual annotation also includes information concerning degree of sentiment orientation (D), negation (N) and the association of sentiment expression to particular topic descriptors (SE-TD). The following figure summarizes the characteristics of the two text classes.

Text class	Texts	Sentences	Words	SE	D	N	SE-TD
Bahn-News	25	484	8420	266	39	15	148
Bahn-Group	25	183	2863	160	15	6	59

Figure 3 : Characteristics of the test texts

We used the *Bahn-News* texts to tune our system to the specific topic *railway*. During tuning we extended the sentiment lexicon and the sentiment template database in order to cover as many sentiment expressions and typical patterns found in the *Bahn-News* texts as possible. Then, we compared the gold standards for both text classes with the automatic annotation of the test texts and computed precision and recall (cf. Figure 4) for detecting sentiment expressions and the corresponding sentiment orientation values. The association of sentiment expression to topic descriptors was evaluated at a later stage (cf. Figure 5).

Text class	Recall	Precision
Bahn-News	99%	97%
Bahn-Group	75%	92%

Figure 4 : Recall and precision values for the test texts

As expected, recall and precision for the *Bahn-News* texts are very high. Precision in case of the *Bahn-Group* texts is fairly high whereas recall is considerably lower than for the *Bahn-News* texts. Decrease in recall is mainly due to sentiment expressions that are not covered in the sentiment lexicon. Many of these expressions are rather colloquial and thus more typical to newsgroup postings than to the online articles on which the system was trained. Other missing sentiment expressions as *einfach* (*simple*) are ambiguous and can be used with different polarity depending on the context.

Figure 5 presents recall and precision values for all the resources involved in the sentiment analysis process considered independently. For all resources both recall and precision values for the *Bahn-Group* texts decrease. Decrease in recall is mainly due to sentiment expressions or patterns not yet covered in the resources. Precision in case of the degree and negation patterns is relatively high, 88% respectively 80%. However the number of such patterns detected in the test texts is too low to attempt any kind of generalization at the moment.

	SE		D		N		SE-TD	
Text class	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
Bahn-News	99%	97%	97%	97%	93%	100%	82%	61%
Bahn-Group	75%	92%	88%	88%	67%	80%	64%	57%

Figure 5 : Recall and precision values SE, D, N, und SE-TD

As shown in Figure 5 recall and precision values in case of the topic relevant sentiment patterns are quite low. In many cases, templates based on vicinity information lead to wrong associations which could be avoided by implementing more templates that are linguistically motivated.

5 Conclusions and Outlook

In this paper we described an issues monitoring system which uses linguistic resources to perform a fine-grained analysis and classification of sentiments related to user-specific topics. The evaluation of the sentiment analysis component shows that the performance of **SentA** can be compared to that of the CELI system (Dini, Mazzini, 2002) and **Sentiment Analyzer** (Yi et al., 2003). Even though a direct comparison of the approaches is not possible due to the differences in the technical approach, development, test corpora or covered patterns, figures show that the overall precision for the *Bahn-Group* texts (92%) is comparable to the values obtained with the CELI system (92%) and the **Sentiment Analyzer** (87%). Recall with **SentA** is, however, higher: 75% vs. 52% with the CELI system and 56% with **Sentiment Analyzer**. The evaluation also shows that the patterns associating sentiment expressions to the main topic of a text or to its various aspects need to be further refined. This seems to be a major problem in the CELI approach, too (Beermann et al., 2004). The experiment reported in (Beermann et al., 2004) showed that by integrating deep analysis with HPSG inappropriate sentiment expression – topic associations could be filtered out and thus noise could be reduced. However, the fact that recall could not be increased and the relatively high percentage of non-parsed text units (27%) along with longer processing times required by

deep analysis in general raises the question if deep analysis is the appropriate technique to be used to increase the overall efficiency of an issues monitoring system. Therefore, future work on **SentA** will rather concentrate on further refining the resources used for shallow processing, more specifically on the implementation of further linguistically motivated templates for the assignment of sentiment expressions to specific topics. Moreover, topic association needs to be extended to the text level and therefore, such phenomena as anaphora resolution or ellipsis must be addressed.

References

BEERMANN D., ought/downloads_pubEREID P, HELLAN L., GONELLA D., KURZ D., MAZZINI G., PLAEHN O., SIEGEL M. (2004). DeepThought - Hybrid Deep and Shallow Methods for Knowledge-Intensive Information Extraction (IST-2001-37836). D5.10. Evaluation report on efficiency, accuracy and usability of the new approach. http://www.eurice.de/deepthought/downloads_public/D5.10.pdf.

CARL M., SCHMIDT-WIGGER A. (1998). Shallow Post Morphological Processing with KURD. In Proceedings of the *Conference on New Methods in Natural Language Processing, NeMLaP'98*, Sydney.

DINI L., MAZZINI G. (2002), Opinion Classification through Information Extraction, In ZANASI A., BREBBIA C.A., EBECKEN N.F.F., MELLI P. (eds.), *Data Mining III*, WIT Press, Southampton.

MAAS H.-D. (1996). MPRO-Ein System zur Analyse und Synthese deutscher Wörter, In HAUSSER, R. (ed). *Linguistische Verifikation Sprache und Information*, Max Niemeyer Verlag, Tübingen.

MULLEN T., COLLIER N. (2004), Sentiment Analysis using Support Vector Machines with Diverse Information Sources, In Proceedings of the *EMNLP 2004*.

PANG B., LEE L. (2004), A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, In Proceedings of the *ACM 2004*.

RIPPLINGER B., SCHMIDT P. (2001), AUTINDEX: an automatic multilingual indexing system, In Proceedings of the 24th annual international *ACM SIGIR conference on research and development in information retrieval*.

YI J., NASUKAWA T., BUNESCU R., NIBLACK W. (2003), Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Technolgies, In Proceedings of the *Third IEEE International Conference on Data Mining (ICDM)*.

TALN 2005, Dourdan, 6–10 juin 2005

Parsing de l'oral : traiter les disfluences

Marie-Laure Guénot
Laboratoire Parole et Langage – CNRS / Université de Provence
{prenom.nom}@lpl.univ-aix.fr

Mots-clefs : Disfluences, Parsing, Linguistique de corpus, Linguistique formelle, Développement de grammaires, Grammaire de Construction (CxG), Grammaires de Propriétés (GP).

Keywords: *Disfluencies, Parsing, Corpus linguistics, Formal linguistics, Grammar development, Construction Grammar (CxG), Property Grammars (PG).*

Résumé Nous proposons une réflexion théorique sur la place d'un phénomène tel que celui des disfluences au sein d'une grammaire. Les descriptions fines qui en ont été données mènent à se demander quel statut accorder aux disfluences dans une théorie linguistique complète, tout en conservant une perspective globale de représentation, c'est-à-dire sans nuire à la cohérence et à l'homogénéité générale. Nous en introduisons une représentation formelle, à la suite de quoi nous proposons quelques mécanismes de parsing permettant de les traiter.

Abstract *We propose a theoretical reflexion about the place of a phenomenon like disfluencies, in a grammar. The precise descriptions that are available leads to a question : what status shall we give to disfluencies into a complete linguistic theory ?, keeping a global point of view and without compromising the coherence and the homogeneity of its representation. We introduce a formal representation of the phenomenon, and then we propose some parsing mechanisms in order to treat it.*

Introduction

On s'intéresse ici au traitement automatique des disfluences : phénomène non négligeable puisque très fréquent en oral spontané, la linguistique descriptive en a fourni un certain nombre d'études fines, présentant son organisation interne et ses caractéristiques. Cependant ces descriptions, quoique très précises dans leurs propositions, ne sont souvent pas exploitées en TALN, sans doute en partie parce que le statut des disfluences dans une grammaire n'y est pas défini de manière claire et formalisable. En effet les applications symboliques de traitement automatique qui s'efforcent d'analyser des données orales font appel à des techniques différentes pour traiter les disfluences, techniques qui sont pourtant basées pour la plupart sur les mêmes descriptions initiales.

Nous proposons ici une réflexion théorique concernant la place de phénomènes tels que les disfluences de l'oral dans une grammaire, laquelle grammaire a pour objet d'être représentée formellement, en vue notamment d'une exploitation en TALN. Nous conduisons notre réflexion en nous basant sur les travaux de linguistique descriptive, et dans le cadre du développement de plusieurs parseurs, aux caractéristiques et aux objectifs différents, mais qui sont tous basés sur une représentation formelle de descriptions linguistiques du français. Nous commencerons donc par exposer les études faites des disfluences en linguistique, puis nous montrerons comment nous avons interprété ces descriptions pour les rendre formalisables dans un modèle de représentation, avant de proposer un ensemble de mécanismes d'analyse automatique qui permettront d'exploiter cette grammaire et d'en tirer les résultats les plus efficaces possibles.

1 Situation du problème

1.1 Typologie(s) des disfluences

Dans la littérature linguistique, une disfluence est un endroit dans un énoncé où "*le déroulement syntagmatique est brisé*" (Blanche-Benveniste *et al.* (1990)) : on occupe une même place syntaxique avec plusieurs objets (ex. (1a)[1]).

(1) a.
il
il a quand-même **un :**
une fibre pédagogique **assez :**
assez euh enfin réelle quoi

b.
tu as toujours un rapport **(il) y a** **un directeur de fouilles**
(il) y a
(il) y a **les chouchous du directeur de fouille** et puis
les crétins de base enfin bon

Ce mécanisme n'est pas propre aux disfluences puisque l'on retrouve un même entassement dans les énumérations (ex. (1b)) ; en revanche, alors que dans ces dernières chaque occurrence de la même place ajoute un élément à la sémantique de l'énoncé, l'accumulation de *il* dans (1a) n'en modifie pas les caractéristiques sémantiques. Il serait de même abusif d'inscrire cette ré-

[1]Pour plus de clarté dans la lecture des disfluences, nos exemples sont représentés en grille (Blanche-Benveniste (1987)), et l'on notera **en caractères gras** les éléments illustratifs. Sauf mention contraire, tous les exemples de cet article sont tirés du Corpus d'Interactions Dilogiques (Bertrand & Priego-Valverde (2005)).

pétition comme étant une méthode de constitution syntagmatique (*i.e.*, l'accumulation paradigmatique ne forme pas de syntagme), ou de lui affecter des relations de dépendance syntaxique.

Parmi ces disfluences, on distingue deux grandes classes générales : les *bribes* qui sont des reprises à partir de syntagmes inachevés (ex. (1a)), et les *amorces* qui sont des reprises à partir de morphèmes inachevés (p. ex. *paran-* dans (2)).

(2) s'il n'y a pas d'éléments à mon avis euh il
il tombe **dans la paran-**
dans la parano quoi

Au sein des amorces Pallaud & Henry (2004) identifient trois formes différentes (formes que l'on peut, d'après elles, appliquer également aux bribes) : les amorces qui sont laissées *inachevées* (ex (3a)), celles qui sont *complétées* (ex. (2)), et celles qui sont *modifiées* (ex. (3b)).

(3) a. tu sais **j'ai v-** enfin
dans mon champ visuel (il) y a eu quelque chose tu vois

b. ils ont des ouvriers euh payés
spécialisés **sup-**
sur les chantiers de fouille

Pour sa part, Shriberg (1994), inspirée par Levelt (1983), a décrit l'organisation interne des disfluences en un ensemble d'espaces distincts : le *reparandum* qui est le lieu de la première production, inachevée au niveau du point d'interruption (*interruption point*), suivi de l'*interregnum* au sein duquel il peut se produire soit rien, soit une marque d'hésitation, soit une à plusieurs nouvelles tentatives de formulation (inachevées), jusqu'au *repair* qui correspond à la reprise du déroulement syntagmatique.

Toutes ces études, qui décrivent l'organisation interne des disfluences, peuvent être prises en considération dans le développement d'une formalisation. Elles les présentent comme un phénomène unique, avec des caractéristiques régulières (l'entassement paradigmatique, l'absence de fonction syntaxique et de fonction sémantique, les espaces internes), et des caractéristiques plus spécifiques à certains cas (bribes *vs.* amorces, inachèvement *vs.* complétion *vs.* modification, composition de l'interregnum). Cependant elles n'indiquent pas comment l'on différencie une disfluence d'une autre construction, ni comment l'on doit les traiter lors de l'analyse d'un énoncé.

1.2 Les disfluences en TALN

Observons maintenant comment le phénomène est traité en TALN[2] : en dépit de la prise en considération, dans la plupart des cas, de tout ou partie des descriptions exposées ci-dessus, les solutions concrètes proposées pour le traitement automatique sont nettement différentes suivant la tâche à accomplir et le type d'approche.

La première technique que l'on peut rencontrer consiste à "effacer" les disfluences de l'entrée qui sera analysée, en effectuant un pré-traitement des données dont l'objet est de reconnaître les disfluences et de les remplacer par une forme considérée comme "équivalente" ne présentant pas de rupture du déroulement syntagmatique (p.ex. chez Dowding *et al.* (1999)). On peut se

[2]Parce que l'on se place nous-même dans la perspective générale de la représentation formelle de la langue, on ne s'intéresse ici qu'aux méthodes de TALN qui sont basées sur des descriptions linguistiques, et non sur les techniques probabilistes, qui certes proposent des approches intéressantes, mais ne font pas partie de notre cadre de recherche.

demander quel est précisément le niveau d'"équivalence" recherché, et quelles sont les limites imposées par des résultats d'analyse pour des utilisations ultérieures, si ceux-ci sont basés sur des entrées qui ne contiennent plus la totalité des informations linguistiques produites.

La deuxième technique consiste en quelque sorte à "ignorer" les disfluences, *i.e.* à ne pas les prendre en compte lors de l'analyse (cf. par exemple Pérennou (1996)). Ceci permet d'obtenir un résultat de parsing dit "robuste", mais ne pose pas la question du statut des unités qui n'ont pas été considérées dans l'analyse : bien que les disfluences n'aient pas de fonction syntaxique en tant que telles, chaque élément qui occupe une place syntagmatique remplit en lui-même la fonction syntaxique de cette place, et il semble difficile d'admettre dans ce cas que seule une occurrence de chaque place sera considérée dans l'analyse. En d'autres termes, si l'on n'analyse que le *repair*, quel est le statut des constituants des autres espaces de la disfluence ?

La troisième technique est celle qui consiste à regrouper les disfluences en un groupe. Antoine *et al.* (2003) proposent dans cette perspective des analyseurs qui forment des disfluences en rassemblant des chunks (chacun d'eux devant être une occurrence de la même place syntaxique) en vertu de "*relations de dépendance sémantico-pragmatiques*". Dans un même ordre d'idées, Godfrey *et al.* (1992) proposent une méthode d'annotation des disfluences qui consiste à effectuer un parenthésage de la totalité de l'accumulation paradigmatique. Dans ce cas on peut se demander comment est déterminée la catégorie syntagmatique de cet ensemble (indispensable à l'analyse), qui peut être constitué de plusieurs répétitions ne comptant pas toujours les mêmes constituants, et qui ne correspond pas systématiquement à un syntagme complet.

On voit que les techniques qui prennent en compte les disfluences, bien que se basant sur tout ou partie des descriptions données ci-dessus, diffèrent nettement dans leurs traitements, qui sont tous limités par la nature de leur représentation du phénomène. Nous allons donc commencer par expliquer quelle place nous lui donnons dans notre grammaire, avant d'en montrer la formalisation proposée à partir de là.

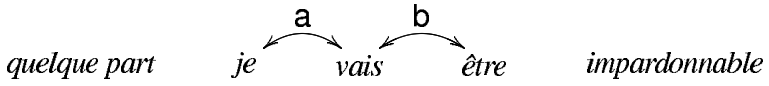
2 Description et représentation

Le fait de vouloir intégrer un phénomène tel que celui des disfluences dans une grammaire, met en avant une différence fondamentale entre les descriptions linguistiques et leur formalisation : là où la première s'attache à décrire finement le fonctionnement interne et les propriétés d'un phénomène donné, le modèle formel qui a pour tâche de le représenter doit également intégrer, en plus de sa description interne, les propriétés plus générales du phénomène, *i.e.* conserver un point de vue plus global sur l'articulation entre celui-ci et les autres, sur la façon dont le tout s'articule et conserve une cohérence générale.

C'est ce à quoi nous allons nous attacher ici. Nous allons proposer une réflexion générale sur la place des disfluences au sein d'un système grammatical. La grammaire du français que nous développons et dans laquelle nous tâchons d'intégrer ce que nous présentons ici a pour cadre théorique celui de la *Construction Grammar* (CxG, cf. p.ex. Kay & Fillmore (1999)), et pour modèle formel celui des *Grammaires de Propriétés* (GP, cf. p.ex. Blache (2005)). La description et la représentation formelle qui suivent sont donc basées sur ce cadre de travail CxG/GP, cependant nous espérons les présenter de telle sorte qu'elles puissent s'appliquer à d'autres cadres théoriques et formels que celui qui fait l'objet de notre travail.

Ce que l'on a pu voir jusqu'ici met en avant le fait que pour traiter des disfluences, il est né-

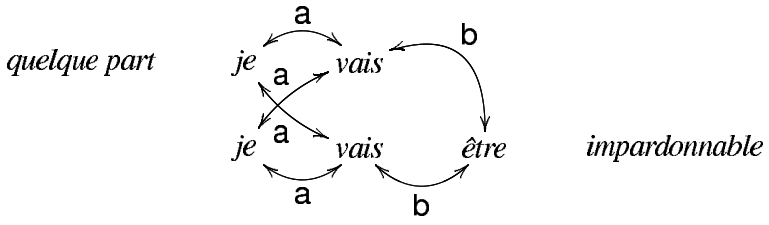
cessaire de commencer par répondre à une question d'ordre plus général : *Quelle est leur place dans la grammaire ?* Et pour répondre à cette question, la première chose que l'on doit se demander est ce que l'on représente au juste dans une grammaire : quels sont les objets que l'on y manipule ? Y représente-t-on des relations entre les occurrences possibles d'un énoncé, ou alors entre les places syntaxiques occupées par ces occurrences ? Dans le cas d'énoncés sans disfluences, ces questions ne sont pas si évidentes ; ainsi, dans un énoncé tel que celui de la figure 1[3], les *occurrences possibles* et les *places syntaxiques* sont confondues, puisque chaque place syntaxique n'est occupée que par une unique occurrence.



(1)

FIG. 1 – Relations dans un énoncé sans disfluences.

Par contre, quand on traite des énoncés avec des disfluences, ces questions prennent toute leur importance. Considérons d'abord que l'on représente dans une grammaire des *relations entre des occurrences possibles*. L'analyse d'une disfluence peut alors être représentée selon l'illustration de la figure 2.



(2)

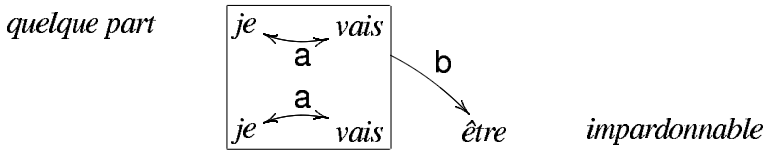
FIG. 2 – Relations entre occurrences.

On voit que dans ce cas on multiplie le nombre de chaque relation par le nombre d'occurrences de la même place syntagmatique : la relation a figure quatre fois et la b deux fois, au lieu d'une seule dans la figure 1. La conséquence de cela est que l'ensemble de caractéristiques, spécifique au syntagme qui contient la disfluence (ici, *je vais je vais être impardonnable*), varie non seulement en fonction de la présence, mais aussi de la forme d'une disfluence : l'ensemble {a, b} de la figure 1 devient {a, a, a, a, b, b} pour la disfluence précise de la figure 2. De plus, un certain nombre de propriétés définitoires du syntagme (*e.g.* l'unicité du pronom clitique nominatif, ou l'ordre linéaire entre ce même pronom et le verbe) sont faussées par la présence de la répétition, et la définition dans la grammaire doit tenir compte de ces variations de caractéristiques. Pourtant comme on l'a vu plus haut, les caractéristiques spécifiques à la présence d'une disfluence n'ont pas d'incidence sur l'analyse syntaxique d'un énoncé, et donc ne devraient pas avoir d'incidence sur la définition (syntaxique) d'un syntagme.

Représenter des relations entre occurrences ne semble donc pas être le fait d'une grammaire ; considérons alors que l'on y représente des *relations entre des places syntagmatiques*. Dans ce cas, si l'on suit l'approche de Godfrey *et al.* (1992), on peut illustrer le traitement des disfluences comme dans la figure 3.

Ici l'on résout le problème de la multiplication injustifiée des caractéristiques, mais on se trouve face à un autre problème : pour pouvoir faire une analyse syntaxique il faut qu'à chaque place

[3]Les "relations" que l'on représente ici sont des représentations intuitives, et n'illustrent pas une théorie donnée.



(3)

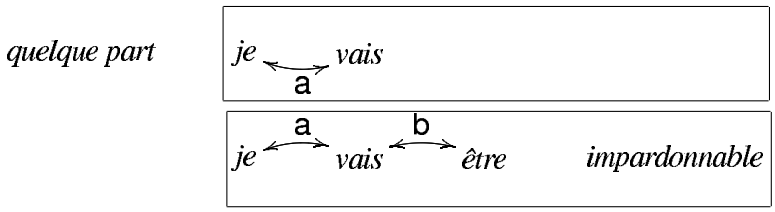
FIG. 3 – Relations entre places syntagmatiques.

corresponde une catégorie, et dans l'exemple quelle est la catégorie de *je vais je vais*, ou de *c'était je crois qu'il était* dans l'ex. (4) ?

(4) **c'était**
je crois qu'il était autrichien ou un truc comme ça

En effet, s'il est simple de traiter de cette façon les reprises simples telle que *il il* dans l'exemple (1a) en lui affectant la place de "pronom clitique", il devient plus difficile de s'accorder sur le statut d'un groupe constitué d'un fragment de début de syntagme, qui ne correspond à aucune étiquette syntaxique. Comment intégrer dans une analyse syntaxique, des objets qui ne sont pas des éléments syntaxiques ? Il faudrait pour cela les ajouter artificiellement à la grammaire, intégrer ces groupes qui ne sont pas vraiment des syntagmes, mais dont la seule raison d'y figurer est qu'ils peuvent apparaître en tant qu'occurrence. Au-delà du problème linguistique de fond que ce type d'artefact suppose (quelle analyse fait-on ? quelle est la nature des objets que l'on introduit ?), on en revient également au problème posé par la technique précédente : les disfluences ne sont pas décrites en tant que phénomène, mais chacune des possibilités de disfluence devra être l'objet d'une construction particulière, et l'on sera limité à un moment ou à un autre par l'itération limitative des possibilités *a priori* infinies.

Une façon de remédier à ce problème sans fabriquer de catégories *ad hoc* est de ne pas rassembler les différentes occurrences d'une disfluence en un groupe unique, mais de considérer que chaque reprise est une occurrence, achevée ou non, du syntagme complet (figure 4).



(4)

FIG. 4 – Les disfluences comme occurrences complètes de syntagmes.

Pour ne pas recourir à des catégories vides[4], on peut tout simplement considérer (ce qui est tout à fait cohérent dans le cadre CxG/GP) que la caractérisation des occurences inachevées de syntagme seront le reflet de leur constitution : un certain nombre de propriétés seront, à juste titre, non évaluées, et d'autres seront évaluées et non-satisfaites, en comparaison avec le *repair*. Il s'agit ensuite, pour ne pas se contenter de déplacer au niveau supérieur le problème posé par la première possibilité envisagée ici, de mettre en relation ces occurrences du même syntagme au sein de la grammaire, en tant que "phénomène de disfluence". Il est possible de le repérer par un ensemble de caractéristiques telles que la différence de caractérisation entre les premières occurrences, incomplètes, et le *repair*, ainsi que l'occupation des mêmes fonctions syntaxique et sémantique. D'autres éléments peuvent ensuite permettre de distinguer les différentes formes

[4]Nous n'avons pas la place de justifier notre position sur ce point ici, cependant nous développons une grammaire qui n'a pas de recours au postulat de catégories vides, que ce soit pour ce cas ou pour tous les autres.

de disfluences : les *inachevées* n'auront pas les mêmes constituants que le repair, les *modifiées* s'en écarteront par un sous-ensemble variable de traits (différences de genre, de nombre, etc.), contrairement aux *complétées* dont les constituants (présents) seront identiques à ceux du repair.

Ce qui pose un problème plus délicat dans cette représentation, c'est l'expression de relations entre parties différentes d'occurrences différentes. Observons par exemple l'énoncé suivant :

(5)

lesquels registres		sont très	euh
		sont	
	doivent	être	
	doivent pouvoir	être	contrôlés

tiré de Blanche-Benveniste *et al.* (1990), p. 24, où l'auteur dit que l'on "*vise à dégager la séquence maximale qui a été donnée par le locuteur, en tenant compte de toutes les bribes qu'il a fournies ; dans l'exemple précédent, on retiendra comme séquence maximale :* lesquels registres doivent pouvoir être très contrôlés", ce qui implique de pouvoir établir des relations entre *très* apparu dans le reparandum, et *contrôlés* apparu dans le repair. Or, bien qu'une analyse en GP repérera cette relation, il paraît par contre bien difficile de la représenter au sein même de la grammaire et donc d'en tenir compte dans la description de la "*séquence maximale*". Difficile, parce que cela demanderait d'utiliser une relation entre constituants de syntagmes différents, lorsque ces syntagmes sont liés entre eux par une relation de disfluence, et que l'on n'utilise pas de telles relations dans d'autres cas. On devrait donc créer artificiellement, non pas de nouveaux syntagmes comme précédemment, mais de nouvelles propriétés (ou de nouvelles *sémantiques* de propriétés, cf. Vanrullen *et al.* (2003)) uniquement pour traiter le phénomène.

La représentation que l'on propose conserve la vision des disfluences en tant que construction à part entière, telle que montrée ci-dessus, mais ne la considère pas systématiquement comme mettant en relation des syntagmes : on reliera en tant que disfluence chaque occurrence répétée d'une même place syntaxique, qu'elle soit construite ou non (figure 5).

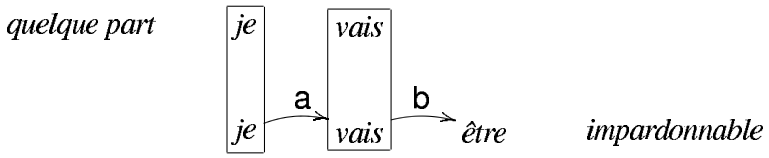
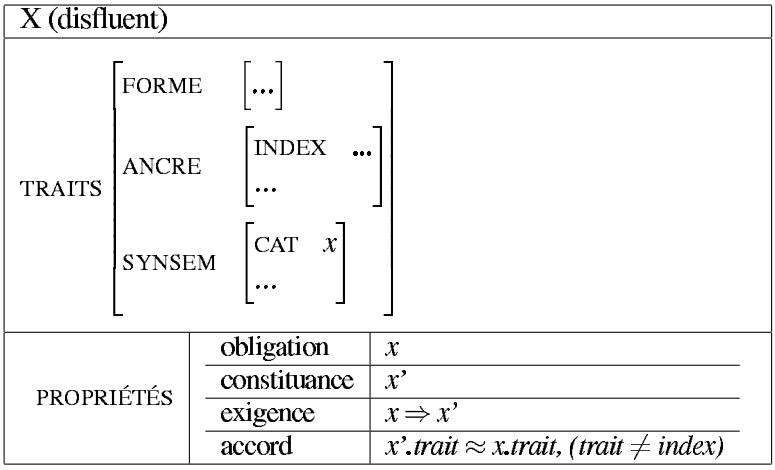


FIG. 5 – Les disfluences comme accumulation de places syntagmatiques.

Ainsi, toutes les places syntaxiques qui doivent être mises en relation lors de l'introduction du syntagme qui les contient (le "syntagme maximal" de Blanche) sont séparées les unes des autres et peuvent être prises en compte par le reste de la grammaire de la même façon, qu'elles soient disfluentes ou non, sans introduire ni formes de constructions ni relations *ad hoc*. Concrètement, une bribe *complétée* aura la forme de la figure 4, et une *modifiée* aura cette même forme à ceci près que les éléments d'une disfluence auront quelques différences de valeurs de traits. Les *inachevées*, enfin, seront traitées comme des reprises de syntagmes (donc comme décrites dans le cas précédent, mais uniquement quand aucune identité de forme ne permet de relier des unités plus petites entre elles).

Dans la grammaire, la construction correspondant aux disfluences aura la forme proposée dans la figure 6[5]. Ces quelques caractéristiques nous permettent de signifier qu'une disfluence consiste en la mise en relation d'un *x* et un ou plusieurs *x'*, dont chacun des traits, mis à part l'INDEX,

[5]Il s'agit d'une représentation extrêmement simplifiée pour les besoins de l'article. La développer ici sous sa forme réelle eût demandé des explications superflues dans ce contexte précis.



(6)

FIG. 6 – représentation formelle de la construction "disfluente".

sont de valeurs identiques un à un. La satisfaction complète de ces propriétés caractérise une bribe *complétée* ; si un certain nombre de propriétés d'accord sont évaluées et non satisfaites alors on caractérise une bribe *modifiée*, et l'application de *x* et *x'* en tant que syntagmes nous permet de reconnaître les bribes *inachevées*. Les *accords*, dans notre grammaire, font référence à la totalité des traits d'un objet, que ceux-ci soient morphologiques ou syntaxiques ou sémantiques (ou autres), ce qui nous permet de traiter avec une même description tout aussi bien les disfluences dont la modification est syntaxique (*un une* dans (1a)), que celles dont la modification est sémantique (ex. (6)).

(6) ils sont pas à l'abri de ça quoi mais c'est **un peu**
pas mal d'hypocrisie quand-même à ce niveau-là

3 Mécanismes de parsing

Dans notre cadre on s'intéresse principalement à des tâches de parsing non-déterministe, cependant l'on suppose (et l'on espère) que les quelques mécanismes qui suivent pourront s'appliquer aussi bien à du parsing déterministe. En outre, l'implémentation de cette grammaire est en cours, mais à l'heure actuelle elle n'est pas suffisamment avancée pour que l'on puisse la considérer comme évaluable. Nous verrons donc ici les idées générales qui dirigent la phase d'implémentation qui est en cours de réalisation.

Nous avons vu jusque là comment représenter les disfluences au sein de la grammaire, et pourquoi ; voyons maintenant quelles conséquences cette introduction peut avoir sur le parsing. En effet, les tenants et les aboutissants de l'automatisation d'une analyse basée sur une grammaire formelle ont ceci de différent du développement de la grammaire elle-même, qu'ils doivent fournir un résultat exploitable à l'issue d'un traitement le plus efficace et le plus robuste possible. L'obstacle principal ici est un problème d'explosion combinatoire : la description des disfluences proposée est tellement large qu'elle va engendrer l'introduction de constructions disfluentes non seulement dans les cas pertinents, mais également dans une quantité déraisonnable de cas superflus. Plusieurs façons de limiter les introductions superflues sont envisageables :

– Borner l'introduction d'une disfluence à une distance arbitrairement définie (qu'elle soit fixe ou relative à la longueur de l'énoncé). Même si les disfluences peuvent être non bornées (et elles le sont dans de nombreux cas selon notre description), elles ne sont probablement que

très rarement séparées de plus d'une certaine distance.
– Introduire une série de marques linguistiques permettant de différencier les disfluences et les énumérations (c'est ce que font p.ex. Johnson *et al.* (2004)). Les différentes parties des disfluences auraient tendance à être séparées par des pauses oralisées, des connecteurs, alors que les énumérations seraient plutôt séparées par des coordonnants. De plus, les énumérations sont des entassements de syntagmes dont des occurrences sont (normalement) toutes achevées, contrairement aux disfluences comme on l'a vu.

A terme, l'observation des résultats fournis en parsing en faisant varier ces différentes possibilités devrait nous permettre de faire remonter des informations exactes à ce propos, que l'on pourra intégrer directement à la grammaire comme autant de propriétés supplémentaires des disfluences.

Un autre mécanisme à traiter lors du parsing est celui de l'instanciation des traits de la construction disfluente en vertu des traits de ses constituants. Notre méthode consisterait, pour le cas où l'accord n'est pas satisfait, à affecter à la construction disfluente la valeur de trait de l'occurrence du repair, pour justifier de l'analyse par exemple de *un une* dans (1a) comme étant un déterminant de genre féminin (et non simplement indéterminé). Ceci permettrait d'affiner l'analyse et donc de réduire la possibilité d'introduction de relations non pertinentes par la suite.

Conclusion et perspectives

Les disfluences dont nous traitons ici ont été décrites comme des entassements paradigmatiques qui ont ceci de particulier qu'ils n'ont ni de fonction syntaxique, ni de fonction sémantique. Les objets accumulés sur une même place syntaxique peuvent être parfaitement identiques ou partiellement différents, mais ont toujours un certain nombre de caractéristiques communes, et l'on peut en décrire une organisation interne assez précise. Cependant l'exploitation très variable que l'on peut en voir en TALN montre qu'en plus de tout cela, il est nécessaire avant de traiter les disfluences dans un modèle basé sur un formalisme linguistique, de répondre aux questions suivantes : Quelle est la place des disfluences dans une grammaire ? Quand et comment les analyse-t-on ? Nous avons donc ici proposé une réflexion sur la place des disfluences dans une grammaire formelle, à la suite de laquelle nous avons introduit notre représentation de ce phénomène. Au delà du problème posé par cette représentation particulière, nous avons mené une réflexion plus générale sur la place de ce type de phénomènes, propres à l'oral, et par là nous avons présenté une vision globale du développement de grammaire et de ces spécificités. Nous avons proposé ensuite quelques mécanismes de traitement de la grammaire proposée, qui devraient permettre de garantir des résultats plus pertinents et robustes pour la tâche précise du parsing, tout en mettant en avant la méthode de développement de grammaire assisté par l'informatique, qui permet d'effectuer un va-et-vient entre les hypothèses formées à partir de l'étude de corpus, leur formalisation et leur vérification sur une quantité importante de données de manière automatisée.

Ce travail ne s'arrête bien évidemment pas là. Nous avons présenté ici une étude dont l'implémentation est en cours, qui a pour objectif à terme de proposer une grammaire formelle du français oral, basée sur des descriptions fines de corpus. Un autre objectif qui transparaît à travers cet article, et qui sera sans aucun doute nécessaire à la finalisation de cette grammaire, est l'intégration au sein même de la grammaire formelle, d'informations de plusieurs domaines différents : nous avons évoqué la syntaxe et la sémantique ici, mais pour traiter de l'oral il nous semble évident qu'à cela nous devrons ajouter des informations prosodiques (p. ex. Morel &

Danon-Boileau (1998)). Nous avons également pour projet d'ajouter à cela les informations gestuelles pertinentes[6], suite à l'étude du corpus qui nous a servi de base pour cet article, et qui est disponible sous forme audiovisuelle. Il serait également intéressant d'intégrer au traitement des éléments concernant l'interprétation (pragmatique et/ou psycholinguistique) de ces disfluences, et qui justifient leur apparition et leur statut lors de la perception d'un message. Enfin, outre le développement de grammaire, cette étude s'inscrit dans un projet d'annotation multi-niveaux de corpus, et dans ce cadre permet de réfléchir aux différences et aux liens existants entre les différents niveaux à représenter, en se basant sur l'étude de phénomènes réels et concrets.

Références

Jean-Yves Antoine, Jérome Goulian, & Jeanne Villaneau. Quand le tal robuste s'attaque au langage parlé : analyse incrémentale pour la compréhension de la parole spontanée. In *Actes de TALN 2003*, 2003.

Roxanne Bertrand & Béatrice Priego-Valverde. Le corpus d'interactions dilogiques : Présentation et perspectives. Technical report, Laboratoire Parole et Langage – CNRS / Université de Provence, 2005.

Philippe Blache. Property grammars : A fully constraint-based theory. In H Christiansen, P Skadhauge, & J Villadsen, editors, *Constraint Satisfaction and Language Processing*. Springer-Verlag, 2005.

Claire Blanche-Benveniste. Syntaxe, choix du lexique et lieux de bafouillage. *DRLAV*, 36-37 :123–157, 1987.

Claire Blanche-Benveniste, Mireille Bilger, Christine Rouget, & Karel Van Den Eynde. *Le français parlé : Etudes grammaticales*. Sciences du langage. CNRS Editions, Paris, 1990.

J. Dowding, J. Gawron, D. Appelt, J. Bear, L. Cherny, R. Moore, & D. Moran. Gemini : A natural language system for spoken language understanding. In *Proceedings of ARPA Workshop on Human Language Technology*, pages 21–24, 1999.

J. J. Godfrey, E. C. Holliman, & J. McDaniel. Switchboard : A telephone speech corpus for research and development. In *Proceedings of the IEEE*, pages 517–520, 1992.

Marie-Laure Guénot & Emmanuel Bellengier. Quelques principes pour une grammaire multimodale du français. In *Proceedings of RECITAL 2004*, 2004.

Mark Johnson, Eugene Charniak, & Matthew Lease. An improved model for recognizing disfluencies in conversational speech. In *Rich Transcription 2004 Fall workshop*, 2004.

Paul Kay & Charles J. Fillmore. Grammatical constructions and linguistic generalizations : The *what's X doing Y ?* construction. *Language*, 75(1) :1–33, March 1999.

W Levelt. Monitoring and self-repair in speech. *Cognition*, 14 :41–104, 1983.

Mary-Annick Morel & Laurent Danon-Boileau. *Grammaire de l'intonation*. Bibliothèque de faits de langues. Ophrys, Paris, 1998.

Berthille Pallaud & Sandrine Henry. Amorces de mots et répétitions : des hésitations plus que des erreurs en français parlé. In *Actes de JADT 2004*, 2004.

G. Pérennou. Compréhension du dialogue oral : le role du lexique dans l'approche par segments conceptuels. In *Actes de Lexique et Communication Parlée*, pages 169–178, 1996.

Elizabeth Shriberg. *Preliminaries to a theory of speech dislfuencies*. PhD thesis, University of California, Berkeley, 1994.

Tristan Vanrullen, Marie-Laure Guénot, & Emmanuel Bellengier. Formal representation of property grammars. In *Proceedings of ESSLLI Student Session*, 2003.

[6]Ce point avait été introduit dans Guénot & Bellengier (2004).

Description détaillée des subordonnées non dépendantes Le cas de *quand*

Christophe Benzitoun

Equipe DELIC – Université de Provence
29, Av. Robert Schuman 13100 Aix-en-Provence
Christophe.Benzitoun@up.univ-aix.fr

Mots-clés : syntaxe, subordination, dépendance, topologie
Keywords : syntax, subordination, dependency, topology

Résumé

De nombreux linguistes ont mis en évidence des cas de « subordonnées » non dépendantes dans de multiples langues dans le monde (Mithun, 2003 ; Haiman & Thompson (eds), 1988). Ce phénomène a aussi été relevé en français, notamment pour un « subordonnant » tel que *parce que* (Debaisieux, 2001 ; Ducrot et al., 1975). Nous nous proposons de décrire un cas de « subordonnée » en *quand* non dépendante et de le représenter dans le cadre formel de Gerdes & Kahane (à paraître)[1].

Abstract

Many linguists have pointed out instances of non dependent clauses "subordinate in form" in various languages in the world (Mithun, 2003 ; Haiman & Thompson (eds), 1988). Such cases have been found and informally analysed in French, for instance *parce que* (Debaisieux, 2001 ; Ducrot et al., 1975). We propose here to extend the analysis to cases of non dependent subordinate clauses involving *quand* and to integrate it in the formal framework of Gerdes & Kahane (to appear).

1 Introduction

La syntaxe formelle s'est longtemps préoccupée de la seule « phrase simple » de sorte qu'elle est désormais en mesure d'analyser avec précision les divers arrangements et relations qu'entretiennent les éléments constitutifs de cette unité. Il n'en est pas de même pour la « phrase complexe » qui a été beaucoup moins étudiée. De plus, lorsque le problème de la subordination en français est abordé, le cas des subordonnées non dépendantes n'est généralement pas envisagé ou bien traité en terme de relations entre unités discursives (Delort, 2004). Il est fort possible que cela soit dû à l'idée d'une correspondance directe entre

[1] Article disponible à l'adresse : http://www-erssab.montaigne.u-bordeaux.fr/IMG/pdf/GerdesKahane.Long.pdf

la présence d'un « subordonnant » et la relation de « subordination » (ou dépendance), critère que l'on retrouve dans à peu près toutes les grammaires du français. Notons au passage que dans une perspective de traitement automatique, il est très pratique de repérer les relations de « subordination » à partir de la présence d'un « subordonnant » car il s'agit d'éléments facilement repérables dont la liste des plus fréquents peut être établie assez facilement. Malheureusement, dès que l'on ne se base plus sur cette seule présence et que l'on essaie de déterminer si une subordonnée est dépendante à partir d'autres critères, on s'aperçoit qu'il y a des constructions indépendantes qu'il est difficile de négliger compte tenu de leur fréquence.

A notre connaissance, c'est Brunot, dès 1922, qui le premier a constaté pour le français que le terme « subordination » désignait le procédé de rattachement par certains morphèmes (conjonction de subordination) d'une construction verbale à une autre et non une relation de dépendance grammaticale. Cette analyse n'a cessé d'être affinée jusqu'au célèbre article de Haiman & Thompson (1984) prônant l'abandon du concept de « subordination » à cause de son incapacité à embrasser la diversité des relations qui peuvent exister entre deux constructions verbales, comme nous le révèlent les usages attestés. Et il est vrai que l'on observe des structures particulières, sans avoir recours à une théorie syntaxique donnée ou à des usages de langue parlée ou « familière ».

1) *Frappé de tout ce qu'il vient de voir, le philosophe réfléchit profondément à ces terribles scènes et se demande où donc est la vérité ?* ***Quand*** *tout à coup une voix se fait entendre dans les airs, prononçant distinctement ces mots : « C'est ici le fils de l'homme ! que les cieux se taisent et que la terre écoute sa voix. »* [Gaberel, Rousseau]

2) ***Quand*** *je pense que quelqu'un qui livrerait cet homme-ci gagnerait soixante mille francs et ferait sa fortune !* [Hugo, Quatre-vingt-treize]

Comme le suggère la ponctuation, ces deux constructions introduites par *quand* sont vraisemblablement des indépendantes, ce qui infirme de manière évidente la correspondance entre « subordonnant » et dépendance.

Dans cet article, nous ne proposerons pas un traitement effectif de ces phénomènes. Il serait en effet trop tôt pour une telle approche vu que l'on s'attaque à des exemples complexes peu envisagés jusqu'ici. Nous envisagerons seulement un type de constructions en *quand* ayant l'apparence d'éléments dépendants d'un verbe (à cause de la présence du « subordonnant » *quand*). Dans un premier temps (partie 2), nous montrerons que la distinction entre « subordonnées » dépendantes et indépendantes peut non seulement être mise en évidence par des tests syntaxiques (manipulation des données) mais aussi grâce à certaines propriétés de la construction pouvant elle-même se réaliser en contexte (partie 3). En outre, cette distinction de statut syntaxique se traduira par une différence sémantique et un relâchement des contraintes du « subordonnant » sur la construction qu'il introduit, phénomène parfois qualifié de « main clause phenomena » (Green, 1976) à cause de sa ressemblance avec la structure d'une « principale ». Dans un second temps (partie 4), nous en proposerons une illustration dans le cadre de Gerdes & Kahane (à paraître), modèle qui, selon nous, permet d'articuler une syntaxe de disposition des unités et une syntaxe de dépendance.

2 Tests syntaxiques

Nous partons de l'idée que les « subordonnées » sont généralement dépendantes d'un verbe,

quand elles sont dépendantes[2]. Nous nous plaçons donc dans une perspective de grammaire de dépendance. Afin de déterminer si un élément est dépendant d'un verbe, nous allons utiliser les tests proposés par Blanche-Benveniste et al. (1987) dans le cadre de l'Approche Pronominale. Il a été démontré par ces auteurs que ces tests permettent de mettre en évidence qu'un élément est dépendant d'un verbe dans le cas où les énoncés résultants sont grammaticaux. En voici un exemple en guise d'illustration.

3) *Je suis arrivé **quand** le soleil se levait.*

4) Equivalence avec un interrogatif : ***Quand** suis-je arrivé ? Quand le soleil se levait.*

5) Extraction : ***C'est** quand le soleil se levait **que** je suis arrivé.*

6) Adverbe : *Je suis arrivé **juste** quand le soleil se levait.*

En face de ces exemples largement présents dans la littérature traitant de la « subordination circonstancielle », on trouve des énoncés comportant un *quand* qui ne peut être analysé comme une marque de dépendance. Il s'agit d'exemples tels que le suivant dans lequel les tests appliqués ci-dessus ont pour résultat des énoncés agrammaticaux[3].

7) *J'attendais dans ma chambre sur mon lit **quand** tout à coup ça frappe à ma porte "c'est maman ouvre".* [Enfants]

8) ***Quand** attendais-je dans ma chambre sur mon lit ? Quand tout à coup ça frappe à ma porte.*

9) ***C'est** quand tout à coup ça frappe à ma porte **que** j'attendais dans ma chambre sur mon lit.*

10) ***J'attendais dans ma chambre sur mon lit **juste** quand tout à coup ça frappe à ma porte.*

Malgré leur caractère généralement littéraire ou narratif, les *quand* non dépendants se retrouvent dans divers genres avec néanmoins des fréquences assez variables. Nous les avons justement retenus à cause de cet aspect littéraire car les « subordonnées » non dépendantes ont souvent été étudiées dans des corpus oraux, si bien que cela aurait pu faire penser, à tort, à une caractéristique de l'oral ou de la langue « relâchée ». De plus, ce phénomène est largement indépendant du seul *quand* et a été mis en évidence pour des mots comme *que* (Deulofeu, 1999), *parce que* (Debaisieux, 2001), *bien que*, *alors que, tandis que*… Il paraît donc inenvisageable de ne pas le prendre en compte car *que* est parmi les mots les plus fréquents de la langue française et Debaisieux (2001) a recensé 78% de constructions non dépendantes, dans ses corpus oraux, parmi les « subordonnées » en *parce que*.

Il est bien évident que toutes ces « subordonnées » doivent néanmoins être rattachées au contexte mais autrement que par une relation de dépendance. Nous pensons que ce lien reste

---

[2] On observe aussi des cas de « subordonnées » dépendantes d'un nom : *Je pense à Paul quand il avait vingt ans* => ***C'est** quand il avait vingt ans **que** je pense à Paul.* Ces tournures ont été étudiées par Jeanjean (1985). Nous ne les aborderons pas dans le cadre de ce travail.

[3] Nous n'avons trouvé aucun exemple de la sorte dans les nombreux corpus que nous avons consultés, ce qui peut confirmer l'intuition de leur agrammaticalité.

syntaxique car il se fonde notamment sur des contraintes d'ordre linéaire. Et il est vrai que ces éléments observent des règles de placement assez rigides dont ne pourrait pas rendre compte la dimension sémantique ou une hypothétique dépendance avec la « phrase », composantes auxquelles on aurait envie d'avoir recours pour décrire ces relations. Rendre compte de ce phénomène en terme d'arrangement et non plus en terme de dépendance permet de traiter en une seule fois l'enchâssement dans l'énoncé et la position contrainte de l'élément (cf. ex. 19).

Maintenant que nous avons rapidement montré que des tests syntaxiques mettent en évidence que des « subordonnées » peuvent être non dépendantes d'un verbe, il nous faut détailler ce qui, formellement, distingue les dépendantes des non dépendantes en vue de leur futur traitement automatique. En premier lieu, le rattachement au contexte par un autre procédé que la dépendance permet d'expliquer certains phénomènes observés.

3 Propriétés distinctives

Tout d'abord, la construction en *quand soudain*[4] apparaît après une ponctuation forte de manière tout à fait significative. Plus du tiers des exemples recueillis sont précédés de cette marque voire débutent un paragraphe distinct. De plus, on remarque une très forte tendance à trouver dans la « principale » un verbe à l'imparfait et dans la « subordonnée » un verbe au passé simple, ce qui n'est pas le cas de la plupart des « subordonnées circonstancielles » en *quand*.

11) *Il y **avait** à peine quelques minutes qu'il avait atteint son but, reprenant son rôle diurne, quand soudain, il **entendit** les grincements caractéristiques des marches de l'escalier.* [Fantastique]

Hormis la présence prégnante du couple imparfait – passé simple, on observe des décalages temporels qui seraient difficiles dans une construction dépendante d'un verbe.

12) *J'en **étais**, cher lecteurs et amis à ces reflexions (et oui, il m'arrive de penser à Arnaud), quand soudain, alors que j'étais à la fenêtre, un individu d'une vingtaine d'années **m'interpelle** : LOUVET !* [Journal_lycéen]

13) *? Je **partais** quand il **arrive**.*

Les éléments pouvant se retrouver dans la subordonnée peuvent être assez divers, un peu comme si le « subordonnant » ne posait aucune contrainte. Cela va de la construction nominale :

14) *J'étais assis, dans le bus Orléans-Denfert bondé, je somnolais frileusement... Ding-ding, régulièrement la chevillette du contrôleur marquait les arrêts... Quand soudain, dérangeant ma torpeur juvénile, **un vague remue-ménage** : «... mais ce jeune homme va certainement se faire un plaisir de vous céder sa place». Aïe. Emmerdeur.* [Bayon, Le lycéen]

à l'interrogative et au discours direct :

---

[4] Par raccourci d'écriture, nous résumons par les termes *quand soudain* l'ensemble des constructions non dépendantes du type *quand soudain* et *quand tout à coup*. Nous verrons par la suite que ces mêmes constructions peuvent apparaître sans *soudain* ou *tout à coup*, sans pour autant que cela modifie leur statut syntaxique de construction non dépendante.

15) *Ils étaient donc chez lui, ce soir-là, dans cette chambre qu'ils appelaient « le tombeau de la femme inconnu ». Quand soudain...* ***Qui donc sonne si tard, compagnons de la Marjolaine ?*** [Montherlant, cité par Sabio (2003)]

Toutes ces configurations sont normalement impossibles dans une « vraie subordonnée » (dépendante) car un « subordonnant » est généralement associé à un verbe à temps fini. Muller (1996) en fait d'ailleurs une propriété des « subordonnants » et y voit un paradigme <*que* + temps fini>.

Contrairement aux « subordonnées canoniques », ces *quand* ne peuvent être en relation avec un verbe enchâssé.

16) *Je demandais au serveur qu'il* ***débarrasse*** *la table quand il aura fini le service.*

17) ***Je demandais au serveur qu'il débarrasse la table*** *quand soudain une bagarre éclata.*

Dans 16), *quand il aura fini le service* est dépendant du verbe *débarrasser* alors que dans 17), il est impossible de proposer la même analyse. *Quand soudain* enchaîne sur la totalité de ce qui précède.

Les *Quand soudain* ne peuvent pas non plus s'antéposer, contrairement aux autres « subordonnées » en *quand.*

18) *Quand on fait le vide dans le vase, la pomme est "attirée" vers le bas* [Sciences]

19) **Quand tout à coup ça frappe à ma porte, j'attendais dans ma chambre sur mon lit.*

Il serait facile de penser que c'est l'adverbe *soudain* ou *tout à coup* qui contraint l'emploi de *quand.* Mais on trouve des exemples en *quand* non suivis de *soudain* ou *tout à coup* qui acceptent pourtant la même analyse. Cette propriété est particulièrement visible dans des exemples comportant un complément temporel car on observe alors une sorte de « décalage » et l'antéposition, notamment, est visiblement exclue.

20) *Je dormais à poings serrés le lendemain matin* ***quand*** *aux pieds de mon lit se dressa un monsieur en habit noir* [exemple extrait de Sandfeld (1936)]

21) **Quand aux pieds de mon lit se dressa un monsieur en habit noir je dormais à poings serrés le lendemain matin.*

En fait, une construction dépendante peut être coordonnée à un complément circonstanciel de temps comme *la nuit* alors qu'une construction non dépendante peut difficilement rentrer dans un paradigme temporel.

22) *Les pistes sont totalement désertes* ***la nuit et quand il pleut*** [Forum]

23) **Je dormais à poings serrés* ***le lendemain matin et quand aux pieds de mon lit se dressa un monsieur en habit noir***

De plus, les *quand soudain* ne constituent pas une circonstance de l'action en cours mais deux actions successives ou en coïncidence si bien que le cadre temporel peut être ressenti comme étant posé par la première construction. Cela rejoint l'idée sous-jacente à la « subordination inverse », catégorie traditionnelle dans laquelle sont classées ces constructions. La distinction

syntaxique permet donc d'illustrer formellement une différence sémantique, ce qui ajoute un intérêt supplémentaire à l'étude que nous menons. On peut notamment étendre cette analyse à des « subordonnants » comme *alors que* (temporel vs adversatif), *si* (hypothèse vs prémisse), *comme* (manière vs causal)…

24) *Je parle comme je veux. / Je parle comment ?*

25) *Comme il avait soif, il a du aller boire à la fontaine. / *Il a du aller boire à la fontaine comment ?*

Il serait d'ailleurs intéressant de se pencher sur la part que prend la sémantique dans la détermination du statut de ce type de constructions. Mais nous pensons que la méthodologie souffrirait d'une trop grande proximité entre syntaxe et sémantique, cette dernière ne devant normalement pas intervenir dans le calcul des relations syntaxiques. Nous reportons donc le traitement de l'interface syntaxe-sémantique à un travail ultérieur[5].

Nous allons maintenant montrer qu'il est possible d'intégrer la distinction que nous avons opérée dans un cadre formel qui soit en adéquation avec l'analyse que nous avons proposée. Selon nous, le modèle de Gerdes et Kahane permet de récupérer l'élément grammaticalement indépendant en l'intégrant au reste de l'énoncé grâce à la dimension topologique.

4 Intégration dans un modèle

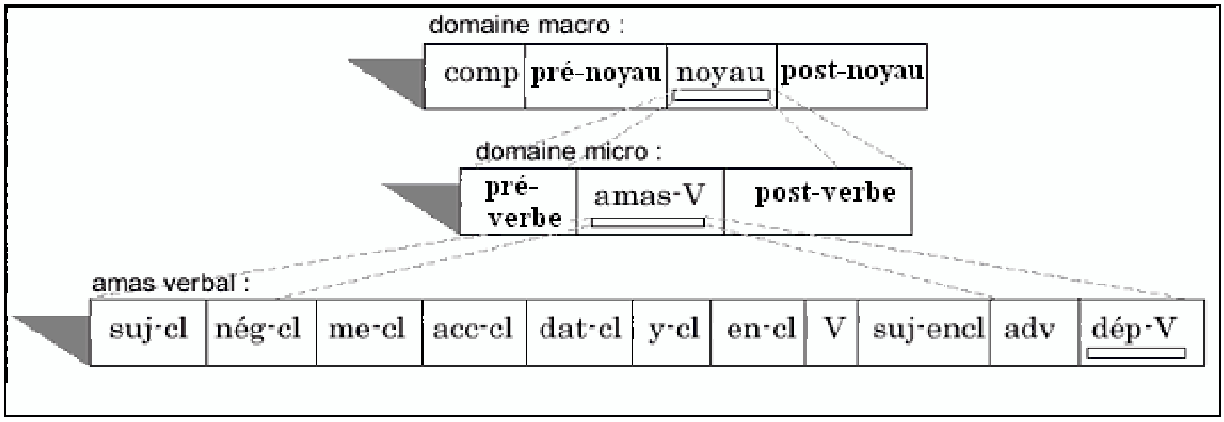


Figure 1 : Modèle de Gerdes & Kahane (à paraître)

Gerdes et Kahane proposent de distinguer topologie et dépendance. Dans la dimension topologique, les auteurs spécifient trois domaines d'enchâssement : l'amas verbal, le domaine micro-syntaxique et le domaine macro-syntaxique. Cela se traduit par le fait qu'un même arbre de dépendance peut se projeter de diverses manières en fonction de la réalisation linéaire des constituants et donc avoir des analyses topologiques distinctes. Avec des termes légèrement différents, nous avons reproduit (Figure 1) le schéma proposé dans Gerdes & Kahane (à paraître).

Nous avons choisi les termes de pré-noyau/post-noyau et pré-verbe/post-verbe au lieu de préfixe/postfixe et sujet/complément afin de conserver une symétrie pré/post dans les deux

[5] Pour un bref aperçu des propriétés sémantiques et pragmatiques, cf. Vogeleer (1998).

domaines et de bien distinguer topologie et dépendance[6]. Pour l'instant, nous laissons de côté l'amas verbal car il nous intéresse moins directement pour résoudre notre problème. Le domaine micro rend compte des éléments non « détachés » et non « extraits » dépendants du verbe. Ils sont dans les champs pré-verbe ou post-verbe en fonction de leur position relativement au verbe (à gauche ou à droite). Le domaine macro, quant à lui, comprend les compléments détachés à gauche (pré-noyau) ou à droite (post-noyau) ainsi que le champ complémenteur (comp). Par exemple, les énoncés a) *Pierre a déjà donné un bonbon à Marie* et b) *A Marie, Pierre a déjà donné un bonbon* auront la même analyse en dépendance mais des représentations topologiques distinctes : *à Marie* sera dans le champ post-verbe pour a) et pré-noyau pour b) (Gerdes & Kahane, à paraître).

Les « subordonnées » dépendantes du verbe et postposées à celui-ci se projettent donc soit dans le champ post-verbal (micro), soit dans le champ post-noyau (macro) si elles sont « détachées ». Quant aux « subordonnées » non dépendantes, il faut aménager le modèle pour que l'on puisse projeter deux arbres de dépendance distincts dans une même entité topologique. Les deux arbres seront associés par l'intermédiaire de la macrostructure topologique. On serait alors en présence d'un cas de constituance sans dépendance.

Il faut aussi distinguer un arbre de base et un arbre associé, l'arbre de base se projetant dans le champ noyau et l'arbre associé dans le champ post-noyau. Cette analyse est assez générale et regroupe notamment nos *quand soudain* mais aussi *quoique, de sorte que*, *parce que* (non dépendants) et d'autres « subordonnées » non dépendantes bloquées en position post-noyau. Certaines « subordonnées » non dépendantes peuvent se trouver avant ou après le noyau. C'est le cas par exemple des « subordonnées » en *puisque*, qui peuvent se trouver dans le champ pré-noyau ou post-noyau. Cette analyse ne se limite évidemment pas aux seules « subordonnées » mais à tous les éléments non dépendants, que leur position soit libre ou contrainte[7].

On obtient donc les deux représentations ci-dessous pour les exemples suivants.

26) *J'arriverai quand le soleil pointera le bout de son nez.*

27) *Je rédigeais mon article quand soudain l'ordinateur planta.*

L'exemple 26) comporte une construction en *quand* dépendante du verbe *arriver*. Il s'agit donc d'une « subordonnée circonstancielle » tout à fait canonique. On peut en faire les représentations topologique et en dépendance suivantes (Figure 2), pour lesquelles, dans un souci de présentation, nous nous sommes limité à la structure plate et seulement aux champs utilisés par les éléments constituant la « subordonnée », pour la représentation topologique. Il faut donc imaginer des structures s'emboîtant les unes dans les autres, comme cela apparaît dans la Figure 1. De plus, nous avons bien mis en évidence le noyau afin de mieux distinguer l'analyse des deux énoncés. Pour la représentation en dépendance, nous nous sommes permis de faire un arbre horizontal, dans un souci de gain de place.

---

[6] Nous disposons d'une version non définitive de Gerdes & Kahane (à paraître), les termes sont donc susceptibles de changer d'ici la parution. La terminologie « complément » et « sujet » nous semblait trop proche de la dépendance, étant des fonctions syntaxiques et non des places relatives à l'amas verbal. De plus, dans le champ sujet, il peut y avoir un complément et réciproquement, ce qui nuit à la clarté l'argumentation.

[7] Pour une description de ces éléments, cf. Blanche-Benveniste et al. (1990, chapitre macro-syntaxe).

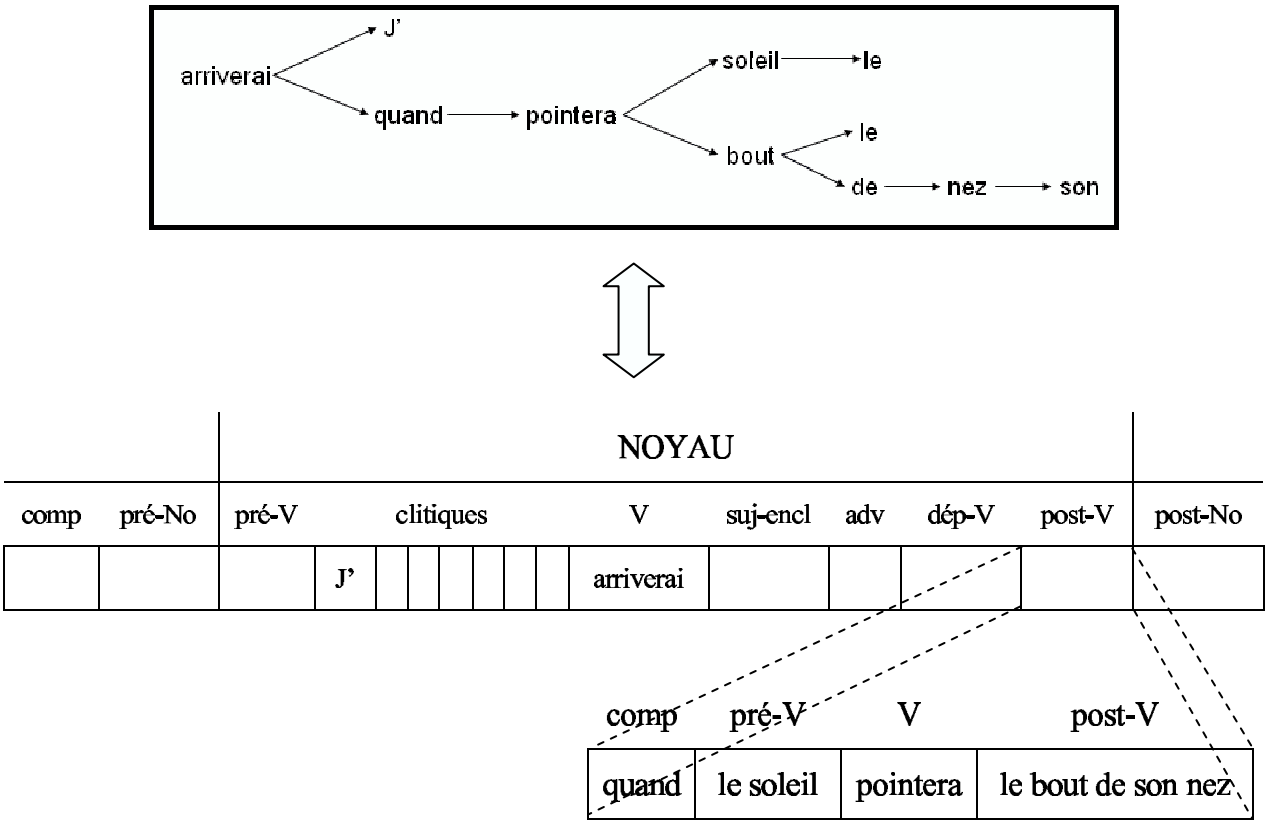


Figure 2 : Arbre de dépendance et structure topologique de l'énoncé 26)

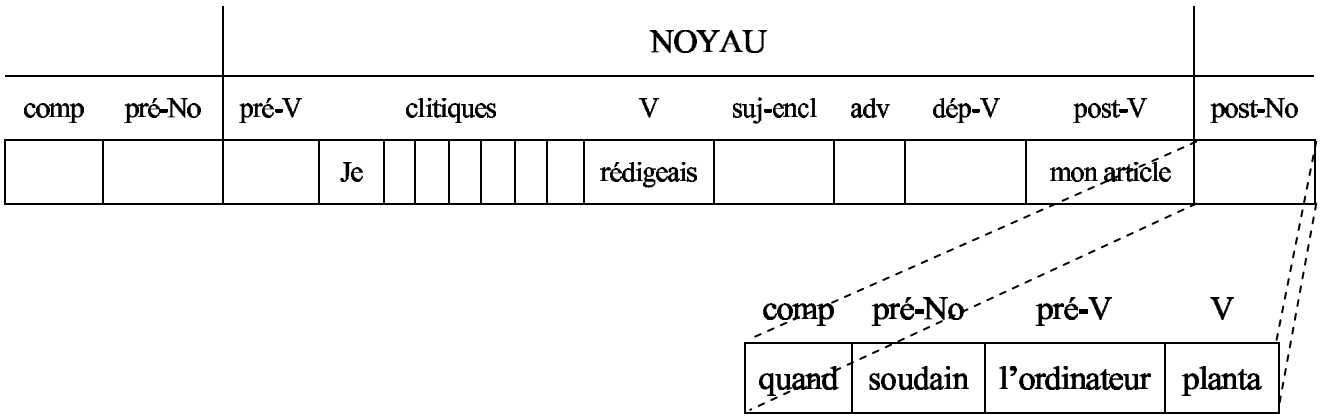


Figure 3 : Structure topologique de l'énoncé 27)

En ce qui concerne l'énoncé 27), il faut distinguer deux arbres de dépendance car il n'y a pas de lien grammatical permettant de mettre l'un dans la dépendance de l'autre. L'un est constitué de *Je rédigeais mon article* et l'autre est formé par *quand soudain l'ordinateur planta*. Etant non dépendante, la construction en *quand soudain* est donc éjectée à l'extérieur du noyau, dans le champ post-noyau. La structure topologique permet donc de regrouper dans un même énoncé deux constructions grammaticalement indépendantes.

Dans ce cadre, on peut donc proposer les analyses suivantes pour les énoncés comportant une « subordonnées » en *quand* :

- Un seul arbre dont la totalité se projette dans le noyau, la « subordonnée » occupant alors le champ post-V.

28) *La vie n'est pas un privilège quand on est l'enfant de deux ennemis.* [Forum]

- Deux arbres indépendants (l'un base, l'autre associé) se projetant sur une seule entité topologique : par défaut, l'arbre de dépendance dont la racine n'est pas un complémenteur va dans le noyau, l'autre élément allant dans la champ post-noyau[8].

29) *Je pensais à nos deux vieux qui devaient dormir tranquillement, quand tout à coup j'entends souffler une machine sur la double voie.* [Contes]

- Un arbre indépendant formant un énoncé autonome[9].

30) *Quand je pense qu'il suffirait que l'un des deux candidats se retire pour que le MNR tombe définitivement dans les oubliettes !* [Forum]

Dans ce modèle, la syntaxe est donc vue comme le résultat de la topologie (constituance) et de la dépendance, traitées indépendamment l'une de l'autre. Cela permet notamment d'analyser des énoncés dont les seules informations issues de la dépendance ou de la constituance sont insuffisantes. Les deux niveaux apportent donc des informations complémentaires.

5 Conclusion - Perspectives

Dans cet article, nous avons essayé de montrer, à travers l'étude de quelques « subordonnées » en *quand,* qu'il était nécessaire de ne pas dériver la relation syntaxique de la seule présence d'une certaine classe de mots (les « subordonnants »). En effet, si l'on utilise des critères autres que la présence d'un item particulier, on s'aperçoit que des constituants introduits par *quand, comme*, *parce que* etc. peuvent ne pas être dépendants d'un élément du contexte et même que certains « subordonnants » peuvent ne jamais « subordonner » (*quoique*, *puisque*…). Dans une logique scientifique de séparer tout ce qui doit l'être, il faut se prémunir contre la démarche qui consiste à postuler une relation syntaxique de dépendance à partir de la présence d'une catégorie grammaticale particulière.

Ce cas de figure n'est généralement pas considéré. Donner un statut à ces éléments et montrer qu'ils ont des propriétés formelles distinctes des subordonnées dépendantes est un premier pas vers leur traitement automatique. De plus, leur prise en compte lors de l'annotation de corpus permettrait une meilleure connaissance des contraintes qui pèsent sur ces constructions qui, pour l'instant, ne disposent pas de descriptions massives et détaillées.

Ce constat fait, le cadre de Gerdes et Kahane nous semble apte à modéliser une telle distinction. Le recours à une relation de type sémantique ou à un lien tantôt avec le verbe, tantôt avec la phrase est moins adapté, selon nous, notamment à cause des contraintes de linéarité observées. Le fait de distinguer une dimension syntaxique spécialement dédiée aux phénomènes d'ordre des unités linguistiques permet d'intégrer directement les propriétés caractéristiques de ces éléments non dépendants. Cela permet aussi de rendre compte de la

---

[8] Afin que les *quand soudain* aillent obligatoirement dans le champ post-noyau, il faudrait prévoir dans le formalisme le marquage des positions bloquées.

[9] Il faudrait prévoir que certaines « subordonnées » puissent être marquées [±Enoncé] en fonction d'un couplage du complémenteur avec un lexème verbal particulier comme *je pense que*.

possibilité d'avoir une dislocation dans ces constructions non dépendantes grâce à l'ouverture d'un champ préfixe, comme le montre cet exemple de français parlé ou ces exemples anglais empruntés à Miller & Weinert (1998 : 95) :

31) *surtout / il y a une scission entre les les organisateurs / quoique François / c'est lui qui a gardé l'aspect millenium /* [ffamcv02]

32) *She switched off the light when into the kitchen came the dog* vs ***When into the kitchen came the dog, it stole a large slice of beef*

La description à grande échelle de ces phénomènes reste encore à faire et les modalités de l'exploitation effective, dans un système de TAL, des propriétés relevées dans la section 3 sont à définir. Mais il nous semble que les perspectives offertes par une telle approche pour traiter les éléments non dépendants sont d'ores et déjà fort prometteuses.

Références

BLANCHE-BENVENISTE Cl. et al. (1990), *Le français parlé : études grammaticales*, coll. Sciences du langage, Paris, CNRS éditions.

BLANCHE-BENVENISTE Cl. et al. (1987), *Pronom et syntaxe. L'approche pronominale et son application à la langue française*, Paris, SELAF.

DEBAISIEUX J-M. (2001), Le fonctionnement de *parce que* en français parlé : étude quantitative sur corpus, Actes de la *1ère Rencontre fribourgeoise de la Linguistique de Corpus Appliquée aux langues romanes*, Tübingen, Gunter Narr Verlag (éd.).

DELORT L. (2004), Relations subordonnantes et coordonnantes pour la désambiguïsation du discours, Actes de *TALN 2004*, pp. 475-484

DUCROT O. et al. (1975), Car, parce que, puisque, *Revue romane*, X, II, pp.248-280.

GERDES K., KAHANE S. (à paraître), L'amas verbal au cœur d'une modélisation topologique du français, *Ordre des mots dans la phrase française, positions et topologie*, Bordeaux.

GREEN G. (1976), Main clause phenomena in subordinate clauses, *Language*, Vol. 52, n°2, pp. 382-397.

HAIMAN J., THOMPSON S. (eds) (1988), *Clause combining in grammar and discourse,* Typological studies in language 18, Amsterdam/Philadelphia, John Benjamins.

HAIMAN J., THOMPSON S. (1984), 'Subordination' in Universal Grammar, Brugmann, Claudia and Monica Macauley (eds.), *Proceedings of the Tenth Annual Meeting of the Berkeley Linguistics Society*, pp. 510-523.

JEANJEAN C. (1985), "Toi quand tu souris" : analyse sémantique et syntaxique d'une structure du français peu étudiée, *Recherches sur le Français Parlé*, 6, pp.131-165.

MILLER J., WEINERT R. (1998), *Spontaneous spoken language. Syntax and discourse*, Clarendon Press, Oxford.

MITHUN M. (2003), On the sentence as the domain of grammar, Communication dans le cadre de la Fédération de recherche *Typologie et universaux du langage*.

MULLER C. (1996), *La subordination en français*, Collection U, Armand Colin.

SABIO F. (2003), L'écriture cérémonieuse chez les enfants : quelques exemples d'intégration grammaticale, *Rivista di psicolinguistica applicata*, special issue edited by Emilia Ferreiro and Marina Pascucci, pp.79-90.

SANDFELD K. (1936), *Syntaxe du français contemporain*, Tome II : *Les propositions subordonnées*, Copenhague/Paris, Librairie E. Droz.

VOGELEER S. (1998), *Quand* inverse, *Revue Québécoise de linguistique*, Vol. 26, n°1, Montréal, pp.79-101.

Des arbres de dérivation aux forêts de dépendance : un chemin via les forêts partagées

Djamé Seddah - Bertrand Gaiffe
Laboratoire LORIA, Equipe Langue et Dialogue
Campus Scientifique, BP 239
F-54506 Vandœuvre-lès-Nancy Cedex
{djame.seddah,bertrand.gaiffe}@loria.fr

Mots-clefs : TAG, analyse syntaxique, sémantique, arbre de dépendance,forêt partagée, forêt de dérivation

Keywords: TAG, syntax, semantic, dependency tree, shared forest, derivation forest

Résumé L'objectif de cet article est de montrer comment bâtir une structure de répresentation proche d'un graphe de dépendance à l'aide des deux structures de représentation canoniques fournies par les Grammaires d'Arbres Adjoints Lexicalisées . Pour illustrer cette approche, nous décrivons comment utiliser ces deux structures à partir d'une forêt partagée.

Abstract This paper aims describing an approach to semantic representation in the Lexicalized Tree Adjoining Grammars (LTAG) paradigm : we show how to use all the informations contained in the two representation structures provided by the LTAG formalism in order to provide a dependency graph.

1 Introduction

Dans cet article[1], nous montrerons comment construire un graphe de dépendance dont les principales propriétés sont de matérialiser des liens syntaxiques non représentés dans l'arbre de dérivation et de décrire toutes les analyses dans une seule structure compacte.

1.1 Les grammaires d'arbres adjoints

Une grammaire LTAG est essentiellement un lexique où chaque lemme est associé à un ensemble d'arbres. Ces arbres, appelés arbres élémentaires, sont manipulables par deux opérations : la substitution et l'adjonction. La substitution opère sur un ensemble restreint d'arbres appelés *arbres initiaux* et correspond à une dérivation hors-contexte. Cette opération est obligatoire, contrairement, d'une façon générale, à l'adjonction qui opère sur des arbres appelés *auxiliaires* et qui correspond à l'insertion d'un arbre spécifique au sein d'un arbre élémentaire (indifféremment initial ou auxiliaire) .

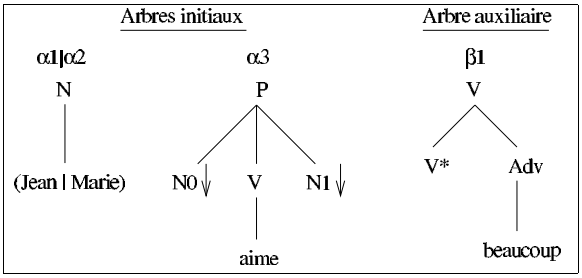


Figure 1: Arbres élémentaires d'une LTAG

Tous les arbres élémentaires sont des projections d'entrées lexicales et par définition décrivent tous les arguments syntaxiques des ancres associées. De fait quand une LTAG suit les principes de bonnes constructions (Abeillé, 1991) tels que le principe de co-occurence prédicat-argument et le principe de minimalité sémantique, les arguments syntaxiques correspondent à des arguments sémantiques. C'est pourquoi l'une des deux structures de représentation des LTAG, **l'arbre de dérivation**, qui formellement n'est que l'enregistrement des opérations résultant d'une analyse syntaxique, peut être vu comme une structure prédicat-argument.
La figure 2 nous montre que l'arbre de dérivation [2] reflète la structure prédicat-argument de la phrase *"Jean aime beaucoup Marie"*.

Étant donné que chaque nœud de l'arbre de dérivation correspond à une projection d'un arbre élémentaire et que chacune de ses branches décrit l'opération de combinaison entre deux nœuds, on associe à ces nœuds une adresse de Gorn[3] indiquant où l'opération a eu lieu. Ainsi, l'arbre de dérivation décrit de façon univoque **l'arbre dérivé** qui est l'arbre syntagmatique d'un énoncé (fig. 2).

[1] *Nous tenons à remercier vivement les reviewers de TALN 2005 pour leurs commentaires et leurs aimables corrections.*

[2] Pour des raisons de simplicité, chaque arbre dont le nom commence par β, resp. α, est un arbre auxiliaire, resp. initial. γ quand ce type est indifférent.

[3] Cette adresse correspond à une séquence d'entiers positifs définie par induction de la façon suivante : la séquence vide notée 0 est l'adresse du nœud racine de l'arbre et p.k est l'adresse du k-ième fils du nœud d'adresse p.

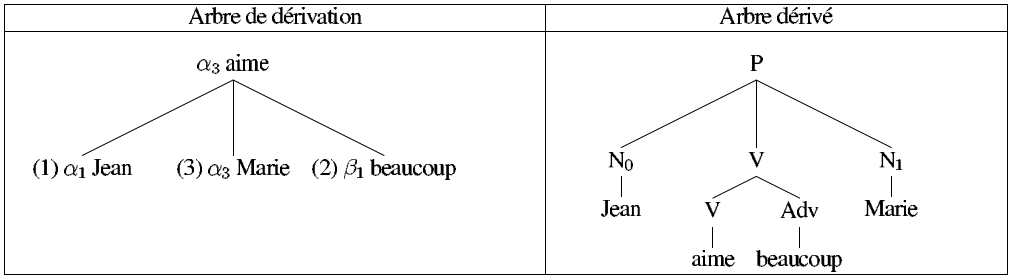


Figure 2: Arbre dérivé et arbre de dérivation pour *Jean aime beaucoup Marie*

1.2 Un formalisme idéal pour une interface syntaxe sémantique idéale ?

Dans un monde idéal il serait possible de travailler à partir de cette structure afin de construire une représentation logique dans l'optique d'une sémantique compositionnelle *à la Montague*. Pour établir les fondements d'un tel modèle il suffirait d'associer un λ-terme à chaque arbre élémentaire et d'utiliser l'arbre de dérivation comme support aux différentes β-réductions induites par les combinaisons syntaxiques de l'analyse. Ainsi, si l'on se base sur la propriété obligatoire de complétion d'un nœud de substitution, on peut considérer ce type de nœuds comme support à des variables argumentales. Comme l'adjonction a comme propriétés d'être non prédictible et optionnelle, on pourrait considérer l'arbre où l'adjonction a lieu comme un argument à la fonction associée à l'arbre auxiliaire qui s'adjoint. De cette façon, on pourrait interpréter l'arbre de dérivation figure 2 de la façon suivante (figure 3) sachant que dans ce mini modèle, on remplace les variables classiques du λ-calcul par des positions argumentales entre crochets[4].

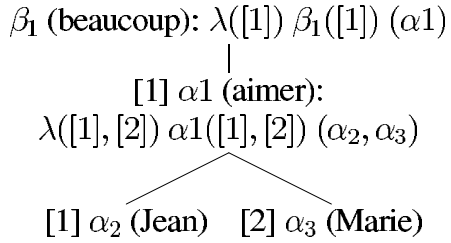


Figure 3: Interface syntaxe sémantique dans un monde idéal

Le problème de ce type d'interfaces syntaxe-sémantique est qu'elles fonctionnent uniquement sur un sous-ensemble restreint du langage avec des structures prédicatives simples. Les LTAG en tant que formalisme de départ ne font pas de distinctions entre adjonctions prédicatives et modifieures (les premières inversant le sens des dépendances sémantiques (Candito & Kahane, 1998)) ; de plus il n'y a pas de mécanismes permettant de résoudre les problèmes d'ambiguïtés des quantifieurs et, pire, certains liens syntaxiques entre des compléments verbaux et leurs sujets n'apparaissent pas dans l'arbre de dérivation.
Les solutions analysant les déficiences de l'arbre de dérivation sont nombreuses (Candito & Kahane, 1998; Schabes & Shieber, 1994; Rambow & Joshi, 1994). Ces solutions peuvent être divisées en deux groupes : celles où l'arbre de dérivation est considéré comme inutilisable, l'arbre dérivé est donc utilisé comme support à une interface syntaxe-sémantique (Gardent & Kallmeyer, 2003; Franck & van Genabith, 2001) ; et celles où l'on considère que les informations portées par l'arbre de dérivation doivent être enrichies, le formalisme LTAG étant modifié

[4]Les λ-termes lexicaux sont ici, bien sûr, simplifiés

sinon remplacé par les TAG Ensemblistes[5] afin de pouvoir gérer les informations de portées des modifieurs (Kallmeyer & Joshi, 1999).

Les solutions basées sur l'arbre dérivé ont pour principal problème d'être basées sur l'unification de structures de traits et sur l'utilisation conjointe d'une sémantique plate comme *liant* via unification des arguments afin d'obtenir une structure prédicative correcte ou une formule logique. Le problème est que chaque événement qui a lieu durant ce processus advient sur un nœud où une dérivation a pris place (adjonction ou substitution) ; ces solutions simulent donc de façon implicite l'arbre de dérivation au sein d'une structure que lui-même décrit sans équivoque[6].

Nous partons du principe que pour établir une interface syntaxe sémantique à partir des LTAG, nous devons d'abord nous assurer que tous les liens argumentaux sont présents dans la structure de représentation.
Sur la base de l'analyse de la problématique des verbes à contrôle, nous proposons une façon de construire un graphe de dépendance à partir des informations contenues tant dans l'arbre dérivé que dans l'arbre de dérivation à l'aide d'une structure décrivant ces deux arbres : la forêt partagée.

2 La problématique posée par les verbes à contrôle

Nous rappellerons brièvement dans cette section la difficulté d'analyse que posent les verbes à contrôle dans le formalisme LTAG[7].
On utilise souvent les verbes à contrôle comme témoin d'un hiatus entre syntaxe et sémantique en TAG pour la simple raison que s'il existe un lien de sous-catégorisation entre un sujet et le verbe qui le sous-catégorise, ce lien devrait être représenté dans l'arbre de dérivation (principe de co-occurence prédicat-argument).
Or l'arbre de dérivation (fig. 5) issu de l'analyse de la phrase *(1) Jean espère dormir* , suivant la grammaire jouet figure 4[8], ne contient pas de lien entre le sujet non réalisé de "dormir" et le sujet qu'il sous-catégorise : "Jean". Or ce lien est présent via la structure de traits dans l'arbre dérivé[9] figure 5.

En réalité la structure que nous voudrions obtenir est un graphe dans lequel ce lien est présent (figure 6).

Dans cette analyse, un verbe à contrôle ancre un arbre auxiliaire (*i.e arbre à contrôle*) qui s'adjoint sur la racine d'un arbre initial ayant au plus un nœud de substitution non-réalisé. On peut donc considérer qu'un verbe à contrôle *transfère* l'un de ses arguments vers l'arbre sur lequel il s'adjoint. L'objectif est donc de formaliser ce processus à travers une opération

[5](Weir, 1988) pour les TAG ensemblistes.

[6]Un autre problème de ce type de solutions basées sur l'arbre dérivé et les structures de traits est l'emploi d'un nombre non fini de traits ce qui a pour conséquence d'accroître la capacité générative du formalisme(Kallmeyer, 2004).

[7]On pourra se reporter à (Abeillé, 1999) pour une analyse complète mise en perspective avec les phénomènes des verbes à montées et comparée à d'autres formalismes.

[8]On notera la position vide dominée par le nœud N de l'arbre α_5, nous appelons ce nœud **"nœud de substitution non réalisée"** ou sujet non réalisé

[9]comme le trait d'accord pour la phrase *(2) Marie espère être belle.* A des fins de lisibilité, nous marquons cet accord par un indice de co-indiciation.

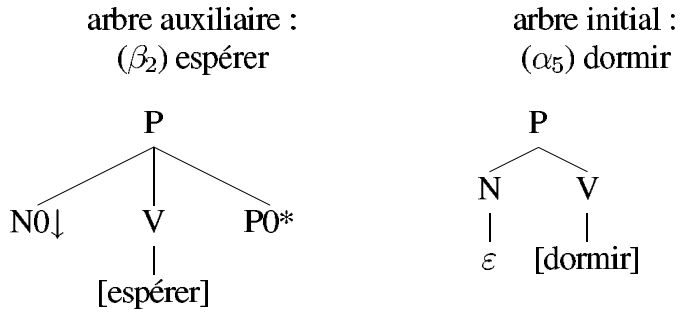


Figure 4: Grammaire jouet verbe à contrôle

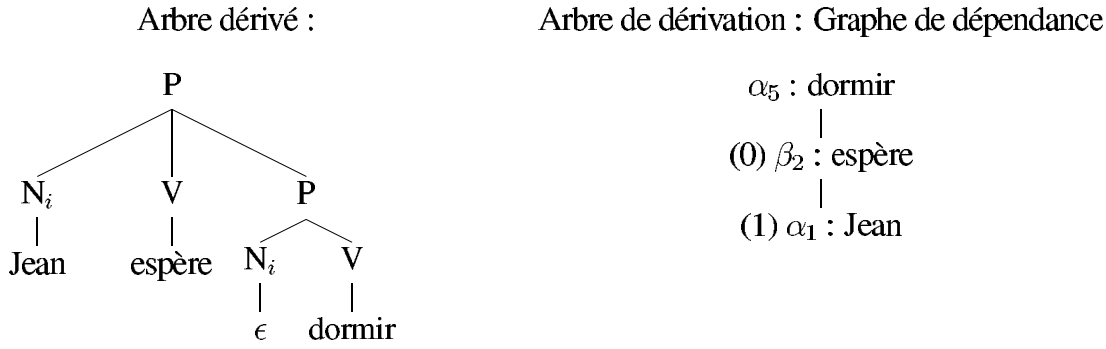


Figure 5: Analyse LTAG : "Jean espère dormir"

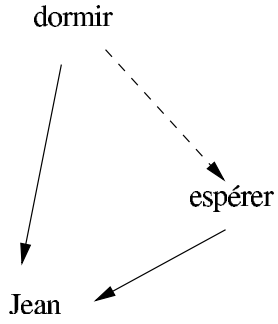


Figure 6: Graphe pour "Jean espère dormir"

appelée : fusion argumentale.

2.1 Informations préalables

Les seules opérations visibles sur un arbre de dérivation sont la substitution et l'adjonction parce qu'elles témoignent du passage d'un arbre à un autre. Si nous voulons faire apparaître ce lien manquant, nous devons donc simuler la dérivation qu'il induit forcément. De quelles informations disposons-nous ?

- Nous connaissons le nombre d'arguments dans un arbre (nœuds de substitution)
- Nous connaissons quel argument doit être transféré car il s'agit d'une information lexi-

cale, **le canevas de contrôle**, que nous marquons directement sur le nœud concerné. Le canevas de contrôle associé au nœud N d'un arbre ancré par un verbe à contrôle transférant son i-eme argument vers le j-eme argument est noté $\boxed{\mathrm{N}_{i\rightarrow j}}$.

Exemple :	*Jean interdit à Marie de dormir*	*Jean espère dormir*
canevas de contrôle	$N_1 \rightarrow N_0$	$N_1 \rightarrow N_0$

- Nous savons qu'un arbre dont un nœud domine une position vide, un argument non réalisé, va recevoir un argument provenant de l'adjonction d'un arbre à contrôle via l'opération de fusion. Ce transfert d'argument marquera la création effective du lien.

Dans ce cas, si nous savons à l'avance que ce lien va être créé, nous pouvons d'ores et déja prévoir la position de ce lien dans le graphe de dérivation. C'est pourquoi nous créons une nouvelle dérivation que nous appelons **Dérivation incomplète** (marquée par la variable libre $\boxed{?}$ dans la figure 7) et qui témoigne d'une substitution non réalisée sur le nœud N d'un arbre α.

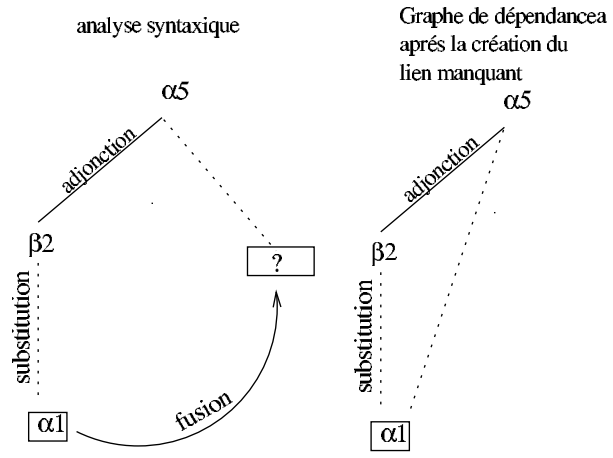


Figure 7: Vision synthétique de l'opération de fusion

Avant de définir formellement l'opération de fusion sous la forme de règles d'inférence, la section suivante va expliciter le processus d'extraction des items de dérivations.

3 Forêt partagées et items de dérivations

Cette section présente l'extraction d'une forêt de dérivation classique, c'est-à-dire sans prise en compte du processus de fusion. La forêt partagée sur laquelle nous basons notre processus résulte de l'intersection d'une grammaire TAG et d'un automate d'entrée, (Vijay-Shanker & Weir, 1993) et (Billot & Lang, 1989). La définition donnée par (Vijay-Shanker & Weir, 1993) est étendue par (Seddah, 2004) afin de lier la reconnaissance du sous arbre dominé par le nœud site d'une adjonction au nœud pied de l'arbre auxiliaire qui s'adjoint. Ceci afin de permettre un parcours synchrone de la forêt partagée. Nous la représentons sous la forme d'une grammaire hors-contexte augmentée d'une pile bornée contenant les adjonctions en cours. Chaque partie de règle correspond à un item de type Cocke-Kasami-Younger de la forme $< N, POS, I, J, Pile >$ où N est un nœud d'un arbre, Pos indique si l'on se situe après une adjonction prévisible ou

pas (marquée $\top$ si une adjonction sur le nœud est possible et $\bot$ si elle ne l'est plus, symbolisée par un point gras en position haute ou basse d'un nœud sur les schémas), I et J sont les indices de début et de fin de la chaîne reconnue par le nœud N et $Pile$ est la pile contenant les appels des sous arbres ayant démarré une adjonction et qui doivent être reconnus par la règle de reconnaissance du pied.
Le processus a lieu en deux temps : 1) La forêt partagée est générée à partir de la grammaire TAG initiale et d'une chaîne d'entrée 2) la forêt partagée est ensuite parcourue afin d'en extraire les dérivations.
L'extraction est simple : si une règle témoignant d'une dérivation est validée, un item de dérivation est inféré.

3.1 Forme des items de dérivation

A chaque occurrence d'une règle de dérivation, nous produisons un item de dérivation en témoignant. Cet item que nous appelons **Deriv** est de la forme $< N, \gamma, \alpha, type >$ avec N, le nœud qui reçoit la substitution ; γ, l'arbre de ce nœud, α l'arbre qui va s'y substituer, $type$ est le type de dérivation.
Trois dérivations sont possibles :

- une première évidente lors d'une substitution : on passe d'un nœud N d'un arbre γ à un nœud d'un arbre α :
  Item de substitution : $< N, \alpha, \gamma, subst >$

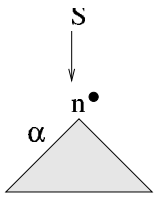
- lors d'une adjonction : on passe d'un nœud d'un arbre γ au nœud racine d'un arbre β. Nous ne considérons pas comme une dérivation le passage du nœud pied de l'arbre β au sous arbre du nœud site de l'adjonction de l'arbre α. Cette information est de fait évidemment redondante : une adjonction étant une insertion, il doit y avoir retour et analyse de l'arbre appelant.
  Item d'adjonction : $< N, \beta, \gamma, adj >$

- lors de la reconnaissance d'un arbre initial α ayant portée sur toute la chaîne, on crée une dérivation de type "tête".
  Item axiomes : $< N, \alpha, -, - >$

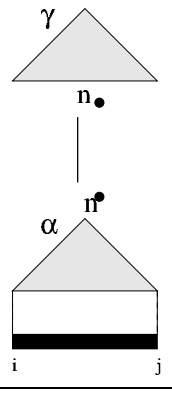
3.2 Règles d'inférence de l'algorithme d'extraction du graphe de dérivation

Pour toute règle $r \in R$ se situant sur le chemin succès, nous appliquons les règles d'inférence suivantes :

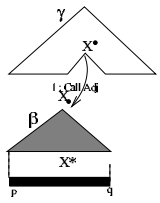
- **Dérivation de l'axiome**
  Si la règle r = S $\longrightarrow < \top, N_\alpha, 0, n, -, -, \emptyset >$ est validée lors de l'exécution de la grammaire, alors nous produisons l'item $< N, \alpha, -, - >$.

$S \longrightarrow < \top, N_\alpha, 0, n, -, -, \emptyset >$	
Reconnaissance d'un axiome $\dfrac{S \longrightarrow < \top, N_\alpha, 0, n, -, -, \emptyset >}{< N, \alpha, -, - >}$ *avec n=longueur de la chaine*	

- **Dérivation d'une substitution**
  Si la règle $r = < \bot, n_\gamma, i, j, -, -, Pile > \longrightarrow < \top, n_\alpha, i, j, -, -, Pile >$ est validée alors nous produisons l'item $< N, \alpha, \gamma, subst >$.

$< \bot, n_\gamma, i, j, -, -, Pile > \longrightarrow < \top, n_\alpha, i, j, -, -, Pile >$	
Reconnaissance d'une substitution $\dfrac{< \bot, n_\gamma, i, j, -, -, Pile > \longrightarrow < \top, n_\alpha, i, j, -, -, Pile >}{< N, \alpha, \gamma, subst >}$	

- **Dérivation d'une adjonction**
  Si la règle $r = < \top, n_\gamma, i, j, p, q, Pile > \longrightarrow < \top, n_\beta, i, j, p, q, Pile' >$ est validée alors nous produisons l'item $< N, \beta, \gamma, adj >$.

$< \top, n_\gamma, i, j, p, q, Pile > \longrightarrow < \top, n_\beta, i, j, p, q, Pile' >$	
Reconnaissance d'une adjonction $\dfrac{< \top, n_\gamma, i, j, p, q, Pile > \longrightarrow < \top, n_\beta, i, j, p, q, Pile' >}{< N, \beta, \gamma, adj >}$	

Les variables Pile et Pile′ représentent l'état de la pile d'appel des adjonctions lors de l'exécution d'une forêt produite. Leurs valeurs n'est pas prise en compte par cette règle d'inférence, car l'appel d'une adjonction est une opération hors contexte. Quelque soit le contenu de ces 2 variables, la dérivation sera celle témoignant de l'adjonction de l'arbre β sur le nœud X de l'arbre γ.

3.3 item de dérivation pour la dérivation incomplète

Nous avons décrit (section 2.1) un nœud dominant une position vide comme témoignant d'une substitution non réalisée et, par conséquent témoin d'une **dérivation incomplète**. La règle d'inférence suivante définit cette nouvelle dérivation :

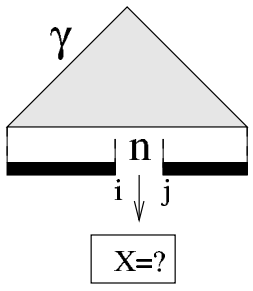
- Si la règle $< \bot, N_\gamma, i, j, -, -, Pile > \longrightarrow$ vrai est validée alors on produit la dériva-

tion suivante :

$< \bot, N_\gamma, i, j, -, -, Pile > \longrightarrow$ vrai

Dérivation d'une substitution non-réalisée

$$\frac{< \bot, N_\gamma, i, j, -, -, Pile >}{< N, \alpha, X, subst >} \textit{ avec } X, \textit{ non instancié}$$



L'ensemble des items de dérivation est généré dans un *chart* spécifique et correspond à l'ensemble des analyses possibles. Partant d'un item axiome, nous décrivons très exactement un arbre de dérivation. Si la grammaire suit les recommandations de (Rambow & Joshi, 1994), cet arbre peut être vu comme un arbre de dépendance, ainsi l'ensemble des arbres décrits par cette forêt d'items est dans ce cas une forêt de dépendance.

4 Formalisation du processus de fusion argumentale

L'analyse de l'adjonction d'un arbre à contrôle sur un arbre élémentaire met en jeu 3 dérivations : la dérivation $D1$ témoignant de la substitution d'un arbre α_1 sur le nœud N,dont le canevas de contrôle est $\boxed{N_{i \rightarrow j}}$, d'un arbre à contrôle β_2 ; la dérivation $D3$ témoignant de l'adjonction de l'arbre à contrôle β_2 sur la racine d'un arbre élémentaire γ[10] et la dérivation incomplète $D2$ témoignant de la présence d'un nœud N_j de substitution non réalisée sur l'arbre γ. L'opération de fusion est donc la règle d'inférence permettant de simuler la création du lien manquant via un nouvel item de dérivation $D4$ qui remplace l'item de dérivation incomplète $D3$.

$$\frac{D3 :< X, \beta_2, \gamma, adj > \; D1 :< N_{i \rightarrow j}, \alpha_1, \beta_2, subst > \; D2 :< N_j, \boxed{?}, \gamma, subst >}{D4 :< N_j, \alpha_1, \gamma, subst >}$$

5 Conclusion

Ce travail a été implémenté dans (Seddah, 2004). Sa caractéristique principale est de faire un usage intensif des propriétés des forêts partagées afin de parcourir simultanément les nœuds de l'arbre dérivé et de l'arbre de dérivation. Les nœuds qui sont utilisés lors de l'opération de fusion proviennent de l'arbre de dérivation pour les dérivations complètes et de l'arbre dérivé pour la dérivation incomplète. Si l'on considère que chaque item de dérivation correspond à un argument sémantique, le graphe de dérivation correspond à un graphe sémantique similaire au DSyntS de la théorie sens-texte (Mel'Cuk, 1997). Une extension du modèle pour traiter des phénomènes liés aux coordinations elliptiques est actuellement en cours d'élaboration. Pour bâtir ce modèle nous avons dû modifier légèrement le formalisme TAG pour inclure une information lexicale nécessaire aux règles d'inférence. La méthodologie consistant à travailler

[10] L'arbre élémentaire γ généralise le cas présenté figure 7 où la fusion opérait sur l'arbre initial α_5.

systématiquement au coeur de la forêt partagée nous permet de travailler sur toutes les analyses à la fois et donc de générer tous les graphes de dépendance au sein d'une forêt compacte.

References

ABEILLÉ A. (1991). *Une grammaire lexicalisée d'arbres adjoints pour le français.* PhD thesis, Paris 7.

ABEILLÉ A. (1999). Verbes "à monté" et auxiliaires dans une grammaire d'arbres adjoints. *LINX,Linguistique Institut Nanterre Paris X.*

BILLOT S. & LANG B. (1989). The structure of shared forests in ambigous parsing. In *33rd Conference of the Association for Computional Linguistics (ACL'89).*

CANDITO M.-H. & KAHANE S. (1998). Can the TAG derivation tree represent a semantic graph ? In *Proceedings TAG+4, Philadelphie*, p. 21–24.

FRANCK A. & VAN GENABITH J. (2001). Gluetag : Linear logic based semantics for LTAG -and what it teaches us about LFG and LTAG-. In *Proceddings of the LFG01 Conference, University of Hong Kong, Hong Kong.*

GARDENT C. & KALLMEYER L. (2003). Semantic construction in feature-based tag. In *Procedings of EACL 2003.*

KALLMEYER L. (2004). LTAG Semantics with Semantic Unification. In *Proceedings of TAG+7, To appears.*

KALLMEYER L. & JOSHI A. (1999). Factoring predicate argument and scope semantics: Underspecified semantics with LTAG. In *Proceedings of the 12th Amsterdam Colloquium, December.*

MEL'CUK I. (1997). *Vers une linguistique Sens-Texte. Leçon inaugurale.* College de France, Paris.

RAMBOW O. & JOSHI A. K. (1994). *A Formal Look at Dependency Grammar and Phrase Structure Grammars, with Special consideration of Word Order Phenomena.* Leo Wanner, Pinter London, 94.

SCHABES Y. & SHIEBER S. (1994). An alternative conception of tree-adjoining derivation. *Computational Linguistics*, **20**(1), 91–124.

SEDDAH D. (2004). *Synchronisation des connaissances syntaxiques et sémantiques pour l'analyse d'énoncés en langage naturel à l'aide des grammaires d'arbres adjoints lexicalisées.* PhD thesis, Université Henry Poincaré, Nancy.

VIJAY-SHANKER K. & WEIR D. (1993). The use of shared forests in tree adjoining grammar parsing. In *EACL '93*, p. 384–393.

WEIR D. (1988). *Characterizing Midly Context-Sensitive Grammar Formalisms.* PhD thesis, University of Pennsylvania.

Évaluation des Modèles de Langage n-gramme et n/m-multigramme

P. Alain, O. Boëffard
IRISA – Université de Rennes 1 / ENSSAT
6, rue de Kerampont, 22305 Lannion
{pierre.alain,olivier.boeffard}@irisa.fr

Mots-clefs : Modèles de Langage statistiques, n-gramme, multigramme, évaluation

Keywords: Statistical Language Models, n-grams, phrase multigrams

Résumé Cet article présente une évaluation de modèles statistiques du langage menée sur la langue Française. Nous avons cherché à comparer la performance de modèles de langage exotiques par rapport aux modèles plus classiques de n-gramme à horizon fixe. Les expériences réalisées montrent que des modèles de n-gramme à horizon variable peuvent faire baisser de plus de 10% en moyenne la perplexité d'un modèle de n-gramme à horizon fixe. Les modèles de n/m-multigramme demandent une adaptation pour pouvoir être concurrentiels.

Abstract This paper presents an evaluation of statistical language models carried out on the French language. We compared the performance of some exotic models to the one of the more traditional n-gram model. The experiments show that the variable n-gram models can drop more than 10% of the average perplexity for a fixed n-gram model. n/m-multigram models require an adaptation to be able to compete.

1 Introduction

La modélisation du langage est un problème crucial et très largement abordé en traitement automatique de la langue écrite ou parlée[1]. À partir de l'observation de séquences de mots, il s'agit de construire un modèle dont l'objectif est de prédire avec succès de nouvelles séquences. On peut distinguer déjà deux problèmes, d'une part celui du choix du modèle et de sa méthodologie de construction et d'autre part celui de la méthodologie d'évaluation d'un modèle de langage. Concernant le premier point, on peut distinguer des approches déterministes qui tiennent compte de l'organisation profonde des mots liées notamment à la syntaxe, des approches probabilistes qui s'intéressent essentiellement à la forme de surface (Rosenfeld, 2000).

L'évaluation est un point relativement délicat dans la mesure où elle peut être dépendante du modèle choisi. La mesure la plus communément adoptée consiste à calculer l'entropie croisée entre un modèle de langage et la distribution réelle des données observées, mais inconnue. En supposant que les données suivent une distribution stationnaire et ergodique, le calcul de l'entropie-croisée peut être estimé à partir d'un corpus suffisamment grand[2]. La perplexité d'un modèle de langage n'est qu'une autre manière de représenter le degré d'incertitude d'un modèle et se calcule directement à partir de l'entropie-croisée du modèle sur un jeu de phrases de test. Pour un mot à prédire, la valeur de la perplexité représente le

[1] On peut citer le domaine de la reconnaissance automatique de la parole mais aussi celui de la reconnaissance de texte manuscrit ou encore celui de la traduction automatique.

[2] Théorème de Shannon-MacMillan-Brieman, (Shields, 1998). En respectant ces hypothèses de stationnarité et d'ergodicité, un corpus de parole de longueur finie peut refléter la distribution réelle des données.

nombre d'hypothèses moyennes de branchement[3]. Plus la perplexité est faible, plus le facteur moyen de branchements d'un mot vers un autre est bas et plus le modèle de langage est efficace. Pour les modèles n-gramme, un mot est prédit en tenant compte d'un historique relativement limité des mots qui le précèdent. Ces modèles connaissent finalement très peu des raisons profondes de l'organisation des mots dans une phrase. En revanche, l'utilisation de probabilités conditionnelles et un apprentissage réalisé à partir de quelques millions de phrases permettent d'obtenir de bonnes performances. Leur principal défaut réside dans la complexité spatiale sous-jacente. Théoriquement, plus la séquence de l'historique du modèle s'allonge (n augmente), plus le modèle répartit efficacement la masse de probabilités sur des mots qui reviennent souvent après une valeur particulière de l'historique. Cependant, plus n augmente, plus les observations se raréfient compte-tenu de la nature hyperbolique de la distribution de ces événements[4]. Pour des situations expérimentales réelles, les valeurs courantes de n dépassent rarement 4 (Siu & Ostendorf, 2000). De nombreuses solutions ont été apportées au problème de l'explosion combinatoire et à celui de la raréfaction des événements (Rosenfeld, 2000). Des techniques de lissage permettent notamment de répondre à la difficulté de l'estimation d'une distribution de probabilité lorsque les événements sont rares. On peut citer le principe du lissage qui n'effectue l'estimation des points de la densité au sens du maximum de vraisemblance que pour des événements dont l'occurrence est supérieure à un seuil de *cut-off*. Une partie de la masse de probabilité est répartie sur des événements dont l'occurrence est inférieure au seuil, (Katz, 1987). (Chen & Goodman, 1999) propose une évaluation des principales techniques de lissage les plus utilisées.

Les modèles de n-gramme pour lesquels la longueur de l'historique est variable sont une alternative aux n-gramme classiques pour lesquels la longueur de l'historique reste fixe. Le principe consiste à ne pas retenir un historique de longueur n si la contribution du n-gramme correspondant n'améliore pas la performance du modèle. Toute la difficulté réside dans la décision d'abandon d'un n-gramme pour un $(n-k)$-gramme avec $1 \leq k < n$, (Niesler & Woodland, 1994)(Siu & Ostendorf, 2000).

Les modèles multigramme sont des modèles de type n-gramme où la tête peut avoir une longueur supérieure à 1.(Bimbot *et al.* , 1995)(Deligne & Bimbot, 1995) présentent un cadre théorique pour des multigramme formés sur des modèles d'uni-gramme (longueur d'historique nulle). Les expériences rapportées concernent une application avec un vocabulaire limité de 900 mots pour un corpus d'apprentissage de 100 000 phrases et un corpus de test de 1 000 phrases (dont 52 occurrences de mots hors-vocabulaire). Les modèles de type multigramme obtiennent une perplexité meilleure que les n-gramme classiques lorsque $n > 3$. Compte-tenu de la taille relativement limitée des corpus, les conclusions sont difficilement transposables directement sur des corpus plus importants. (Deligne & Sagisaka, 2000) se place dans un contexte de multigramme de classes de mots sur des modèles de bi-gramme. Les expériences sont menées avec un vocabulaire d'environ 3 000 mots, 100 000 phrases pour le corpus d'apprentissage et environ 700 phrases pour le test. Deux types de mesure sont rapportés : d'une part la perplexité pour les modèles de type multigramme et d'autre part le taux d'erreur d'un système de reconnaissance de la parole. Les résultats entre multigramme et n-gramme classiques (bi- et tri-gramme) sont difficilement comparables. En effet, pour ces derniers, les valeurs de perplexité sont absentes et les modèles de n-gramme semblent avoir été non réduits[5]. (Zitouni, 2002) propose des modèles de multigramme où les probabilités de co-occurrence de mots sont conditionnées par rapport à des classes. Les expériences concernent deux années du journal "Le Monde" (années 1987 et 1988) pour un vocabulaire de 20 000 mots. L'utilisation de ces multigramme permet de réduire de 7% la perplexité des tri-gramme classiques. Encore une fois, il est difficile de retrouver sur cette expérience une comparaison entre modèles à nombre de paramètres constant.

Cet article propose une étude expérimentale sur les performances relatives des modèles de language

[3]Il s'agit d'une moyenne géométrique.

[4]Il s'agit de distributions à queue lourde où beaucoup d'événements sont extrêmement rares et peu sont très fréquents. La loi de Zipf est un cas particulier de lois puissances caractéristiques de ce phénomène.

[5]Le comportement d'un n-gramme est non-linéaire, il est possible de réduire de façon importante le nombre de paramètres sans trop dégrader ses performances.

de type n-gramme à horizon fixe, à horizon variable et multigramme. La section 2 présente un cadre théorique pour ces trois types de modèles de langage statistiques. La section 3 expose la problématique d'une telle expérimentation ainsi que nos hypothèses de travail. Une évaluation a été menée sur environ un million de phrases en français. La section 4 décrit la méthodologie suivie. La section 5 expose les expériences mises en œuvre. Enfin, la section 6 présente les résultats et une interprétation du comportement des modèles en fonction des données traitées.

2 Cadre théorique

Un modèle de langage statistique est un ensemble de distributions de probabilité sur des séquences observées de symboles. Comme, en pratique, il est impossible de caractériser de telles distributions, les modèles de langage se distingueront entre eux par les hypothèses choisies pour réduire la complexité combinatoire et améliorer leur capacité de généralisation. Après une présentation des notations utilisées, nous discutons du modèle de n-gramme à horizon fixe, du modèle de n-gramme à horizon variable et enfin du modèle n/m-multigramme.

Soit une séquence de mots $W = (w_1, w_2, ..., w_N)$ avec w_i une variable représentant le mot de rang i dans la séquence W. Les valeurs possibles pour w_i appartiennent à un vocabulaire $\mathcal{V}$. Il peut s'agit souvent d'un vocabulaire fermé dans le cadre de systèmes de dialogue, nous considérons ici l'étude de la langue naturelle, nous choisissons un vocabulaire ouvert. Nous pouvons décrire cette séquence par une suite de variables aléatoires w_i. La probabilité conjointe des variables de la séquence W peut se développer de la manière suivante en faisant apparaître des probabilités conditionnelles :

$$p(W) = p(w_1) \times \prod_{i=2}^{N} p(w_i|w_1, ...w_{i-1}) \quad (1)$$

L'objectif d'un modèle de langage consiste à calculer cette probabilité conjointe, c'est-à-dire à estimer des valeurs pour chacune des probabilités conditionnelles. L'estimation de ces probabilités conditionnelles est en pratique impossible car le nombre de paramètres croît de manière exponentielle avec la longueur de la suite de mots. Pour contrer cette difficulté, un modèle de langage pose une probabilité conditionnelle approchée $p^*(.)$ en simplifiant la loi conjointe, équation 1.

On note $\mathcal{G}$ l'ensemble des groupes de mots formés sur le vocabulaire $\mathcal{V}$. On note $\mathcal{S}$ l'ensemble des séquences formées sur les éléments de $\mathcal{G}$. On note $\mathcal{S}^* \subset \mathcal{S}$, l'ensemble des séquences de $\mathcal{S}$ qui correspondent à W. On note S une séquence particulière de $\mathcal{S}^*$. Par exemple, pour $W = (w_1, w_2, w_3)$, on a :

$$\mathcal{S}^* = \begin{cases} [w_1][w_2][w_3] \\ [w_1, w_2][w_3] \\ [w_1, w_2, w_3] \\ [w_1][w_2, w_3] \end{cases}$$

On note $|S|$ le nombre de groupes de mots dans la séquence S. Soit k le groupe de mots de rang k dans la séquence S, on note $i(S(k))$ l'indice dans la séquence W du premier mot de $S(k)$. On note $l(S(k))$ le nombre de mots de $S(k)$.

On note $h_{i,j}(W)$ la chaîne des variables aléatoires représentant l'apparition conjointe de tous les mots w_u de W pour $u \in [i, i + (j - 1)]$.

$$h_{i,j}(W) = \begin{cases} w_i, w_{i+1}, ..., w_{i+(j-1)} & \text{si } i + (j-1) \leq N \\ w_i, w_{i+1}, ..., w_N & \text{sinon} \end{cases}$$

On définit également l'opérateur $t_{i,j}(W)$ qui représente les j mots précédents le mot w_i. On a donc $t_{i,j}(W) = h_{i-j,j}(W)$. $t_{i,j}(W)$ correspond à un horizon ou historique (un groupe de mots qui précède

l'observation d'un mot). $h_{i,j}(W)$ correspond à la tête d'un paramètre du modèle de langage (pour les modèles n-gramme à horizon fixe ou variable, la variable aléatoire de tête est dégénérée, et ne contient qu'un seul mot). Les modèles de langage cherchent d'une part à réduire au maximum la longueur d'un horizon (minimisation du nombre de paramètres) et d'autre part, pour un horizon donné, à estimer la distribution de probabilité des historiques pour calculer la probabilité d'apparition du mot w_i. Soit W associée à une séquence de découpage S, la loi conjointe estimée par le modèle de langage peut alors se réécrire sous la forme suivante avec n l'ordre du n-gramme :

$$p_S^*(W) = p(h_{i(S(1)),l(S(1))}(W)) \times \prod_{k=2}^{|S|} p(h_{i(S(k)),l(S(k))}(W)|t_{i(S(k)),n-1}(W)) \quad (2)$$

2.1 Les modèles n-gramme à horizon fixe

Pour un n-gramme à horizon fixe, on fait une hypothèse d'indépendance conditionnelle du mot w_i avec les mots présents dans la séquence à une distance de plus de $n-1$ mots (pour $n=2$, ce modèle est un modèle de bi-gramme ; la probabilité $P(W)$ correspond à celle d'une chaîne de Markov. Pour $n=3$, on parle de tri-gramme et pour $n=4$ de quadri-gramme). Comme nous l'avons déjà souligné, ce modèle est très simple, mais le nombre de paramètres croît de manière exponentielle avec n. Pour cette raison, les modèles de n-gramme les plus utilisés le sont pour des valeurs de n de l'ordre de 3 ou 4. Pour corriger le problème des événements rares, il existe des techniques de lissage des probabilités conditionnelles, couplées à des techniques de *back-off* permettant de corriger celles d'événements manquants lors de l'apprentissage, (Katz, 1987). Cette correction s'effectue en pondérant la probabilité du $(n-1)$-gramme par un coefficient de *back-off* de telle manière que la distribution de probabilité des n-gramme somme toujours à 1. Le terme produit de l'équation 2 se simplifie alors :

$$p(h_{i(S(k)),l(S(k))}(W)|t_{i(S(k)),n-1}(W)) \triangleq p(h_{i,1}(W)|t_{i,n-1}(W)) \quad (3)$$

Pour ce modèle, $S=W$, on obtient simplement :

$$p_{ML}(W) = p_S^*(W)$$

2.2 Les n-gramme à horizon variable

Forcer l'estimation du terme produit de l'équation 2 à un historique de longueur n introduit un double biais. D'une part les occurrences sont plus faibles, on a donc tendance à faire du lissage et à être moins précis. D'autre part, on introduit des distributions conditionnelles sur w_i qui ne servent pas à grand chose (augmentation injustifiée du nombre de paramètres). Autoriser une variation de la longueur de l'historique pour prédire w_i permet de régler ce problème de sur-apprentissage. Les n-gramme à horizon variable définissent une probabilité en adaptant une longueur d'historique optimale en fonction de w_i. L'approche traditionnelle pour ce type de modèles consiste à déterminer au moment de l'apprentissage les longueurs optimales à retenir, (Bonafonte & Mariño, 1996)(Siu & Ostendorf, 2000). Dans cette situation un n-gramme est remplacé par un $(n-k)$-gramme avec $1 \leq k < n$. Les n-gramme à horizon variable peuvent apparaître intéressants pour un double enjeu : d'une part à nombre de paramètres fixé, il peuvent répondre à une amélioration de la performance des n-gramme à horizon fixe et d'autre part, à perplexité fixée, ils peuvent être utiles à la diminution du nombre de paramètres d'un modèle de langage.

Au moment du test, lors du calcul de la perplexité d'une phrase, pour ce modèle de n-gramme à horizon variable, le terme produit de l'équation 2 s'écrit :

$$p(h_{i(S(k)),l(S(k))}(W)|t_{i(S(k)),n-1}(W)) \triangleq \max_{1 \leq v \leq n-1} \{p(h_{i,1}(W)|t_{i,v}(W))\} \quad (4)$$

Cette écriture signifie que pour chaque mot w_i à prédire, on cherche à maximiser la probabilité en se basant sur des modèles allant du bi-gramme au n-gramme (équation 3). Pour ce modèle, $S = W$, on obtient simplement :

$$p_{ML}(W) \quad = \quad p^*_S(W)$$

2.3 Les n/m-multigram

Un n/m-multigramme correspond à une probabilité conditionnelle où la tête du n-gramme peut être plus longue qu'un mot unique. m représente le nombre maximum de mots dans un groupe de mots en tête. Lors du test du modèle de langage, pour une découpe S donnée, nous cherchons la meilleure probabilité suivant l'équation :

$$p(h_{i(S(k)),l(S(k))}(W)|t_{i(S(k)),n-1}(W)) \quad \triangleq \quad \max_{1\leq u\leq m, 1\leq v\leq m\times(n-1)} \{p(h_{i,u}(W)|t_{i,v}(W))\} \qquad (5)$$

Il suffit ensuite de prendre la meilleure solution sur toutes les séquences $S \in \mathcal{S}^*$:

$$p_{ML}(W) \quad = \quad \arg\max_{S\in\mathcal{S}^*}\{p^*_S(W)\}$$

3 Problématique et hypothèses méthodologiques

Notre objectif est de vérifier l'intérêt des modèles n-gramme à horizon variable par rapport à des modèles à horizon fixe et à des modèles de type n/m-multigramme. La difficulté de mise en œuvre d'une telle évaluation réside dans le problème du contrôle explicite des paramètres lors de la construction des modèles. Différents facteurs sont responsables de la qualité d'un modèle de langage. Certains influent directement le processus d'apprentissage alors que d'autres déterminent la mesure de performance d'un modèle.

Tout d'abord l'estimation des probabilités conditionnelles provient directement de la détection de n-uplets. Avec peu de séquences, on défavorise notamment les modèles de n-gramme d'ordre supérieur. Le nombre de paramètres d'un modèle de langage de type n-gramme est proportionnel à $|\mathcal{V}|^n$. Le *cut-off* est une technique simple et relativement efficace pour limiter le nombre de paramètres (Chen & Goodman, 1999). Il s'agit de ne pas retenir les n-uplets qui apparaissent sous un seuil d'occurrence. Ainsi, un *cut-off* à 1 signifie qu'un mot doit apparaître au moins 2 fois pour être intégré au modèle de langage. Cependant, compte-tenu de la forme des distributions de probabilité (fonction puissance), la réduction conséquente du nombre de paramètres n'est pas linéaire en fonction de la valeur de *cut-off*. En introduction, nous avons souligné le rôle de la perplexité comme outil de mesure de la qualité d'un modèles de langage.

Un autre facteur clé que l'on doit maintenir entre les différents modèles de langage pour pouvoir comparer les valeurs de perplexité est le nombre de mots hors-vocabulaire. Plus la taille du vocabulaire est faible (et donc plus le nombre de paramètres est faible), plus le taux des mots hors-vocabulaire augmente avec des valeurs de perplexité qui s'améliorent. Il s'agit d'un facteur calculé a posteriori, une fois le modèle construit. Il est donc difficile d'intervenir explicitement sur cette valeur.

Le calcul de la perplexité peut varier notamment par la prise en compte ou non des mots hors-vocabulaire sur l'ensemble de test. On peut décider de ne pas prédire un mot hors vocabulaire ; dans ce cas l'accumulation de la perplexité est plus faible mais le nombre de mot prédit n'augmente pas. Le calcul de la perplexité fait intervenir une hypothèse de stationnarité et d'ergodicité qu'il faudrait vérifier en pratique. La performance d'un modèle de langage dépend donc étroitement du couple ensemble d'apprentissage/ensemble de test. Il faut que ces ensembles contiennent un nombre suffisant de

séquences pour pouvoir conclure à des résultats stables. Au cours de nos expériences, nous avons essayé de minimiser l'influence de chacun de ces facteurs de manière à favoriser la comparaison entre structures de modèles (n-gramme à horizon fixe, n-gramme à horizon variable et n/m-multigramme).

Nous avons considéré les hypothèses méthodologiques suivantes. Une année du journal "Le Monde" a été choisie comme univers linguistique (année 1997). Après extraction des phrases et tirage aléatoire, ce corpus est reparti en deux sous-corpus : 70% pour le corpus d'apprentissage et 30% pour le corpus de test. Le choix d'un corpus fixe est suffisant pour valider une comparaison entre modèles, mais ne permettra pas de conclure sur la robustesse des résultats. Des analyses complémentaires seront donc nécessaires. Nous avons considéré trois ensembles de mots : un premier vocabulaire à 3 000 mots, un deuxième à 30 000 mots et un dernier à 60 000 mots (il s'agit à chaque fois des plus fréquents sur l'ensemble d'apprentissage). Les valeurs de perplexité et les taux de mots hors vocabulaire dépendent directement de ces trois ensembles. Nous avons cherché à contrôler explicitement le nombre de paramètres de nos modèles. Deux approches complémentaires ont été mises en œuvre : d'une part par application de seuils de coupure sur les différents types de n-gramme et d'autre part par la conservation des co-occurrences de m-uplets de mots les plus fréquentes pour les n/m-multigramme. Dans le premier cas, nous balayons un spectre de valeurs de *cut-off* et nous observons a posteriori le nombre de paramètres. Ce nombre nous sert ensuite à ajuster le nombre de multigramme autorisés à entrer dans le vocabulaire et se placer ainsi à nombre de paramètres constant (avec une tolérance de 1%). La perplexité calculée ne tient pas compte des mots hors vocabulaire qu'ils soient présents dans la tête ou dans l'historique d'un n-gramme.

Notre système de référence est celui des n-gramme classiques (que nous avons nommé n-gramme à horizon fixe). Nous avons choisi des valeurs communément admises pour n et introduit des modèles de bi-, tri- et quadri-gramme. L'estimation de ces modèles utilise le lissage des probabilité de Good-Turing, selon les recommandations classiques de lissage, *discounting*, et *back-off* (Chen & Goodman, 1999). Nous cherchons tout d'abord à comparer les n-gramme à horizon fixe avec des n-gramme à horizon variable. Les n-gramme à horizon variable sont mis en œuvre lors du test, en appliquant l'équation 4. Notre objectif n'est pas de valider une technique de réduction de paramètres au moment de la construction du modèle, (Siu & Ostendorf, 2000)(Niesler & Woodland, 1994), mais plutôt de débrider un modèle de n-gramme à horizon fixe pour en faire un modèle de n-gramme à horizon variable. Notre manière de procéder introduit un coût de calcul supplémentaire, mais il reste acceptable car les longueurs des historiques sont faibles devant le nombre de mots à traiter.

Nous cherchons enfin à situer les modèles n/m-multigramme par rapport aux deux approches précédentes. L'intérêt du multigramme réside dans sa capacité à prédire une séquence de mots avec un seul paramètre. En moyenne on baisse le nombre de termes impliqués dans le calcul de la perplexité ; il s'agit alors d'une situation favorable. Cependant, le risque est de répartir une masse de probabilités sur plus de termes[6]. Pour que la compétition entre modèles reste équitable, nous avons choisi de travailler avec des modèles n/m-multigramme dont la taille maximale (en nombre de mots) est soumise à une contrainte.

4 Estimation des paramètres des modèles

Les expériences sont réalisées à partir de la suite de programme *HTK* (Woodland & Young, 1993). Cet ensemble de librairies et d'outils correspond à une chaîne complète permettant de construire et de tester un modèle de langage. La gestion des n-gramme à horizon variable n'est pas écrite dans la distribution standard de *HTK*. La modification du programme de test du modèle de langage a été nécessaire pour introduire le traitement proposé équation 4. La gestion des n/m-multigramme n'est pas non plus écrite. Les modifications à faire sont d'une part dans le programme d'apprentissage, afin de

[6]Un n-gramme classique estime, pour chaque historique, une densité de probabilité dont la complexité spatiale est celle du vocabulaire. Les multigramme avec des têtes de longueur au plus m ont une complexité spatiale bornée par $|\mathcal{V}|^m$, les probabilités tendent vers 0.

parcourir systématiquement toutes les unités de multigramme possibles Il est également nécessaire de modifier, comme pour les n-gramme à horizon variable, le programme de test, pour pouvoir parcourir toutes les têtes et tous leurs historiques possibles.

Pour les n-gramme à horizon fixe, la perplexité du modèle de langage est déterminée directement grâce à l'équation 6. Si le mot de tête du n-gramme n'est pas dans le vocabulaire sélectionné, il est alors compté comme mot hors vocabulaire.

$$\begin{aligned} PP &= 2^{H^*} \quad \text{avec} \\ H^* &= -\frac{1}{m}\log_2\left(P(w_1, w_2, \ldots, w_m)\right) \end{aligned} \tag{6}$$

Pour les n-gramme à horizon variable, la situation est différente : pour chaque mot plusieurs choix sont possibles (le choix se fait entre un 2-gramme, un 3-gramme, ..., ou un n-gramme). Il suffit de choisir *le meilleur* k-gramme parmi les $n-1$ possibles (choix parmi toutes les longueurs d'historique autorisées). Cet algorithme est appliqué phrase par phrase (hypothèse d'indépendance des phrases entre elles). En fin de traitement d'une phrase on connaît la perplexité évaluée sur cette phrase, le nombre de mots prédits ainsi que le nombre de mots hors vocabulaire.

Pour les n/m-multigramme, la situation est encore différente. Maintenant plusieurs têtes sont disponibles, et pour chacune d'elles, plusieurs choix sont possibles. Nous avons volontairement limité la taille maximale du n/m-multigramme à un nombre fixe de mots : avec des multigramme de taille au plus 2, nous pourrions former 4 bi-gramme : $P(w_i|w_{i-1})$, $P(w_i|[w_{i-2}\,w_{i-1}])$, $P([w_i\,w_{i+1}]|w_{i-1})$, $P([w_i\,w_{i+1}]|[w_{i-2}\,w_{i-1}])$. En limitant le nombre maximum de mots dans le n/m-multigramme, nous pouvons choisir le modèle de langage avec lequel nous entrons en concurrence. Par exemple, avec des multigramme de taille au plus 2, et une somme à 3, nous n'avons plus que 3 choix possibles : $P(w_i|w_{i-1})$, $P(w_i|[w_{i-2}\,w_{i-1}])$, et $P([w_i\,w_{i+1}]|w_{i-1})$. Dans le programme de test, afin de sélectionner la meilleure découpe de la phrase selon le $\max$ de l'équation 5, nous avons mis en place une recherche du meilleur chemin dans un graphe[7] orienté et valué selon l'algorithme de Dijkstra.

5 Méthodologie expérimentale

Les expériences sont réalisées sur un corpus de texte du français : tous les articles parus pendant l'année 1997 dans le journal "Le Monde" (ressource ELRA). Ce corpus est découpé en phrases par un logiciel d'analyse syntaxique (logiciel Cordial de Synapse). Les phrases sont uniformisées par une réécriture systématique en majuscules et la suppression de toute ponctuation. Le corpus ainsi obtenu contient 1 131 135 phrases pour un vocabulaire de 219 034 mots. Il s'agit de la taille exacte du vocabulaire (mots variants en genre et en nombre, ainsi que les verbes rencontrés sous une forme conjuguée), le nombre d'occurrence des mots est de 23 999 626. L'apprentissage se fait sur 70% du corpus, le test est réalisé sur les 30% restant. La répartition des phrases a été réalisée de manière aléatoire à partir du corpus d'origine.

Pour faire baisser le nombre de paramètres d'un modèle de taille n, on fait évoluer la valeur de *cut-off* sur des n-gramme à horizon fixe. On conserve une valeur de *cut-off* à 1 sur les paramètres d'ordre inférieur (horizon de longueur inférieure à n). Ainsi, pour faire baisser le nombre de paramètres d'un modèle de tri-gramme, on va augmenter la valeur du *cut-off* sur les probabilités conditionnelles des tri-gramme, et laisser constante la valeur de *cut-off* pour les probabilités de bi-gramme et d'uni-gramme. Pour les $2/2$-multigramme, on peut faire baisser le nombre de paramètres en limitant le nombre de multigramme autorisés dans le modèle de langage[8]. On peut ainsi fixer le nombre de paramètres du modèle n/m-

[7] L'algorithme de Viterbi permet de rechercher la meilleure solution a priori, nous lui préférons l'algorithme de Dijkstra qui permet d'obtenir la meilleure solution a posteriori.

[8] avec un nombre de multigramme à 0, on obtient un modèle de bi-gramme ; cela est visible sur la figure 1 en prolongeant la courbe de perplexité des n/m-multigramme.

multigramme de façon précise grâce à un algorithme de dichotomie qui sélectionne le bon nombre de multigramme à prendre en compte dans la suite.

Si au moins un des mots de l'historique n'est pas présent dans le vocabulaire alors le modèle de langage déclare ne pas pouvoir prédire le n-gramme. Le mot de tête du n-gramme est alors déclaré non prédit, et la perplexité n'évolue pas. Dans la situation où tous les mots sont présents dans le vocabulaire, mais où la probabilité du n-gramme n'a pas été apprise par le modèle de langage, le système de *back-off* déjà présenté s'applique. Dans le cas des n-gramme à horizon variable, la probabilité est évaluée de manière identique, le mot en tête du n-gramme est déclaré non prédit si au moins un mot de son horizon est hors vocabulaire. Si tous les mots de l'horizon sont dans le vocabulaire, le choix de la meilleure probabilité est réalisé selon l'équation 4. Pour les multigramme, l'algorithme de Dijkstra permet de déterminer la meilleure solution au sens de l'équation 5, parmi toutes les solutions possibles.

6 Résultats et commentaires

La figure 1 présente l'évolution de la perplexité en fonction du nombre de paramètres des différents modèles pour différentes tailles de vocabulaire. Le modèle de bi-gramme à horizon variable est exactement le modèle de bi-gramme à horizon fixe, les courbes de perplexité sont donc confondues. On peut observer que le modèle de n/m-multigramme tend à avoir un comportement de bi-gramme de mots quand le nombre de multigramme autorisés diminue.

La perplexité d'un modèle de langage augmente quand le nombre de paramètres utilisé baisse. Cela est parfaitement normal, car le pouvoir de prédiction d'un mot de la langue est moins important avec un nombre de paramètres inférieur. Un modèle de n-gramme semble avoir une perplexité plus importante qu'un modèle de $n+1$-gramme. Cependant (Bonafonte & Mariño, 1996) rapporte que la perplexité des n-gramme augmente à partir de $n = 5$. Un modèle de tri-gramme avec un seuil de *cut-off* à 2 a une perplexité et un nombre de paramètres plus faible qu'un modèle de bi-gramme avec un seuil à 0 ; le modèle de tri-gramme est donc préférable dans ce cas. Selon (Rosenfeld, 2000), l'intérêt comparé d'un modèle de langage apparaît lorsque la mesure de perplexité baisse de plus de 10%. Le modèle de tri-gramme est donc notablement plus intéressant que le modèle de bi-gramme. Tout comme le modèle de quadri-gramme est plus intéressant que le modèle de tri-gramme.

Un modèle de n-gramme à horizon variable, comparativement au n-gramme concurrent, à horizon fixe, obtient une perplexité[9] plus faible. Cette baisse de la perplexité est due pour partie au calcul de la probabilité maximum ; en effet, par construction, on obtient une probabilité au moins supérieure à celle déterminée par le modèle à horizon fixe. Le gain obtenu par des n-gramme à horizon variable provient également de l'utilisation du coefficient de *back-off* par le modèle de n-gramme. En effet, le modèle de n-gramme utilise un coefficient de *back-off* pour obtenir une probabilité de n-gramme à horizon fixe à partir de la probabilité du $n-1$-gramme qui lui correspond si le n-gramme n'est pas trouvé. Le modèle de n-gramme à horizon variable permet de mélanger les probabilités des différents $(n-k)$-gramme avec $m \in [1, n-2]$, et ce sans pénaliser des $(n-k)$-gramme d'ordre inférieur.

Les n/m-multigramme se montrent moins performants que le modèle de n-gramme de même ordre (c'est à dire a nombre de mot considérés constants). En effet, l'équation 5 semble indiquer que le choix de la meilleure probabilité se fait entre un bi-gramme de mots, un tri-gramme de mots, et un bi-gramme ayant 2 mots en tête (dans le cas où la taille maximum d'un multigramme est de 2 mots, et la somme des mots du bi-gramme est d'au plus 3). Le choix ne peut donc par construction qu'être au moins aussi bon qu'un tri-gramme de mots. Cependant, nous pouvons constater que pour obtenir un nombre de paramètres équivalent afin de comparer les différents modèles, il faut interdire un nombre conséquent de multigramme parmi ceux disponibles. Nous devons alors chercher à améliorer ce modèle n/m-multigramme.

[9] Bien sûr, il ne s'agit pas exactement d'une mesure de perplexité qui devrait être calculée à partir d'une distribution de probabilité.

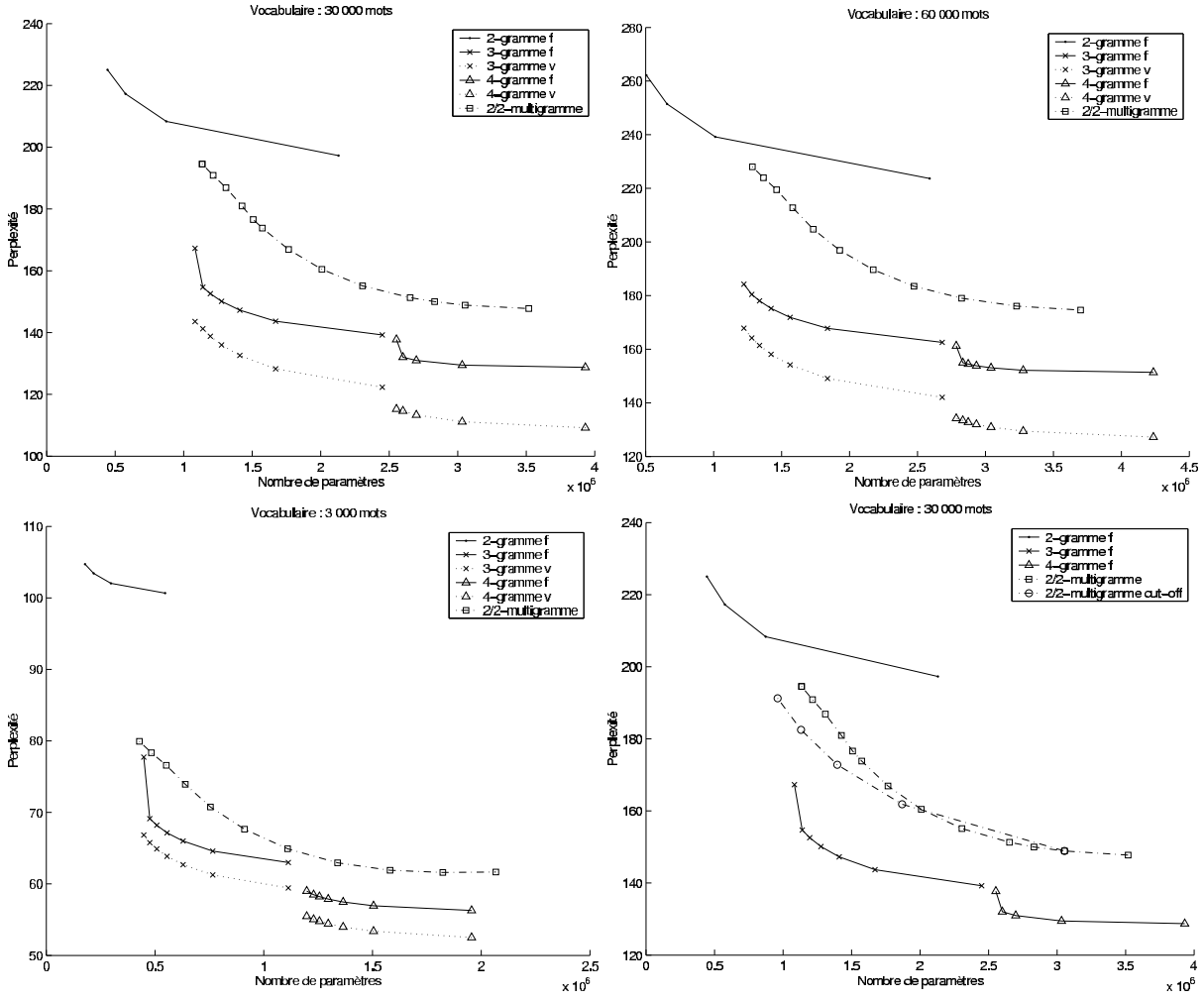


Figure 1: Comparaison de l'influence du nombre de paramètres sur la perplexité des modèles de n-gramme à horizon fixe (n-gramme-f) ou variable(n-gramme-v) pour $n \in [2, 4]$, et du $2/2$-multigramme pour différentes tailles de vocabulaire, et influence sur la perplexité de la méthode de *cut-off* pour réduire les paramètres du modèle de $2/2$-multigramme avec un vocabulaire de 30 000 mots.

Pour améliorer la situation, on peut tout d'abord chercher à n'inclure dans les multigramme autorisés que ceux qui apportent un gain vis à vis de l'équation5. Nous avons constaté par des expériences que ces multigramme n'améliorent pas significativement la perplexité (nous n'avons pas la place pour rapporter ces expériences). Cela semble indiquer que les multigramme qui apportent le plus gros gain en terme de perplexité sont déjà inclus dans la liste des plus fréquents. Une expérience similaire consiste à définir la liste des multigramme en changeant le seuil de *cut-off*. En effet, on peut observer une baisse significative du nombre de paramètres, qui s'accompagne d'une augmentation de la perplexité (environ 10 points) quand on passe d'un bi-gramme de mots avec un *cut-off* à 0 (respectivement 1) à un bi-gramme de mots avec un *cut-off* à 1 (respectivement 2). La figure 1 montre l'évolution de la perplexité en conservant les multigramme les plus fréquents (100 000 multigramme pour un vocabulaire de 30 000 mots). On peut constater une baisse du nombre de paramètres sans hausse de la perplexité ; cette solution semble donc convenir. Enfin, étant donnée la baisse significative de la perplexité observée avec peu de multigramme entre un bi-gramme de mots avec un *cut-off* à 1, et un $2/2$-multigramme avec le même *cut-off*. On peut souhaiter généraliser l'usage des multigramme aux n-gramme. Cependant la complexité risque d'augmenter de manière exponentielle avec n.

7 Conclusion

Cet article a présenté des résultats concernant des modèles de langage statistiques de type n-gramme à horizon fixe ou variable et des n/m-multigramme. À taux de mots hors-vocabulaire fixe, le comportement des n-gramme classiques fait baisser la perplexité pour des valeurs de n de 3 à 4, mais au prix d'une baisse du nombre de mots prédits (environ 7 millions pour un modèle de bi-gramme, 6.8 millions pour un tri-gramme, et un peu plus de 6.6 millions pour un quadri-gramme). Plus on reconnaît des mots, plus la probabilité conjointe va être faible, on peut donc trouver discutable de comparer entre eux des modèles de n-gramme qui ne se trouvent pas tout à fait sur le même pied d'égalité. Ce problème ne se pose pas pour les n-gramme à horizon variable, ou les n/m-multigramme, car le nombre de mots prédits est à chaque fois celui du modèle de bi-gramme. Les résultats de perplexité obtenus avec des vocabulaires de taille plus importante nous montrent à la fois une augmentation de la perplexité, et une augmentation du nombre de paramètres. Cette augmentation est due encore une fois à une augmentation du nombre de mots prédits (pour un modèle de bi-gramme, nous avons près de 4.9 millions de mots prédits pour un vocabulaire de 3 000 mots, 7 millions pour 30 000 mots, et 7.3 millions pour 60 000 mots). Le taux de mots hors vocabulaire sur le corpus de test baisse de 19.38% pour 3 000 mots à 1.65% pour 60 000 mots. Nous avons montré que le modèle de multigramme le plus simple, un 2/2-multigramme (c'est-à-dire un bi-gramme de séquences comprenant au plus deux mots) se comporte comme un modèle situé entre un bi-gramme et un tri-gramme classique. Notre objectif consiste à pousser un peu plus loin ces modèles en augmentant notamment l'ordre et en réglant le nombre de paramètres par des techniques de cut-off.

Références

BIMBOT, F., PIERACCINI, R., LEVIN, E., & ATAL, B. 1995. Variable-Length Sequence Modeling: Multigrams. *IEEE Signal Processing Letters*, **2**(6), 111–113.

BONAFONTE, A., & MARIÑO, J. 1996. Language Modeling Using X-grams. *Pages 394–397 of: Proceedings of the International Conference on Spoken Language Processing.*

CHEN, S.F., & GOODMAN, J. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, **13**(4), 359–394.

DELIGNE, S., & BIMBOT, F. 1995. Language modeling by variable length sequences: theoretical formulation and evaluation of multigrams. *In: IEEE International Conference on Acoustics and Speech Signal Processing.*

DELIGNE, S., & SAGISAKA, Y. 2000. Statistical language modeling with a class-based n-multigram model. *Computer Speech and Language*, **14**, 261–279.

KATZ, S.M. 1987. Estimation of Probailities from Sparse Data for the Language Model Component of a Speech Recognizer. *IEEE transactions on Acoustics, Speech and Signal Processing*, **35**, 400–401.

NIESLER, T.R., & WOODLAND, P.C. 1994. Variabl-length category n-gram language models. *Computer Speech and Language*, **13**, 99–124.

ROSENFELD, R. 2000. Two decades of statistical language modeling: where do we go from here? *Proceedings of the IEEE*, **88**(8), 1270–1278.

SHIELDS, P.C. 1998. The Interactions Between Ergodic Theory and Information Theory. *IEEE Transactions on Information Theory*, **44**, 2079–2093.

SIU, M., & OSTENDORF, M. 2000. Variable n-grams and extensions for conversational speech language modeling. *IEEE transactions on Speech and Audio Processing*, **8**(1), 63–75.

WOODLAND, P.C., & YOUNG, S.J. 1993. The HTK Continuous Speech Recogniser. *Pages 2207–2219 of: Proceedings of the Eurospeech conference.*

ZITOUNI, I. 2002. A Hierarchical Language Model Based on Variable-Length Class Sequences: The MC_n^{ν} Approach. *IEEE Transactions on Speech and Audio Processing*, **10**(3), 193–198.

Y a-t-il une taille optimale des règles de succession intervenant dans l'étiquetage grammatical ?

Fathi Debili (1), Emna Souissi (2)

(1) CNRS – ICAR – ENS LSH
15, Parvis René Descartes
69342 Lyon Cedex 07
France
fathi.debili@wanadoo.fr

(2) ISG – Université de Sousse
BP 763 Sousse 4000 - Tunisie
emna.souissi@isgs.rnu.tn

Mots-clés : Etiquetage grammatical, règle de succession, taille des règles, chaînage de règles, règle attestée, règle simulée, discriminance, couverture, évaluation en usage *vs* évaluation en définition d'un ensemble de règles.

Keywords: Part-of-speech tagging, tag sequences, rule length, rule composition, attested rule, simulated rule, evaluation of generation *vs* evaluation of analysis.

Résumé

La quasi-totalité des étiqueteurs grammaticaux mettent en œuvre des règles qui portent sur les successions ou collocations permises de deux ou trois catégories grammaticales. Leurs performances s'établissent à hauteur de 96% de mots correctement étiquetés, et à moins de 57% de phrases correctement étiquetées. Ces règles binaires et ternaires ne représentent qu'une fraction du total des règles de succession que l'on peut extraire à partir des phrases d'un corpus d'apprentissage, alors même que la majeure partie des phrases (plus de 98% d'entre elles) ont une taille supérieure à 3 mots. Cela signifie que la plupart des phrases sont analysées au moyen de règles reconstituées ou simulées à partir de règles plus courtes, ternaires en l'occurrence dans le meilleur des cas. Nous montrons que ces règles simulées sont majoritairement agrammaticales, et que l'avantage inférentiel qu'apporte le chaînage de règles courtes pour parer au manque d'apprentissage, plus marqué pour les règles plus longues, est largement neutralisé par la permissivité de ce processus dont toutes sortes de poids, scores ou probabilités ne réussissent pas à en hiérarchiser la production afin d'y distinguer le grammatical de l'agrammatical. Force est donc de reconsidérer les règles de taille supérieure à 3, lesquelles, il y a une trentaine d'années, avaient été d'emblée écartées pour des raisons essentiellement liées à la puissance des machines d'alors, et à l'insuffisance des corpus d'apprentissage. Mais si l'on admet qu'il faille désormais étendre la taille des règles de succession, la question se pose de savoir jusqu'à quelle limite, et pour quel bénéfice. Car l'on ne saurait non plus plaider pour une portée des règles aussi longue que les plus longues phrases auxquelles elles sont susceptibles d'être appliquées. Autrement dit, y a-t-il une taille optimale des règles qui soit suffisamment petite pour que leur apprentissage puisse converger, mais suffisamment longue pour que tout chaînage de telles règles

pour embrasser les phrases de taille supérieure soit grammatical. La conséquence heureuse étant que poids, scores et probabilités ne seraient plus invoqués que pour choisir entre successions d'étiquettes toutes également grammaticales, et non pour éliminer en outre les successions agrammaticales.
Cette taille semble exister. Nous montrons qu'au moyen d'algorithmes relativement simples l'on peut assez précisément la déterminer. Qu'elle se situe, compte tenu de nos corpus, aux alentours de 12 pour le français, de 10 pour l'arabe, et de 10 pour l'anglais. Qu'elle est donc en particulier inférieure à la taille moyenne des phrases, quelle que soit la langue considérée.

Abstract

Is there an optimal n for n-grams used in part-of-speech tagging?
Almost all part-of-speech taggers apply rules about permitted successions and collocations of two or three grammatical categories. Their performance amounts to 96 percent of correctly tagged words, and to less than 57 percent of correctly tagged sentences. These binary and ternary succession rules represent a small fraction of succession rules one can extract from sentences in a learning corpus, where most sentences (more than 98 percent of them) have a length of more than three words. In other words, most sentences are processed by rules that are reconstructed, or simulated, from shorter ones, here ternary at best. We show that most such simulated rules are agrammatical, and that, if some inferential benefit comes from the chaining of short rules to compensate inexistent learning, mainly in the case of long rules, this benefit is nullified by the permissive behaviour of this process, in which a variety of weights, scores or probability are ineffective in hierarchizing its production and yield a separation between grammatical and agrammatical rules. So we feel forced to look again at larger-than-ternary rules. However, if we admit a necessity of enlarging succession rules, we must ask the question "up to which limit, and for what profit". For we also decline to argue for rules as long as the longest sentences upon which they might apply. So the real question is, can we define an optimal size for rules, short enough for learning to converge, and long enough for any chaining of rules to deal with larger sentences to be grammatical? A positive result would be that weights, scores or probability would then be invoked only to decide between equally grammatical successions of tags, and no longer to eliminate agrammatical ones.
This optimal size apparently exists. We show that the use of rather simple algorithms leads to its determination. And its value, according to our corpora, is near 12 for French, 10 for Arabic and 10 for English. Therefore, it is less than the average length of sentences, for each of these three languages.

1 Introduction

Cet article rend compte d'une étude critique des règles les plus couramment mises en œuvre dans les étiqueteurs grammaticaux. Il souligne de façon quantitative le caractère infondé de l'emploi généralisé, du moins dans la plupart des modèles probabilistes, des seules règles de succession (ou de précédence) d'ordre deux ou trois que renferment les expressions : *P(étiquette de rang i | étiquette de rang i-1) ; P(étiquette de rang i | étiquette de rang i-2, étiquette de rang i-1).*

La sanction, s'il en est, de cet état de fait est bien connue : un niveau de performance – 96% de mots correctement étiquetés – rapidement atteint (par exemple Debili, 1977 ; DeRose, 1988), mais que l'on peine à dépasser de façon substantielle et reproductible, en dépit d'efforts qui ne se sont point relâchés depuis plus de trente ans (par exemple Andreewsky et Fluhr, 1973 ; Mérialdo, 1994 ; Leech et al., 1994 ; Adda et al., 1999 ; Valli et Véronis, 1999 ; Nasr et Volanschi, 2004). Ce niveau de performance est relativement faible. Il signifie que si la taille moyenne de la phrase est de 20 mots, alors en moyenne 4 phrases sur 5 sont mal étiquetées. Dans la pratique, la situation est sans doute meilleure, mais demeurant à des niveaux qui restent très faibles, les performances au niveau de la phrase sont rarement affichées. 57% de phrases correctement étiquetées est semble-t-il le meilleur résultat publié que l'on ait pu atteindre (Toutanova et al., 2003). Erigés en barrière depuis une trentaine d'années, ces niveaux de performances, ajoutés aux difficultés réelles que pose leur évaluation comparative, ont pu amener jusqu'à douter du statut de l'étiquetage tel que préconisé (Fairon et Sennelart, 1999).

Nous voulons ici revenir sur une hypothèse non toujours formulée liée à la portée des règles mises en œuvre dans les étiqueteurs grammaticaux. Leur extrême localité – examen de contextes très proches portant sur les deux positions qui précèdent, entourent ou suivent l'ambiguïté étudiée – est souvent pointée pour expliquer les échecs constatés, mais elle est également souvent assumée au nom de contraintes techniques liées autant à la puissance des machines qu'à la taille des corpus d'apprentissage, et de fait en partie justifiée sur le plan expérimental, puisque aussi bien l'on a pu en effet enregistrer des résultats meilleurs avec des règles de portées paradoxalement plus courtes.

Il ne s'agit pas évidemment de rejeter ces règles qui restent utiles. Il s'agit de s'attaquer au problème que leur chaînage soulève dès lors que l'on essaie de les appliquer à des phrases dont la taille est supérieure à 2 ou à 3 mots, ce qui arrive dans plus de 98% des cas. Le schéma suivant où les m_i sont les mots de la phrase, et les t_{ij}, leurs diverses étiquettes grammaticales potentielles respectives, rappelle ce qui se passe :

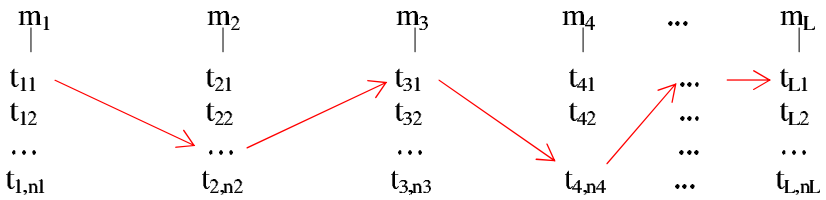


Figure 1

L'étiquetage consiste à reconnaître parmi les N^L chemins combinatoirement possibles (N étant le nombre moyen d'étiquettes pour un mot, et L le nombre de mots de la phrase), le chemin qui permet d'attribuer à chacun des mots l'étiquette grammaticale qui lui sied dans la phrase. La situation est plus complexe en fait car la segmentation en mots n'est pas toujours unique. Mais nous ne nous en préoccuperons pas ici, car cela n'a pas d'incidence sur ce que nous essayons de montrer. Nous ne nous préoccuperons pas non plus de la nature des étiquettes que nous recevrons comme telles.

Appliquer des règles *binaires* pour construire ces chemins, c'est joindre bout à bout des successions *attestées* (c'est-à-dire qui ont été relevées dans un corpus d'apprentissage) de 2 étiquettes, de façon telle que la seconde étiquette de la première règle soit identique à la première de la seconde règle. Ainsi, si *a*, *b*, *c*, …, *z*, sont des étiquettes grammaticales, et (*a b*), (*b c*) des successions attestées, alors (*a b c*) est une succession d'ordre 3 potentiellement valide ou, dirons-nous, potentiellement *attestable*.

Appliquer des règles *ternaires*, c'est chaîner de la même façon des successions attestées de trois étiquettes de façon telle que les deux dernières du premier triplet soient identiques aux deux premières du second triplet. Ainsi, si (*a b c*) et (*b c d*) sont des successions ternaires attestées, alors on peut former la succession quaternaire (*a b c d*). Et de proche en proche (*a b c d e*) si (*c d e*) est une succession attestée, et ainsi de suite. Nous remarquons qu'il y a là production de successions qui à leur tour peuvent être assimilées à des règles d'ordre supérieur. La différence est que les premières sont attestées, alors que les secondes ne le sont pas. Pour les distinguer, nous les appellerons *simulées* dans la suite de l'exposé.

C'est ainsi qu'opèrent basiquement la plupart des étiqueteurs qui essaient d'établir ou de reconnaître des continuités syntaxiques. Le problème réside dans le fait que les successions ainsi obtenues et que l'on peut en effet assimiler à des règles de succession de taille aussi longue que les phrases auxquelles elles s'appliquent, ne se révèlent pas toujours, loin s'en faut, attestables, bien qu'établies à partir de règles binaires ou ternaires attestées. L'on s'aperçoit que ces successions ou règles simulées sont au contraire souvent agrammaticales. Cela est bien connu. Mais à notre connaissance aucune étude à caractère quantitatif n'a été menée pour mesurer la proportion du potentiellement attestable ou grammatical par opposition à ce qui demeurera non attestable ou agrammatical.

La question qui surgit est alors : y a-t-il corrélation entre ces proportions grammaticales *vs* agrammaticales de règles simulées et la taille des règles attestées leur ayant donné naissance ? Sans doute oui, mais là aussi nous n'en connaissons pas la nature. L'on imagine toutefois que la proportion des règles simulées agrammaticales devrait aller en diminuant à mesure que la taille des règles

attestées génératrices irait croissant. Si tel est le cas, quelle est la valeur minimale de cette taille, taille à partir de laquelle l'on n'obtiendrait plus que des successions simulées attestables ? Ne pourrait-on pas élaborer un protocole expérimental pour déterminer de meilleure façon, empirique et non plus apriorique, la taille optimale des règles de succession devant intervenir dans l'étiquetage grammatical ?

C'est à ces questions que nous allons essayer de répondre à partir d'observations et de calculs effectués sur trois corpus étiquetés : • français (MULTITAG du CNRS-LIMSI, un ensemble de textes du *Monde* d'environ 754 000 mots regroupés en 31 points et utilisant 337 étiquettes grammaticales, Paroubek et Rajman, 2000) ; • arabe (un ensemble de 53 textes du *Monde Diplomatique* d'environ 92 000 mots manuellement voyellés et étiquetés au CNRS-ICAR, utilisant 558 étiquettes) ; • et anglais (SUZANNE Corpus, 64 textes d'environ 149 000 mots, utilisant 308 étiquettes).

2 Notations

Ayant à considérer pour les besoins de nos expérimentations différentes tailles de corpus ou de règles :

- $T_{1..x}$: désignera le corpus constitué de l'ensemble de textes $\{T_1, T_2, \ldots, T_x\}$
- $RA_p(T_{1..x})$ ou plus simplement $RA_p(T)$: désignera l'ensemble des règles d'ordre p attestées ou apprises à partir du corpus $T_{1..x}$, ou plus simplement T.
- $RS_q[RA_p(T)]$: désignera l'ensemble des règles simulées d'ordre q engendrées à partir des règles attestées d'ordre p ($p < q$) du corpus T.

Rappelons qu'une règle simulée d'ordre q est obtenue en chaînant $(q - p + 1)$ règles attestées d'ordre p. Exemple avec $q = 7$ et $p = 5$:

		1	2	3	4	5	6	7
Règle attestée	1	*a*	*b*	*c*	*d*	*e*		
Règle attestée	2		*b*	*c*	*d*	*e*	*f*	
Règle attestée	3			*c*	*d*	*e*	*f*	*g*
Règle simulée		*a*	*b*	*c*	*d*	*e*	*f*	*g*

Figure 2 : Règle simulée d'ordre 7 engendrée au moyen de règles attestées d'ordre 5.

3 Règles de succession attestées

Les règles attestées sont les successions d'étiquettes grammaticales de diverses longueurs que l'on peut extraire des phrases prises une à une dans des textes préalablement étiquetés. Une phrase de trois mots respectivement étiquetés (*a b c*) donnera lieu aux six successions suivantes : • unaires : (*a*), (*b*), (*c*) ; • binaires : (*a b*), (*b c*) ; • et ternaire : (*a b c*). Plus généralement, une phrase donnée de longueur L donnera naissance à L(L + 1)/2 successions différentes, de taille allant de une à L étiquettes.

La figure 3 donne les effectifs associés aux trois corpus français, arabe et anglais. Dans les cas d'espèce, si l'on s'intéresse aux règles binaires et ternaires qui sont les seules mises en œuvre dans la quasi totalité des étiqueteurs probabilistes, ces histogrammes révèlent qu'en fait celles-ci cumulées ne représentent au plus que 3,57% du total des règles de succession potentielles que l'on peut extraire d'un texte dûment étiqueté, cas de l'anglais. Et moins d'un pour cent (0,75%) lorsque la taille du corpus est plus importante, cas du français, l'arabe étant à 3,10%. Plus de 96%, voire 99%, des règles potentiellement utiles pour une analyse plus discriminante de ces mêmes corpus ou d'autres ne sont pas retenues, tandis que plus de 98% des phrases ont d'une façon générale plus de trois mots. Ces proportions soulignent que dans une perspective d'analyse, plus de 98% des phrases sont étiquetées au moyen de moins de 4%, voire moins d'un pour cent, du total des règles que l'on peut extraire des corpus. Autrement dit, 98% des phrases ou davantage sont étiquetées aux moyen de règles simulées.

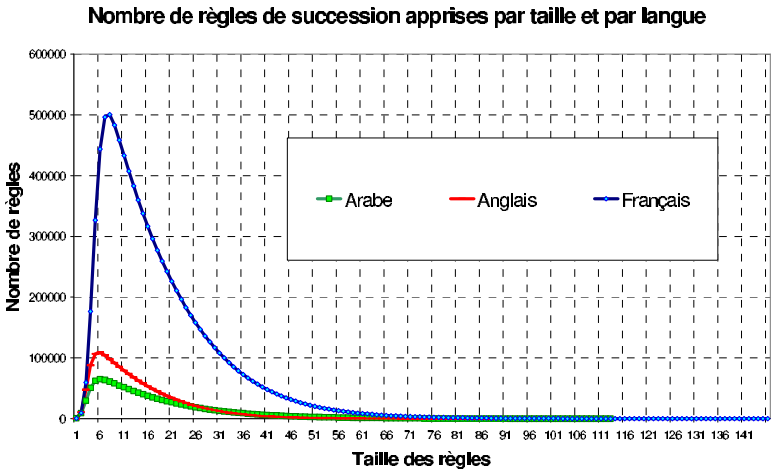


Figure 3 : Histogramme des successions attestées selon leur taille

Ce constat ne signifie pas bien entendu qu'il faille opter d'emblée pour l'apprentissage de toutes les règles de succession, y compris de celles qui seraient aussi longues que les plus longs tronçons ou phrases dont elles sont issues. Quelle valeur aurait une règle hapax qui aurait toutes les chances de ne pouvoir jamais être appliquée qu'au fragment ou à la phrase d'où elle provient ? L'idée n'est donc pas de sauvegarder toutes les règles quelle que soit leur taille, d'autant qu'il s'avère que celles-ci sont toutes précisément systématiquement hapax dès lors que leur taille dépasse un certain seuil, ici respectivement pour les trois corpus : 16 pour l'anglais, 19 pour l'arabe, et 60 pour le français.

Cependant, dans la perspective qui est la nôtre d'argumenter en faveur d'une extension de la taille des règles apprises, ces valeurs peuvent être considérées dès maintenant comme des limites qu'il ne sera pas utile de dépasser. Nous verrons, moyennant d'autres observations, qu'il ne sera pas non plus utile d'aller aussi loin.

Il nous faut auparavant argumenter davantage et en particulier tenter d'évaluer, – en amont des performances d'étiquetage déjà signalées, et que nous qualifierons d'extrinsèques (ou exogènes), dans la mesure où elles renseignent sur la couverture et la discriminance *en usage* des règles binaires et ternaires, – de ces mêmes propriétés mais *en définition*. C'est-à-dire de la propension combinatoire d'un ensemble de règles à n'engendrer intrinsèquement (ou de façon endogène) par chaînage, que des règles simulées attestables, ou en tout cas à minimiser la proportion du généré non attestable. A l'inverse de l'évaluation *en usage*, qui sans doute au final seule compte, l'évaluation *en définition* d'un ensemble de règles a une valeur prédictive. En introduisant des critères et des indices permettant de comparer des ensembles de règles différents sous l'angle de leur fonctionnement interne et en dehors de toute confrontation aux textes à étiqueter, on se dote de moyens permettant d'anticiper précisément de la qualité de l'étiquetage. L'on imagine en effet que de deux ensembles de règles engendrant les mêmes règles simulées attestables par ailleurs, celui qui produirait proportionnellement le moins de règles agrammaticales concomitamment, devrait conduire incontestablement à de meilleures performances *en usage*, c'est-à-dire de étiquetage proprement dit.

4 Règles de succession simulées

L'argument souvent avancé en faveur des règles binaires ou ternaires est lié à leur capacité inférentielle ou « générative » et à la rapidité de leur apprentissage. Les règles de plus longue taille sont par opposition moins productives, et l'on met plus de temps à les acquérir, temps signifiant ici volume des corpus nécessaires à leur apprentissage.

Dans la perspective d'une extension de la taille des règles apprises, ces avantages subsistent, puisqu'il ne s'agit nullement de se départir des règles binaires et ternaires, mais seulement de leur adjoindre les règles de plus longue taille. L'on peut donc s'interroger sur l'utilité de cette extension, d'autant que ces règles de plus longue taille peuvent être totalement simulées à partir des règles binaires ou ternaires, et qu'en outre, cette simulation produira même potentiellement plus encore de règles attestables de longue taille que nous ne pourrions en extraire directement à partir des mêmes corpus d'apprentissage. La figure suivante illustre ces imbrications.

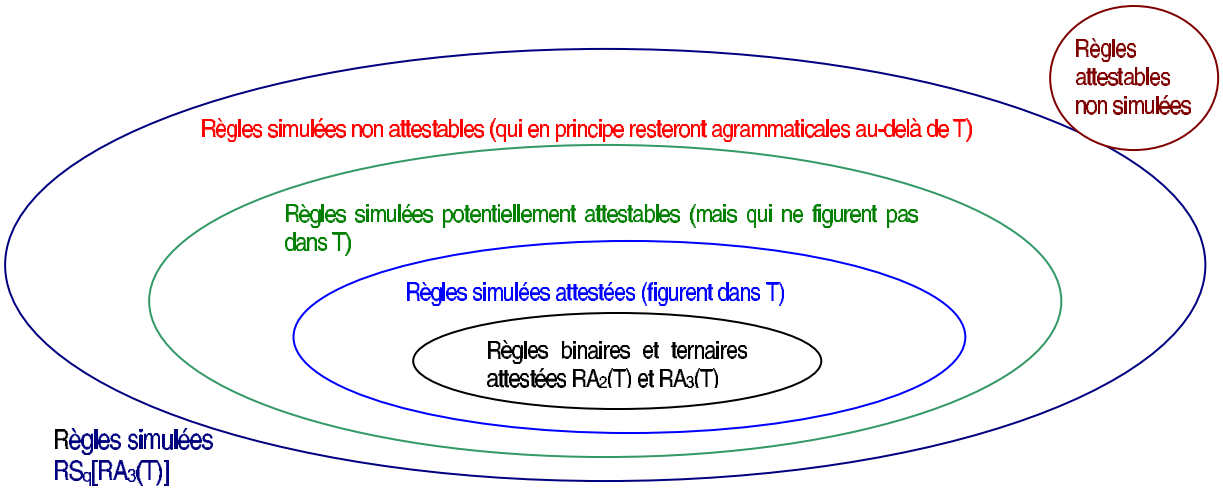


Figure 4 : Imbrication des ensembles de règles simulées $RS_q[RA_3(T)]$ avec $3<q\leq L$

Or, ces capacités productives des règles binaires ou ternaires n'ont à notre connaissance jamais été évaluées autrement qu'au travers de l'utilisation qui en est faite, c'est-à-dire des performances des outils d'étiquetage mettant en œuvre ces règles. Une évaluation en usage donc, qui au demeurant reste à des niveaux peu satisfaisants comme nous avons pu le voir, et non en définition. En particulier, nous n'avons aucune connaissance des proportions relatives des diverses règles simulées : – attestées, – potentiellement attestables, – et non attestables, rapportées au total des règles simulées. Ni *a fortiori* de l'évolution des ces mêmes proportions en fonction de la taille des règles ou des corpus d'apprentissage. Ces proportions et leurs évolutions respectives semblent pourtant être de nature à nous renseigner de façon apriorique sur la couverture et la discriminance d'un ensemble de règles par opposition à un autre. Idéalement, à couverture ou à capacité d'engendrement équivalente, celui des deux ensembles qui produira en proportion le plus de règles attestables sera considéré plus prometteur, car, surgénérant moins, il est plus discriminant. Dans la réalité, concilier couverture et discriminance s'avère contradictoire. C'est pourquoi la comparaison reste difficile, et qu'il nous faudra rechercher non le meilleur, mais un optimum qui ne correspondra qu'à un meilleur local.

Problème : comment mesurer ces proportions et leurs évolutions ?

5 Discriminance, couverture, évaluation d'un système de règles

Nous proposons d'appeler *discriminance en génération* de rang q d'un ensemble de règles attestées d'ordre p, $q>p$, le rapport :

$$\delta_{q,p} = \text{Card}\ \{\ RS_q[RA_p(T)] \cap RA_q(T)\ \} \ /\ \text{Card}\ \{\ RS_q[RA_p(T)]\ \}$$

Ce qui simplement correspond à la proportion des règles simulées d'ordre q, et attestées par le corpus T, rapportée au total des règles simulées d'ordre q, les règles simulées étant engendrées à partir des règles d'ordres p issues du même corpus T.

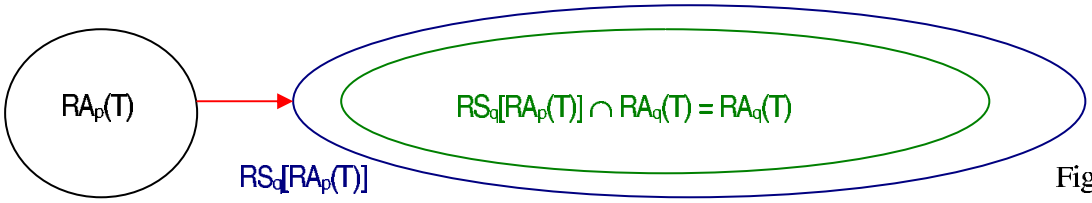


Figure 5a

Nous dirons que cette discriminance est *intrinsèque*, et nous désignerons par :

$$\Delta_{q,p} = \text{Card}\ \{\ RS_q[RA_p(T_a)] \cap RA_q(T_{1...x})\ \} \ /\ \text{Card}\ \{\ RS_q[RA_p(T_a)]\ \} \quad \text{avec } T_a \subset T_{1...x}$$

la discriminance *extrinsèque*, laquelle correspond à la proportion des règles simulées d'ordre q, et attestées par le corpus $T_{1...x}$, rapportée au total des règles simulées d'ordre q, les règles simulées étant engendrées à partir des règles d'ordres p issues du corpus T_a, T_a pouvant être inclus, ou éventuellement, partiellement inclus, ou non inclus dans $T_{1...x}$.

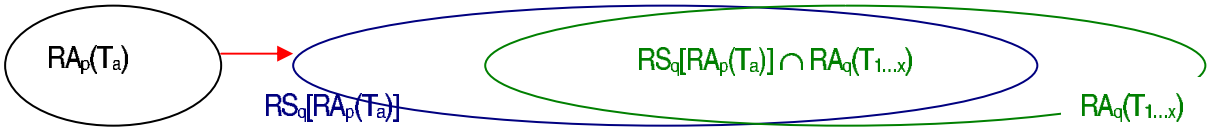


Figure 5b : Intersection des règles simulées $RS_q[RA_p(T_a)]$ avec les règles attestées $RA_q(T_{1...x})$

De même, nous proposons d'appeler *couverture en génération* de rang q d'un ensemble de règles attestées d'ordre p, $q>p$, le rapport :

$$\kappa_{q,p} = \text{Card} \{ RS_q[RA_p(T_a)] \cap RA_q(T_{1...x}) \} / \text{Card} \{ RA_q(T_{1...x}) \}$$

Cela correspond à la proportion des règles simulées d'ordre q, engendrées à partir des règles d'ordres p issues du corpus T_a, et attestées par le corpus $T_{1...x}$, rapportée au total des règles attestées d'ordre q de ce même corpus $T_{1...x}$. Nous remarquons que dans le cas particulier où T_a est identique à $T_{1...x}$, alors $\kappa_{q,p} = 1$.

D'une façon plus générale et par extension, nous appellerons *discriminance en génération* d'un ensemble de règles attestées de diverses longueurs issues d'un corpus d'apprentissage T_a, la proportion des règles engendrées à partir de ces règles, et attestées par le sur corpus $T_{1...x} \supset T_a$, rapportée au total des règles engendrées ; et *couverture en génération* d'un ensemble de règles attestées de diverses longueurs issues du sous corpus $T_a \subset T_{1...x}$, la proportion des règles engendrées à partir de ces règles, et attestées par $T_{1...x}$, rapportée au total des règles attestées de $T_{1...x}$.

Les protocoles expérimentaux permettant de calculer les rapports $\delta_{q,p}$, $\Delta_{q,p}$ et $\kappa_{q,p}$ sont relativement simples. Ils exigent néanmoins des temps de calcul relativement long, d'autant plus long que p est petit, et $(q-p)$ grand. C'est pourquoi nous ne pouvons donner ici, en particulier pour p=3, que les évolutions liées à de faibles écarts, mais que nous continuons d'élargir. Les tendances que nous voulons faire valoir à l'appui d'une extension des règles de succession d'une part, et pour la détermination d'une limite à cette extension d'autre part, sont cependant présentes.

En faisant varier les paramètres p, q, et x (taille du corpus), plusieurs familles de courbes peuvent être tracées. Ces courbes donnent les évolutions des discriminances (intrinsèque $\delta_{q,p}$ et extrinsèque $\Delta_{q,p}$), et de la couverture ($\kappa_{q,p}$) d'un ensemble de règles, soit en fonction de la taille des corpus (x), soit en fonction de la taille des règles génératrices (p), soit en fonction de la taille des règles engendrées (q).

L'observation de l'évolution en interne d'un système de règles au travers de ces diverses courbes et caractérisations recèle ce qui pourrait permettre de comparer entre eux des systèmes de règles différents, et dès lors de procéder à une certaine forme d'évaluation de ces systèmes de règles.

6 Discriminance en génération

Nous focaliserons notre attention ici sur l'évolution des indices de discriminance en génération d'ensembles de règles en fonction de p et de q. Et sur deux constats pressentis, mais désormais chiffrés : 1°) la décroissance, d'autant plus rapide que p est petit, de la discriminance, aussi bien intrinsèque qu'extrinsèque, à mesure que q croit ; 2°) et à l'inverse, la croissance rapide de cette même discriminance à mesure que p croit.

Les courbes ou tableaux $\delta_{q,p}$ et $\Delta_{q,p}$ (voir figure 6 pour le français) montrent qu'en passant simplement de q=4 à q=5 pour p=3, alors la proportion des règles attestées passe de 13,35% à 1,30% pour le

français, de 11,08% à 1,11% pour l'arabe, et de 10,73% à 0,8% pour l'anglais. Et que si l'on tient compte en outre des règles potentiellement attestables, alors l'on passe de 50,13% à 14,78% pour le français, de 50,49% à 18,71% pour l'arabe, et de 56,09% à 18,92% pour l'anglais.

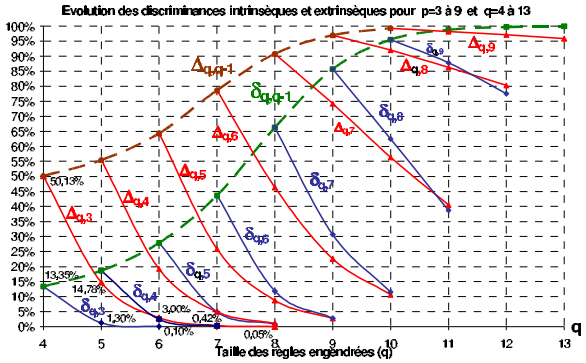


Figure 6 : La discriminance décroît à mesure que q croît ($\delta_{q,p}$ et $\Delta_{q,p}$ du français)

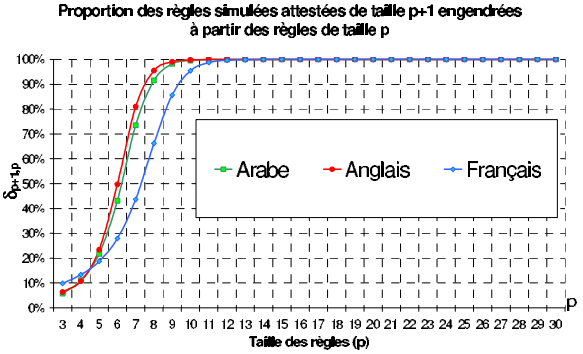


Figure 7 : La discriminance croît à mesure que p croît ($\delta_{p+1,p}$ de l'arabe, de l'anglais et du français)

Une décroissance qui signifie que, dans le meilleur des cas, si pour étiqueter une phrase de quatre mots l'on peut encore compter sur 50,13% de règles attestées ou potentiellement attestables, pour étiqueter une phrase de cinq mots l'on ne peut plus compter que sur 14,78%. Et beaucoup moins si l'on passe à l'étiquetage des phrases de six mots (3%) ou de sept (0,42%), et moins encore au-delà ($\Delta_{8,3}$ = 0,05% !).

Si maintenant nous observons l'évolution de la discriminance en fonction de p, (voir figure 7), nous constatons que ce rapport tend rapidement vers 1 à mesure que p augmente. Ce qui suggère que nous puissions limiter l'apprentissage aux seules règles de taille inférieure ou égale à une certaine valeur de p, valeur au-delà de laquelle les règles simulées se révèlent pour la plupart attestées. La figure 7 donne l'évolution de la proportion des règles simulées attestées rapportées au total des règles engendrées à partir des règles attestées d'ordre immédiatement inférieur, c'est-à-dire de $\delta_{p+1,p}$.

Nous observons que plus de 99% des règles simulées se révèlent attestées à partir de p = 10 pour l'anglais et l'arabe, et à partir de p=12 pour le français (corpus 5 à 7 fois plus important que celui de l'anglais ou de l'arabe). Moins d'un pour cent des règles simulées restent non attestées, sans que l'on puisse affirmer qu'elles sont toutes non attestables.

Nous retrouvons là, de façon empirique, ce que nous savions déjà : plus le contexte avant (ou après) est long, plus le choix sur ce qui suit (ou précède) est contraint. Ce que nous ne savions pas, pour ce qui est de l'étiquetage, c'est la taille de ce contexte, taille minimale à partir de laquelle nous constatons que le choix sur ce qui suit devient si contraint qu'il laisse peu de place à la production de successions qui ne soient pas très probablement attestables.

Nous n'avons pas fini d'exploiter les potentialités qu'offrent ces calculs. Nous pensons en particulier que les évolutions croisées de certaines courbes (discriminance et couverture en fonction de la taille des corpus), que nous n'avons pas pu mentionner ici, pourraient renseigner sur l'état de convergence de l'apprentissage, indépendamment de toute application des règles à l'étiquetage d'un texte nouveau, ainsi qu'il est traditionnellement fait.

7 Couverture en génération d'ensembles de règles de succession

Si la discriminance est à l'avantage des règles de succession longues, la couverture est à l'avantage des règles courtes. Les courbes de la figure 8 le rappellent, même si elles sont incomplètes. Associées aux courbes des figures 6 et 7, elles matérialisent ce que l'on peut intuitivement pressentir : à savoir que plus une règle ou un ensemble de règles est inférent ou productif, moins il est discriminant, et inversement. Mais par leurs formes ces courbes révèlent en même temps que l'avantage inférentiel doit être relativisé, et combien au final la taille des corpus doit être plus importante encore. L'on

constate en effet que si à q constant, la couverture est bien meilleure pour p=3 que pour p=4 ou plus (courbes en rouge), cet avantage ne semble se réaliser de façon notable que pour des valeurs de p petites (inférieures à 5), mais surtout ne se maintenir que pour des valeurs de q, ou plus exactement des différences (q-p) relativement faibles, n'excédant pas quatre en l'occurrence. Au-delà, pour p constant, l'on observe que la couverture, après avoir un temps progressée, décroît à partir de q=p+4 environ (courbes en bleu discontinu). En particulier pour p=3, on note une décroissance à partir de q=8. Décroissance qui indique que les effets inférentiels du chaînage de règles ne semblent assurer une meilleure couverture que pour les phrases dont la taille est légèrement supérieure à p.

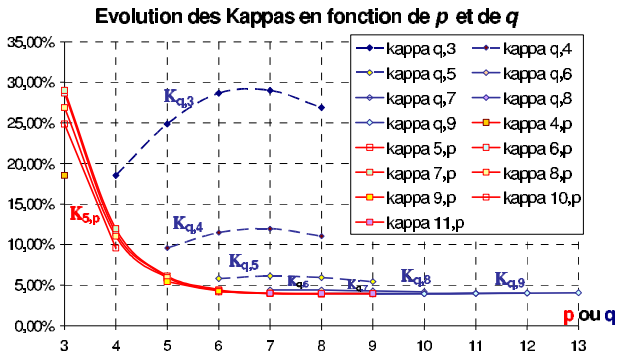


Figure 8 : La couverture de rang q des règles d'ordre p ($\kappa_{q,p}$) décroît à mesure que p croît. Elle croît légèrement puis décroît à mesure que q croît.

Pour les phrases plus longues, les effets semblent s'estomper rapidement, à mesure que q croît, ainsi que les tendances des courbes l'annoncent. L'espoir d'une meilleure couverture au prix d'une moins bonne discriminance se perd, puisque plutôt que d'évoluer en sens inverse, discriminance et couverture semblent, au-delà d'un certain seuil, évoluer dans un même sens, celui d'une dégradation à mesure que q croît. Ce résultat souligne en fait moins l'importance de la capacité inférentielle des règles de succession courtes, que l'insuffisance récurrente des corpus d'apprentissage.

8 Conclusion

Dans le résumé et l'introduction nous avons argumenté en faveur et pour l'extension de la taille des règles de succession couramment mises en œuvre dans les étiqueteurs. Les arguments, détaillés dans les sections 2 à 7, ont été de montrer que ne mettre en œuvre que les règles binaires et ternaires, c'est, dans plus de 98% des cas en situation d'étiquetage, très rapidement surgénérer un très grand nombre de règles non attestées, que l'avantage inférentiel ne permet pas de résorber, la proportion des règles engendrées attestées ou attestables restant à moins d'une fraction de pour cent du total des règles engendrées, dès que l'on dépasse la taille 5 ou 6.

La seconde partie de l'argumentation a été de montrer qu'il n'était pas déraisonnable non plus d'envisager cette extension, et d'ajouter aux règles binaires et ternaires, des règles de tailles supérieures. Pourquoi ? Parce que nous avons pu *découvrir* que cette extension pouvait être limitée, sans perte ou pratiquement, et que cette limite se situait aux environs de 12. Une taille de règle au delà de laquelle l'on constate, pour trois langues différentes, le français, l'arabe, et l'anglais, que l'on n'engendre plus, – ou très peu, c'est-à-dire bien en deçà du pour cent – de règles non attestables. Une taille qui pourrait trouver ses fondements dans la notion de proposition ou de phrase simple, dont la taille moyenne est précisément inférieure à la taille moyenne de la phrase en général, (soit 27 pour le corpus du Monde).

Les protocoles expérimentaux proposés sont simples. Ils n'exigent de disposer que d'un corpus aussi grand que possible, dûment étiqueté et découpé en phrases. En fait, seules les séquences d'étiquettes associées aux phrases sont utiles. C'est pourquoi, disponibilité des corpus aidant, nous espérons voir s'étendre ces observations à d'autres langues d'une part, et pour des volumes de corpus plus grands, d'autre part.

Cette étude des règles en soi, sous l'angle de leur fonctionnement interne, c'est-à-dire sous l'angle de leur assemblage ou aspect génératif, indépendamment de leur utilisation dans l'étiquetage, en analyse donc, nous a amené à poser le problème d'une évaluation intrinsèque, *en définition* avons-nous proposé de dire, d'un système de règles d'une façon générale. Le meilleur argument en faveur de règles ayant une portée supérieure à deux ou à trois est encore d'en montrer l'efficacité en comparant

les performances auxquelles ces règles conduisent, au regard des performances obtenues en utilisant des règles plus classiques, binaires ou ternaires. Bien entendu, et nous aurons à effectuer ces évaluations *en usage*. Mais nous voulons insister sur la complémentarité de ces deux visions de l'évaluation, « *en usage* » *vs* « *en définition* », et des potentialités de la seconde, dans une perspective prédictive ou comparative, comme nous l'avons signalé.

Cette distinction évaluation *en usage vs* évaluation *en définition* semble bien s'appliquer à notre cas qui est un cas d'analyse. Mais qu'en est-il de cette distinction si l'application englobante n'est pas, à l'image de l'étiquetage, une application d'analyse, mais une application de synthèse. Comment évaluer *en définition* un système de règles intervenant dans un programme de génération de phrases par exemple ?

Faut-il voir dans cette dichotomie « en usage » *vs* « en définition », la dichotomie « analyse » *vs* « synthèse » ? Ou encore la distinction « performance » *vs* « compétence » ? Ce qui conduirait à parler d'une évaluation *en analyse* (ou *en performance*) par opposition à une évaluation *en synthèse* (ou *en compétence*). La question est ouverte. Mais bien que nous n'ayons pas pour l'heure une vision claire de ce que pourrait être une évaluation *en définition* d'un ensemble de règles orienté vers un système de génération automatique de phrases comme en traduction par exemple, il nous semble que la distinction « en définition » *vs* « en usage » devrait encore subsister. La génération ou synthèse serait dans ce cas, comme l'est l'analyse, une application particulière qui met en œuvre un ensemble de règles dont l'évaluation indépendante reste pertinente et opératoire. A moins que seule la génération ne puisse faire l'objet d'une grammaire sur laquelle s'appuierait l'analyse. Ce qui ne remet pas en cause l'idée d'une double évaluation d'un système de règles, mais débouche sur une question qui dépasse le cadre de ce papier.

Remerciements

Le présent travail a été réalisé dans le cadre des deux projets EurADic (Action Technolangue du Ministère de la recherche), et MUSCLE (6ème PCRD). Il a en outre bénéficié de l'accueil de la Faculté des Lettres de l'Université de la Manouba, sous les auspices du MRSTDC.

Références

ADDA, G., MARIANI, J., PAROUBEK, P., RAJMAN, M., & LECOMTE, J. (1999). "L'action GRACE d'évaluation de l'assignation des parties du discours pour le français". *Langues*, 2(1).

ANDREEWSKY A., FLUHR C. (1973), "Expérience de constitution d'un programme d'apprentissage pour le traitement automatique du langage", Actes de *COLING 1973*, Volume 2.

DEBILI F. (1977), "Traitements syntaxiques utilisant des matrices de précédence fréquentielles construites automatiquement par apprentissage", *Thèse de Docteur-Ingénieur*, Université Paris VII, U.E.R. de Physique, Septembre 1977.

DEROSE STEVEN J., (1988), "Grammatical Category Disambiguation by Statistical Optimization", *Computational Linguistics*, Vol. 14, Num. 1.

FAIRON C., SENELLART J. (1999), "Réflexions sur la localisation, l'étiquetage, la reconnaissance et la traduction d'expressions linguistiques complexes", Actes de *TALN 1999*.

LEECH G., GARSIDE R., BRYANT M. (1994), "CLAWS4: The tagging of the British National Corpus", *Proceedings of the 15th International Conference on Computational Linguistics, COLING 94,* Kyoto, Japan.

MÉRIALDO B. (1994), "Tagging English text with a probabilistic model", *Computational Linguistics*, 20.2.

NASR A., VOLANSCHI A. (2004), "Couplage d'un étiqueteur morpho-syntaxique et d'un analyseur partiel représentés sous la forme d'automates finis pondérés", Actes de *TALN 2004*.

PAROUBEK P., RAJMAN M. (2000), "MULTITAG, une ressource linguistique produit du paradigme d'évaluation", Actes de *TALN 2000*.

TOUTANOVA K., KLEIN D., MANNING C. D., SINGER Y. (2003), "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network", Actes de *HLT-NAACL*, 2003.

VALLI A., VÉRONIS J. (1999), "Etiquetage grammatical des corpus de parole: problèmes et perspectives", *Revue Française de Linguistique Appliquée*.

Acquisition et évaluation sur corpus de propriétés de sous-catégorisation syntaxique

Didier Bourigault et Cécile Frérot

ERSS – CNRS & Université Toulouse le Mirail
5, allées Antonio Machado
31 058 Toulouse Cedex 1
{didier.bourigault,cecile.frerot}@univ-tlse2.fr

Mots-clés : analyse syntaxique, ambiguïté de rattachement prépositionnel, sous-catégorisation syntaxique
Keywords: syntactic parsing, PP attachment disambiguation, subcategorization lexicon

Résumé

Cette étude est menée dans le cadre du développement de l'analyseur syntaxique de corpus Syntex et porte sur la tâche de désambiguïsation des rattachements prépositionnels. Les données de sous-catégorisation syntaxique exploitées par Syntex pour la désambiguïsation se présentent sous la forme de probabilités de sous-catégorisation (que telle unité lexicale - verbe, nom ou adjectif - se construise avec telle préposition). Elles sont acquises automatiquement à partir d'un corpus de 200 millions de mots, étiqueté et partiellement analysé syntaxiquement. Pour évaluer ces données, nous utilisons 4 corpus de test de genres variés, sur lesquels nous avons annoté à la main plusieurs centaines de cas de rattachement prépositionnels ambigus. Nous testons plusieurs stratégies de désambiguïsation, une stratégie de base, une stratégie *endogène* qui exploite des propriétés de sous-catégorisation spécifiques acquises à partir du corpus en cours de traitement, une stratégie *exogène* qui exploite des propriétés de sous-catégorisation génériques acquises à partir du corpus de 200 millions de mots, et enfin une stratégie *mixte* qui utilisent les deux types de ressources. L'analyse des résultats montre que la stratégie mixte est la meilleure, et que les performances de l'analyseur sur la tâche de désambiguïsation des rattachements prépositionnels varient selon les corpus de 79.4 % à 87.2 %.

Abstract

We carry out an experiment aimed at using subcategorization information into a syntactic parser for PP attachment disambiguation. The subcategorization lexicon consists of probabilities between a word (verb, noun, adjective) and a preposition. The lexicon is acquired automatically from a 200 million word corpus, that is partially tagged and parsed. In order to assess the lexicon, we use 4 different corpora in terms of genre and domain. We

assess various methods for PP attachment disambiguation : an exogeous method relies on the sub-categorization lexicon whereas an endogenous method relies on the corpus specific ressource only and an hybrid method makes use of both. The hybrid method proves to be the best and the results vary from 79.4 % to 87.2 %.

1 Introduction

Les nombreux travaux sur le développement de parseurs statistiques concernent la langue anglaise et tendent à utiliser comme corpus d'apprentissage et comme corpus de test des portions de la section du Wall Street Journal du Penn TreeBank (Charniak, 1997). Outre qu'elle permet d'éviter la tâche l aborieuse de construction de corpus annotés, cette démarche présente l'immense avantage de po uvoir comparer les parseurs entre eux (Ratnaparkhi *et al.*, 1994 ; Pantel et Lin, 1998). Cette exploitation mono-corpus pose cependant la question de la stabilité des performances en fonction du type de corpus, comme le mentionnent Kilgarrif et Grefenstette (2003 :341) : « *there is little work on assessing how well one language model fares when applied to a text type that is different from that of the training corpus* ». Par ailleurs, il est maintenant bien connu que, dans tout corpus, certaines unités lexicales ont des propriétés syntaxiques de sous-catégorisation spécifiques, qui peuvent donc varier d'un domaine à l'autre (R oland, Jurafsky, 1998 ; Basili *et al.*, 1999). Or peu de travaux relatent des expériences sur la variation des performances de l'analyseur en fonction du type de corpus à traiter, sur le problème de la possible variation inter-corpus et sur celui de la nécessaire adaptation des règles de l'analyseur à un corpus donné. On peut néanmoins citer (Sekine, 1997 ; Gildea, 2001 ; Slocum, 1986).

Dans cet article, nous nous intéressons à l'acquisition et à l'évaluation sur corpus de données de sous-catégorisation syntaxique. Cette étude est menée dans le cadre du développement de l'analyseur syntaxique de corpus Syntex et porte sur la tâche de désambiguïsation des rattachements prépositionnels[1] (section 2). Les données de sous-catégorisation syntaxique exploitées par Syntex pour la désambiguïsation se présentent sous la forme de probabilités de sous-catégorisation (que telle unité lexicale - verbe, nom ou adjectif – se construise avec telle préposition). Dans la section 3, nous décrivons comment elles sont acquises automatiquement à partir d'un corpus de 200 millions de mots, étiqueté et partiellement analysé syntaxiquement. La section 4 est consacrée à l'évaluation sur des données acquises sur 4 corpus de test de genres variés, sur lesquels nous avons annoté à la main plusieurs centaines de cas de rattachement prépositionnels ambigus. Dans la section 5, nous présentons plusieurs stratégies de désambiguïsation : une stratégie de base, une stratégie *endogène* qui exploite des propriétés de sous-catégorisation spécifiques, acquises à partir du corpus en cours de traitement, une stratégie *exogène* qui exploite des propriétés de sous-catégorisation génériques, acquises à partir du corpus de 200 millions de mots, et enfin une stratégie *mixte* qui utilise les deux types de ressources. L'analyse des résultats (section 6) montre que la stratégie mixte est la plus performante, et que les performances de l'analyseur sur la tâche de désambiguïsation des rattachements prépositionnels varient selon les corpus.

[1] Nous nous intéressons dans cet article aux prépositions autres que *de*. Le traitement de la préposition *de* repose sur les mêmes principes, mais est sensiblement plus complexe (Frérot *et al.*, 2003).

2 Syntex, un analyseur syntaxique de corpus

Cette expérience est menée dans le cadre du développement de l'analyseur syntaxique de corpus Syntex (Bourigault, Fabre, 2000). Syntex est un analyseur en dépendance qui prend en entrée un corpus de phrases étiquetées[2], et calcule pour chaque phrase les relations de dépendance syntaxique entre les mots. C'est un analyseur en couches (Aït-Moktar et al., 2002) : le corpus est analysé en plusieurs passes, différents modules prenant successivement en charge une relation syntaxique de dépendance donnée, et les sorties d'un module constituant les entrées du module suivant. Chaque module est constitué d'un ensemble de règles construites « à la main ».

Syntex est un analyseur semi-lexicalisé. Le module qui effectue les rattachements prépositionnels exploite des données lexico-syntaxiques de sous-catégorisation, exprimées sous la forme de probabilités qu'une unité lexicale donnée (verbe, nom, adjectif) se construise avec telle ou telle préposition. Le rattachement des prépositions à leur recteur s'effectue en deux passes : (1) recherche des candidats recteurs, (2) choix d'un recteur. Un premier module (*rechercher-candidats*) traite l'ensemble des phrases du corpus, et recherche pour chaque préposition, le ou les mots susceptibles de régir cette préposition. Ce module est constitué de règles qui reconnaissent un certain nombre de configurations linéaires de mots et de catégories morphosyntaxiques à gauche de la préposition au sein desquelles sont identifiés des mots susceptibles de régir la préposition. Ces règles s'appuient sur les relations de dépendance placées par les modules antérieurs, et sont capables d'aller chercher des candidats recteurs dans des configurations relativement complexes, incluant par exemple des structures coordonnées ou des incises. Les configurations d'ambiguïtés, définies comme la succession des catégories grammaticales des candidats recteurs, sont très variées. Sur les 4 corpus de test présentés dans la section 4, la configuration 'V N', où seuls un verbe et un nom sont en compétition - configuration traitée dans beaucoup de travaux dont ceux, fondateurs, de Hindle et Rooth (1993) -, ne représente que 50 % des cas dans le corpus littéraire, environ 35 % dans le corpus journalistique et 15 % dans le corpus juridique et le corpus technique.

Au cours de la seconde étape du traitement des ambiguïtés prépositionnelles, le second module (*choisir-candidat*) revient sur chaque cas ambigu et choisit le recteur de la préposition parmi les candidats. Pour ce faire, ce module exploite des informations de sous-catégorisation associées aux couples (candidat, préposition). Depuis l'origine de nos travaux sur l'analyse syntaxique, ces informations sont acquises de façon endogène sur le corpus en cours de traitement (Bourigault, 1993). En effet, l'analyseur est utilisé dans différents contextes applicatifs, et principalement dans des applications de construction de terminologies ou d'ontologies spécialisées à partir de textes. Il traite des corpus spécialisés, thématiques, de taille moyenne (quelques centaines de milliers de mots, sur des domaines techniques, juridiques, médicaux). Les expériences menées sur de nombreux corpus ont montré que ces corpus renferment des spécificités lexicales, en particulier que certains mots, fréquents dans le corpus, manifestent des comportements syntaxiques spécifiques et imprédictibles. C'est pourquoi, nous avons porté nos efforts depuis une dizaine d'années sur le développement de procédures d'apprentissage endogène sur corpus qui permettent à l'analyseur d'acquérir lui -

[2] Nous utilisons actuellement les versions française et anglaise du Treetager (http://www.ims.uni-stuttgart.de)

même, par analyse du corpus à traiter, des informations de sous-catégorisation spécifiques à ce corpus, acquises à partir des cas non ambigus repérés par le module *rechercher-candidats*.

Devant les limites inhérentes à l'exploitation d'informations de sous -catégorisation acquises exclusivement sur le corpus en cours de traitement, nous travaillons à l'élaboration de ressources générales, susceptibles d'être exploitées pour tout corpus (Frérot *et al.*, 2003). Nous avons expérimenté l'utilisation d'un lexique de sous -catégorisation construit à partir des tables du *Lexique Grammaire* (Frérot, à paraître). Nous présentons dans ce travail une expérience d'acquisition de probabilités de sous-catégorisation à partir d'un corpus de 200 millions de mots.

3 Acquisition de propriétés de sous-catégorisation à partir d'un corpus de 200 millions de mots

Les méthodes d'acquisition de propriétés de sous-catégorisation exploitent classiquement des corpus étiquetés de grande taille (Basili, Vindigni, 1998). Le Web est aussi considéré comme source potentielle d'acquisition (Gala Pavia, 2003). Dans notre étude, nous utilisons comme base d'apprentissage un corpus de 200 millions de mots, constitué des articles du journal *Le Monde*, des années 1991 à 2000 (corpus LM10[3]). Nous ne prétendons pas que ce corpus soit représentatif de la « langue générale », mais nous considérons que sa taille et sa diversité thématique en font un corpus référentiellement et linguistiquement peu marqué, à partir duquel il est possible d'acquérir des données de sous-catégorisation relativement génériques. La procédure d'acquisition est adaptée des méthodes d'apprentissage endogène intégrées dans Syntex. La méthode de calcul des probabilités de sous-catégorisation s'appuie sur un ensemble de triplets (recteur, préposition, régi) extraits d'une analyse syntaxique du corpus LM10 effectuée par Syntex. La procédure d'acquisition se déroule en deux étapes, au cours desquelles la même méthode de calcul de probabilités est lancée successivement sur deux ensembles différents de triplets : une étape d'amorçage et une étape de consolidation.

Au cours de l'étape d'amorçage, le module *rechercher-candidats* traite l'ensemble du corpus LM10, qui a été analysé par les modules antérieurs de Syntex, et construit, à partir des cas non-ambigus, c'est-à-dire ceux pour lesquels il n'a identifié qu'un seul candidat recteur pour la préposition, un ensemble de triplets (w,p,w'), où w est le recteur de la préposition p, et w' le mot (nom, ou verbe à l'infinitif) régi par la préposition. Le module *rechercher-candidats* compte aussi pour chaque mot w le nombre d'occurrences dans le corpus où ce mot n'est candidat *d'aucune* préposition. A l'issue du traitement de l'ensemble du corpus, on dispose des données de fréquence suivantes :

- F(w,0) : nombre d'occurrences non ambiguës où le mot w ne régit aucune préposition ;

[3] Ce corpus a été préparé, à partir de fichiers obtenus auprès de l'agence Elra, par Benoît Habert (LIMSI), qui a effectué les tâches de nettoyage, de balisage et de signalisation nécessaires pour transformer les fichiers initiaux en un corpus effectivement « traitable » par des outils de Traitement Automatique des Langues. Nous remercions Benoît Habert et le LIMSI de nous avoir permis de bénéficier de cette ressource.

- F(*w*,*p*,*w'*) : nombre d'occurrences non ambiguës où le mot *w* régit la préposition *p*, qui elle-même régit le mot *w'*.

A partir de ces données, un premier ensemble de probabilités de sous-catégorisation P(w,p) est calculé, selon la méthode décrite plus loin dans la présente section.

Au cours de l'étape de consolidation, le module *choisir-candidat* exploite ce premier lexique et traite à son tour l'ensemble du corpus LM10, analysé par le module *rechercher-candidats*. Il revient sur les cas ambigus et choisit le candidat recteur dont la probabilité de construction avec la préposition, fournie dans le premier lexique, est la plus importante. A partir de ces nouvelles annotations, un nouvel ensemble de triplets est constitué, qui inclut le précédent et auquel s'ajoutent les triplets (*w*,*p*,*w'*) issus des cas ambigus résolus. De nouvelles données de fréquence F(*w*,*p*,*w'*) et F(*w*,0) sont alors constituées, à partir desquelles un second ensemble de probabilités de sous-catégorisation est calculé, selon la méthode décrite ci-dessous. C'est le lexique construit à l'issue de cette étape de consolidation qui est utilisé dans Syntex.

La méthode de calcul des probabilités est simple. La probabilité est calculée comme une fréquence relative pondérée[4]. Soit T, l'ensemble des triplets (*w*,*p*,*w'*), obtenu à l'issue de l'étape d'amorçage ou à celle de consolidation. Pour un couple (*w*,*p*), on définit $E_{w,p}$ comme l'ensemble des mots *w'* tels que la fréquence F(*w*,*p*,*w'*) est supérieure à 0. On définit la *productivité* du couple (*w*,*p*), Prod(*w*,*p*), comme le cardinal de l'ensemble $E_{w,p}$, c'est-à-dire comme le nombre de mots *différents* que régit la préposition *p* quand elle-même est régie par le mot *w*. Nous utilisons ce coefficient pour pondérer la fréquence totale du couple (*w*,*p*). A fréquence égale, plus le couple (*w*,*p*) a été repéré avec des contextes *w'* différents, plus grande est estimée la propension du mot *w* à régir la préposition *p*. L'expérience montre que, dans des corpus thématiques, la très haute fréquence de certains syntagmes très répétitifs incluant le triplet (*w*,*p*,*w'*) vient biaiser la probabilité d'association lexicale entre *w* et *p*. La pondération proposée ci-dessus vise à limiter une telle surestimation et à accorder un poids non seulement à la fréquence de l'association, mais aussi à sa diversité. La formule de calcul de la probabilité pondérée est donnée dans le tableau 1 : F(*w*,*p*) est la fréquence totale du couple (*w*,*p*), F(*w*) est la fréquence totale du mot *w*, et • est un coefficient de normalisation, choisi de telle sorte que la somme des probabilités associées à un mot donné soit égale à 1.

[4] Nous n'avons pour le moment pas testé d'autres méthode de filtrage, comme celle de la distribution polynomiale (Manning, 1993).

T = { (w,p,w') / F(*w*,*p*,*w'*) > 0 }, ensemble de triplets
F(*w*,*p*,*w'*) : nombre de cas où le mot *w* régit la préposition *p*, elle-même régissant le mot *w'*
F(w,0) : nombre de cas où le mot *w* ne régit aucune préposition
$E_{w,p}$ = { w' / F(*w*,*p*,*w'*) > 0 }, le contexte du couple (*w*,*p*)
Prod(*w*,*p*) = Card($E_{w,p}$), la productivité du couple (*w*,*p*)
F(*w*,*p*) = $\bullet_{w' \bullet Ew,p}$ F(*w*,*p*,*w'*)
F(*w*)=F(*w*,0)+ $\bullet_{p}$ F(*w*,*p*)
P(*w*,0) = F(*w*,0)/F(*w*)
P(*w*,*p*)=F(*w*,*p*) /F(w)*log(1 + Prod(*w*,*p*))/ $\bullet$

Tableau 1. Méthode de calcul des probabilités de sous-catégorisation

Le nombre total d'occurrences de triplets (*w*,*p*,*w'*) à partir desquels les probabilités sont calculées est de l'ordre de 6,7 millions à l'issue de l'étape d'amorçage, et de 12 millions à l'issue de l'étape de consolidation. Le nombre total d'occurrences de mots ne régissant pas de préposition est d'environ 87 millions à l'issue de l'étape d'amorçage , et de 95 millions à l'issue de l'étape de consolidation. Les probabilités ne sont calculées que pour les couples (*w*,*p*) tels que la fréquence totale du mot *w* est supérieure à 20. Un couple n'est retenu dans le lexique de désambiguïsation que si la probabilité dépasse le seuil de 0.01. Le lexique final compte 6 693 verbes différents (chacun pouvant être présent avec plusieurs prépositions), 11 528 noms et 698 adjectifs.

4 Annotation

De façon générale, le développement d'un analyseur syntaxique robuste exige une méthode de travail qui assume la très grande variabilité des corpus sur le plan syntaxique. Les stratégies et règles des différents modules de Syntex sont à chaque expérimentation élaborées à partir de tests effectués sur plusieurs corpus, aussi diversifiés que possible, pour limiter les biais d'implémentation que pourrait introduire une approche mono-corpus. A la variabilité inter-corpus, il faut ajouter la variabilité intra-corpus. Pour éviter d'élaborer des règles trop dépendantes de telle ou telle configuration syntaxique ou unité lexicale, il faut sur chaque corpus annoter à la main un très grand nombre de cas. Dans le cadre de cette étude, nous avons évalué le lexique de sous-catégorisation sur 4 corpus de test, de genres variés, dans lesquels nous avons validé à la main plusieurs centaines de cas :

- BAL. Le roman « Splendeurs et misères des courtisanes », d'Honoré de Balzac (199 789 mots) : 672 cas validés
- LMO. Un extrait du journal *Le Monde* (673 187 mots) : 1 238 cas validés

- REA. Un corpus de comptes-rendus d'hospitalisation dans le domaine de la réanimation chirurgicale (377 967 mots) : 646 cas validés

- TRA. Le *Code du travail* de la législation française (509 124 mots) : 1 150 cas validés

Les règles d'annotation sont les suivantes : (1) ne pas valider de cas où il y a des erreurs d'analyse des modules antérieurs, en particulier des erreurs d'étiquetage, autrement dit on évalue le module de rattachement prépositionnel dans des contextes où les informations sur lesquelles il s'appuie sont justes ; (2) se donner la possibilité de retenir comme valides deux recteurs pour une préposition donnée, en particulier pour les constructions à verbe support (*apporter une aide à*) ; (3) ne pas valider certains cas trop répétitifs, afin de ne pas sur représenter un cas trop spécifique au corpus, comme par exemple dans le corpus CTRA, où les cas de rattachement des participes passés à la préposition sont massifs (ex: *définir les modalités visées à l'article*) ; (4) valider de manière indifférenciée des groupes prépositionnels arguments ou circonstants. Ce dernier point est important, et peut prêter à controverse, si on ne replace pas la tâche d'annotation dans le contexte de l'évaluation des performances d'un analyseur syntaxique. La distinction argument/circonstant, ou complément essentiel/complément circonstanciel, ne fait pas l'objet d'un consensus dans la communauté linguistique. En dehors des cas trivaux, choisis en général soigneusement pour illustrer cette distinction, la confrontation avec des énoncés réels met à mal la clarté de cette distinction (Fabre, Frérot, 2002). Dans ces conditions, la tâche essentielle dévolue à l'analyseur est d'abord de choisir le bon recteur parmi un ensemble de recteurs possibles, et ensuite seulement, et éventuellement, de distinguer le type de complément.

5 Méthode de désambiguïsation

L'algorithme de désambiguïsation mis en œuvre dans le module *choisir-candidat* est simple. Nous comparons 4 stratégies différentes, selon le type des données de sous-catégorisation qu'elles exploitent.

- Mode *base*. En mode base, le module *choisir-candidat* se contente de choisir comme recteur le premier candidat dans l'ordre linéaire de phrase, c'est -à-dire le plus éloigné de la préposition[5].

- Mode *exogène*. En mode exogène, le module *choisir-candidat* exploite le lexique de sous-catégorisation construit à partir du corpus LM10 (section 3). Il choisit le candidat dont la probabilité est la plus élevée. On distingue exogène 1 et exogène 2, selon que le lexique utilisé est obtenu après la phase d'amorçage ou après la phase de consolidation.

- Mode *endogène*. En mode endogène, le module *choisir-candidat* exploite le lexique de sous-catégorisation construit à partir du corpus en cours d'analyse[6]. Avant

[5] Globalement - sur l'ensemble des corpus et sur l'ensemble des configurations d'ambiguïté -, cette stratégie est meilleure que celle qui choisirait le candidat le plus proche.

[6] Selon la méthode décrite dans la section 3, sans l'étape de consolidation.

d'exploiter les probabilités de sous-catégorisation, il exploite la liste des fréquences des triplets (*w*,*p*,*w'*) construite par le module *rechercher-candidats* : si *p* est la préposition et *w'* le mot qu'elle régit, le module choisit le candidat w_i pour lequel la fréquence F(w_i,*p*,*w'*) est la plus élevée. Sinon, il choisit le candidat dont la probabilité endogène est la plus élevée.

- Mode *mixte*. Le mode mixte est analogue au mode endogène, à ceci près que le module *choisir-candidat* choisit le candidat qui a la probabilité endogène *ou* la probabilité exogène la plus élevée.

Dans tous ces modes, la règle par défaut est celle de la stratégie de base, à savoir le choix du premier candidat.

	BAL	LMO	REA	TRA
base	83.0	70.3	59.9	65.5
endogène	83.5	80.1	78.0	82.3
exogène 1	85.7	85.5	65.3	85.9
exogène 2	86.9	86.6	66.3	86.3
mixte	86.6	85.9	78.3	87.3

Tableau 2. Taux de précision (%) des différentes stratégies de désambiguïsation sur les 4 corpus de test

6 Résultats et discussion

Le tableau 2 donne les taux de précision des différentes stratégies de désambiguïsation sur les 4 corpus de test. On peut rapprocher ces résultats de ceux, récapitulés dans (Pantel et Lin, 1998), obtenus sur 3 000 cas ambigus extraits de la partie Wall Street Journal du Penn TreeBank par différentes méthodes : 81,6% avec une méthode supervisée utilisant un modèle d'entropie maximale (Ratnaparkhi *et al.*, 1994), 88,1% avec une méthode supervisée utilisant un dictionnaire sémantique (Stetina, Nagao, 1997) et 84.3% avec une méthode non supervisée utilisant des mots distributionnellement proches (Pantel, Lin, *op.cit.*). Etant donné que les langues, le type de corpus de test et les conventions d'annotations sont différentes, il est délicat de comparer ces chiffres avec ceux que nous présentons dans le tableau 4. Ceux-ci doivent être analysés de façon autonome et contrastive. Notons d'abord que les résultats des stratégies exogènes 1 et 2 justifient l'intérêt d'acquérir les informations de sous-catégorisation en 2 étapes (amorçage et consolidation, section 3). Le corpus médical (REA), qui est le plus spécialisé des 4 corpus de test, présente un comportement particulier. Sur ce corpus, les performances des différentes stratégies sont globalement moins bonnes que sur les 3 autres corpus, ce qui illustre le point que nous avons évoqué au début de cet article, à propos de la sensibilité des résultats des analyseurs aux genres des textes. Par ailleurs, la stratégie de base donne de très mauvais résultats sur ce corpus, alors qu'ils sont particulièrement bons sur le corpus littéraire. C'est uniquement sur le corpus médical qu'apparaît, de façon nette, la

nécessité d'exploiter des probabilités de sous-catégorisation spécifiques au corpus (apprentissage endogène). Sur ce corpus, la stratégie endogène donne de meilleurs résultats que la stratégie exogène, et la stratégie mixte est très légèrement supérieure à la stratégie endogène. Sur les corpus littéraire et journaliste, la stratégie exogène est meilleure que la stratégie mixte.

Les ressources de sous-catégorisation syntaxique construites à partir du corpus LM10 sont exploitées par l'analyseur sans avoir été validées manuellement, et les résultats montrent qu'elles sont performantes pour cette tâche. Il convient de préciser que, sur le plan linguistique, ces propriétés de sous-catégorisation ne sont pas comparables aux descriptions que l'on peut trouver dans des lexiques construits à la main, comme le Lexique Grammaire, dans les dictionnaires de langue ou dans les études de psycholinguistique. C'est particulièrement vrai pour les verbes. La probabilité qu'a un verbe de sous-catégoriser telle préposition est calculée à partir de toutes les occurrences (lemmatisées) de ce verbe, sans distinction des différentes acceptions du verbe, alors que l'on sait qu'un même verbe peut avoir des cadres de sous-catégorisation différents selon ses différents sens. Dans le contexte du développement d'un analyseur syntaxique « tout terrain », l'approximation à laquelle conduit ce lissage des sens est un mal nécessaire.

Références

AÏT-MOKTAR S., CHANOD J.-P, ROUX C. (2002), Robustness beyond shallowness : incremental deep parsing, *Natural Language Engineering Journal*, 8(2/3):121-147

BASILI R., PAZIENZA M.-T., VINDIGNI M. (1999), Adaptative Parsing and Lexical Learning, *Proceedings of VEXTAL'99* , Venise

BASILI R., VINDIGNI M. (1998), Adapting a Subcategorization Lexicon to a Domain, *Proceedings of the ECML98 Workshop TANLPS*, Chemnitz, Germany

BOURIGAULT D. (1993), An endogenous Corpus Based Method for Structural Noun Phrase Disambiguation, In *Proceedings of the 6th Conference of the European Chapter of ACL (EACL),* pp. 81-86, Utrecht, The Netherlands

BOURIGAULT D., FABRE C. (2000), Approche linguistique pour l'analyse syntaxique de corpus, *Cahiers de Grammaire,* 25, Université Toulouse le Mirail, pp. 131-151

CHARNIAK E. (1997), Statistical Parsing with a Contexte-Free Grammar and Word Statistics. *Proceedings of the AAAI97 Conference*, Browne University, Rhode Island, pp.598-603

FABRE C., FRÉROT C (2002), Groupes prépositionnels arguments ou circonstants : vers un repérage automatique en corpus. Actes de la Conférence TALN, pp. 215-224.

FRÉROT C. (2005), *Etude en corpus variés sur l'intégration de ressources linguistiques générales dans un analyseur syntaxique*, Thèse en sciences du langage de l'Université Toulouse le Mirail

FRÉROT C., BOURIGAULT D., FABRE C. (2003), Marier apprentissage endogène et ressources exogènes dans un analyseur syntaxique de corpus. Le cas du rattachement verbal à distance de la préposition « de », in *Revue t.a.l.*, 44-3

GALA PAVIA N. (2003), *Un modèle d'analyseur syntaxique robuste basé sur la modularité et la lexicalisation de ses grammaires*, PhD, University of Paris XI, Orsay

GILDEA D. (2001), Corpus Variation and Parser Performance. In Lillian Lee and Donna Harma, editors, in *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pp. 167-202

HINDLE D., ROOTH M. (1993), Structural Ambiguity and Lexical Relations. *Computational Linguistics,* 19(1):103-120

KILGARRIFF A., GREFENSTETTE G. (2003), Introduction to the special issue of Web as Corpus. *Computational Linguistics*, 29:3, pp. 333-338

MANNING C. (1993), Automatic Acquisition of Large Subcategorization Dictionary from Corpora, *Proceedings of the 31st Meeting of the Association for Computational Linguistics*, Columbus

PANTEL P., LIN D. (2000), An unsupervised approach to prepositional phrase attachment using contextually similar words. In K. VijayShanker and Chang-Ning Huang, editors, *Proceedings of the 38th Meeting of the Association for Computational Linguistics*, pp. 101-108, Hong Kong

RATNAPARKHI A., REYNAR J., ROUKOS S. (1994), A Maximum Entropy Model for Prepositional Phrase Attachment. *Proceedings of the ARPA Workshop on Human Language Technology*, Morgan Kaufmann

ROLAND D., JURAFSKY, D. (1998). How Verb Subcategorization Frequencies Are Affected By Corpus Choice. *Proceedings of Coling-ACL,* pp. 1122-1128

SEKINE S. (1997), The domain dependence of parsing. *Proceedings of the Fith Conference on Applied Natural Language Processing,* pp. 96-102

SLOCUM J. (1986), How one might Automatically Identify and Adapt to a Sublanguage: An Initial Exploration, in Grishman R. and Kittredge R., eds., *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*. Lawrence Erlbaum Associates, Hillsdale, N.J., 1986, pp. 195-210

STETINA J., NAGAO M. (1997), Corpus-based PP attachment ambiguity resolution with a semantic dictionary. In J. Zhou and K. Church editors, *Proceedings of the 5th Workshop on Very Large Corpora,* Beijing and Hongkong.

TALN 2005

12ème conférence annuelle
sur le
Traitement Automatique des Langues Naturelles

POSTER

Induction de règles de correction pour l'étiquetage morphosyntaxique de la littérature de biologie en utilisant l'apprentissage actif

Ahmed Amrani (1), Yves Kodratoff (2) et Oriane Matte-Tailliez (2)

(1)ESIEA Recherche, 9 rue Vésale, 75005 Paris, France.
amrani@esiea.fr
(2) LRI, UMR CNRS 8623, Bât. 490, Université Paris 11, 91405 Orsay.
{yk, oriane}@lri

Mots-clés : Etiquetage morphosyntaxique, Apprentissage de règles, Apprentissage actif, fouille de textes.
Keywords: Part-of-speech tagging, rule learning, active learning, text-mining.

Résumé Dans le contexte de l'étiquetage morphosyntaxique des corpus de spécialité, nous proposons une approche inductive pour réduire les erreurs les plus difficiles et qui persistent après étiquetage par le système de Brill. Nous avons appliqué notre système sur deux types de confusions. La première confusion concerne un mot qui peut avoir les étiquettes 'verbe au participe passé', 'verbe au passé' ou 'adjectif'. La deuxième confusion se produit entre un nom commun au pluriel et un verbe au présent, à la 3ème personne du singulier. A l'aide d'interface conviviale, l'expert corrige l'étiquette du mot ambigu. A partir des exemples annotés, nous induisons des règles de correction. Afin de réduire le coût d'annotation, nous avons utilisé l'apprentissage actif. La validation expérimentale a montré une amélioration de la précision de l'étiquetage. De plus, à partir de l'annotation du tiers du nombre d'exemples, le niveau de précision réalisé est équivalent à celui obtenu en annotant tous les exemples.

Abstract In the context of Part-of-Speech (PoS)-tagging of specialized corpora, we proposed an approach focusing on the most 'important' PoS-tags because mistaking them can lead to a total misunderstanding of the text. After tagging a biological corpus by Brill's tagger, we noted persistent errors that are very hard to deal with. As an application, we studied two cases of different nature: first, confusion between past participle, adjective and preterit; second, confusion between plural nouns and verbs, 3rd person singular present. With a friendly user interface, the expert corrected the examples. Then, from these well-annotated examples, we induced rules. In order to reduce the cost of annotation, we used active learning. The experimental validation showed improvement in tagging precision and that on the basis of the annotation of one third of the examples we obtain a level of precision equivalent to the one reached by annotating all the examples.

1 Introduction

L'étiquetage morphosyntaxique est une étape importante pour la tâche d'extraction d'informations à partir de textes bruts et spécialisés. Cette étape consiste à associer à chaque mot son étiquette grammaticale en fonction de sa morphologie et de son contexte. Les étiqueteurs morphosyntaxiques actuels atteignent des performances très satisfaisantes en précision (plus de 95%) (Paroubek, Rajman, 2000). Ces bons résultats s'expliquent par le fait que les travaux en question se situent dans le domaine de l'apprentissage supervisé où le corpus de test est de même nature que le corpus d'apprentissage. Un pré-requis pour la construction d'un étiqueteur est la disponibilité d'un corpus annoté de taille importante. L'acquisition d'un tel corpus est coûteuse. D'autre part, les systèmes d'étiquetage ont tous des difficultés sur les cas difficiles. Les décisions sont le plus souvent fondées sur l'examen des contextes locaux (tels que les trigrammes de mots), qui résolvent mal les cas qui demanderaient une analyse plus globale et approfondie (Valli, Veronis, 1999).

Il existe deux approches principales pour l'apprentissage de règles : la création de règles à partir d'arbres de décision (Quinlan, 1993) et la technique d'apprentissage directe de règles comme dans l'algorithme RIPPER (Cohen, 1995). L'algorithme d'apprentissage de règles propositionnelles, PART (Frank, Witten, 1998), combine les deux approches précédentes. Chaque règle induite par PART a la forme d'une conjonction de conditions : Si T_1 et T_2 et ... T_n alors la classe est C_x. (Si T_1 et T_2 et ... T_n) est appelé le corps de la règle et (Cx) est la classe cible à apprendre. Chaque condition T_i teste une valeur particulière d'un attribut. La condition a la forme suivante : $A_i = v$, où A_i est un attribut symbolique et v est une valeur possible de A_i.

L'apprentissage actif est une technique qui permet de réduire le nombre d'exemples à annoter. Cette technique consiste à sélectionner les exemples les plus instructifs, pour lesquels le modèle courant est le plus incertain. L'apprentissage actif est de plus en plus utilisé dans des applications de traitement du langage naturel telles que l'étiquetage morphosyntaxique (Engelson, Dagan, 1999), le parsage stochastique (Tang et al, 2002) et la reconnaissance d'entités nommées (Shen et al, 2004).

Dans cet article, nous proposons une méthodologie basée sur l'apprentissage de règles de correction. Ces règles sont employées pour résoudre les erreurs d'étiquetage qui persistent après l'application de l'étiqueteur de Brill (Brill, 1994) et d'ETIQ (Amrani et al, 2004).

2 Méthodologie d'Etiquetage morphosyntaxique

L'approche proposée consiste à adapter un étiqueteur induit à partir d'un corpus généraliste à un corpus de spécialité. Notre système est basé sur l'étiqueteur de Brill (Brill, 1994). Cet étiqueteur utilise un apprentissage supervisé à base de transformations pour engendrer deux listes ordonnées de règles : règles lexicales et règles contextuelles. ETIQ[1](Amrani et al, 2004), l'étiqueteur que nous avons conçu, permet à l'expert de détecter les erreurs de l'étiqueteur de Brill, produites sur les corpus de spécialité. A l'aide d'ETIQ, l'expert visualise le résultat de l'étiquetage de Brill; il peut faire des requêtes lexicales ou contextuelles pour visualiser des

[1] http://ww.lri.fr/ia/genomics/. Téléchargement d'une version de démonstration du logiciel ETIQ.

groupes de mots (et leurs étiquettes) ayant des caractéristiques morphologiques ou contextuelles similaires. En fonction des erreurs détectées, l'expert insère des règles lexicales et contextuelles pour les corriger.

Après l'application de l'étiqueteur de Brill et d'ETIQ (Amrani et al, 2004; Amrani et al, 2005), nous avons remarqué, à l'aide du logiciel ETIQ, que certaines confusions spécifiques et difficiles à résoudre persistent. Voici les confusions les plus sérieuses : (1) JJ (adjectif) et NN (nom commun, singulier) pour quelques mots très fréquents comme *complex*. (2) VBN (verbe participe passé), JJ et VBD (verbe au passé) comme *transformed*. (3) VBZ (verbe au présent, troisième personne du singulier) et NNS (nom commun, pluriel) comme *functions* et *contacts*.

L'expert annote les exemples correspondant aux confusions identifiées. Il corrige ou il confirme l'étiquette du mot cible de chaque exemple. Afin de réduire le nombre d'exemples à annoter, nous utilisons l'apprentissage actif. Pour étudier l'impact de la représentation des exemples sur la performance, nous avons fait varier la taille des contextes aussi bien que les attributs utilisés pour représenter les exemples. Ces exemples servent à apprendre automatiquement des règles qui corrigent l'étiquette du mot en fonction de son contexte. Ces règles sont appliquées à la suite des règles contextuelles existantes.

2.1 Apprentissage actif

Nous calculons une mesure de distance entre chaque couple d'exemples. Puis, un ensemble initial d'exemples est sélectionné puis annoté. A partir de cet ensemble, nous apprenons un modèle. A chaque itération, un nouvel ensemble d'exemples pertinents est sélectionné puis annoté. La stratégie de sélection est basée sur la confiance et la diversité. Chaque étape est détaillée dans les sections suivantes.

2.1.1 Mesure de distance entre deux exemples

Chaque exemple est représenté comme suit: le mot cible est pris dans une fenêtre de n mots de chaque côté. Chaque mot est représenté par un ensemble d'attributs correspondant à son étiquette morphosyntaxique et à ses caractéristiques morphologiques. Soit l'exemple x représenté comme suit, où $(m = 2n+1)$ est le nombre d'attributs et $V_{x,y}$ est la valeur de l'attribut qui est à la position y de l'exemple x.

$$Exemple\ x: [V_{x,-n}\ V_{x,-(n-1)} \ldots V_{x,0} \ldots V_{x,(n-1)}\ V_{x,n}]$$

La mesure globale de distance entre deux exemples A et B ($G\_dist(ex_A,ex_B)$) est basée sur les distances ($(L\_dist(V_{A,k},V_{B,k}))$) entre les valeurs de chaque attribut. Pour chaque attribut (k), nous comparons ses valeurs ($V_{A,k}$ and $V_{B,k}$) dans les exemples: si les valeurs sont égales alors la distance est de 0; si les valeurs sont différentes alors la distance est de 1.

$$si\ (V_{A,k} = V_{B,k})\ alors\ \ L\_dist(V_{A,k},V_{B,k}) = 0,\ \ si\ (V_{A,k} = V_{B,k})\ alors\ \ L\_dist(V_{A,k},V_{B,k}) = 1$$

La mesure globale de distance (G_dist) entre deux exemples A (ex_A) et B (ex_B) est calculée comme suit, où W_k sont les poids donnés aux attributs de sorte que les attributs des mots les plus près du mot central soient les plus importants dans la mesure:

$$G\_dist(ex_A, ex_B) = \frac{\sum_{k=-n}^{n} W_k * L\_dist(V_{A,k}, V_{B,k})}{\sum_{k=-n}^{n} W_k}.$$

2.1.2 Stratégie de sélection des exemples

Tout d'abord, nous sélectionnons un échantillon initial représentatif de tous les exemples. Pour ce faire, nous utilisons l'algorithme des k-moyennes (Jain et al, 1999; Tang et al., 2002). Cet algorithme est basé sur la mesure de distance définie précédemment. Nous obtenons un ensemble composé de *nbc* groupes. Chaque groupe contient des exemples similaires. L'échantillon initial est constitué à partir d'une sélection aléatoire d'un pourcentage α d'exemples de chaque groupe. Cet échantillon nous sert à apprendre un modèle initial. Ensuite, les autres exemples sont sélectionnés de manière itérative. A chaque itération, nous utilisons deux critères pour la sélection : la confiance et la diversité.

L'utilisation du critère de la confiance consiste à choisir les exemples pour lesquels le modèle courant n'est pas satisfaisant. L'incertitude du modèle au sujet d'un exemple peut être due au fait que les exemples semblables sont sous-représentés dans l'ensemble d'apprentissage, ou bien que les exemples semblables sont intrinsèquement complexes. Nous tirons profit de la disponibilité de la confiance en classification du modèle courant. L'algorithme d'apprentissage de règles (par exemple PART (Frank, Witten, 1998)) assigne un degré de confiance à chaque règle induite. Pour chaque exemple non annoté, nous affectons le degré de confiance de la règle de laquelle il vérifie les conditions.

Le but du critère de la diversité (Shen et al., 2004) est de maximiser l'utilité inductive d'un ensemble d'exemples. Nous préférons les ensembles d'exemples hétérogènes. En choisissant un nouvel exemple non annoté, nous le comparons avec tous les exemples précédemment choisis dans l'ensemble courant. Si la similitude entre eux est au dessus d'un seuil β, l'exemple n'est pas ajouté dans l'ensemble. De cette façon, nous évitons de choisir les exemples trop semblables (valeur de similitude $\geq \beta$) dans un ensemble.

La stratégie globale de sélection des exemples est décrite comme suit : les exemples non-annotés sont ordonnés selon la confiance. A chaque itération, nous choisissons un ensemble de *nb* exemples de la manière suivante: D'abord, nous sélectionnons un exemple candidat ($Exemple_i$) avec une valeur de confiance minimale. Ensuite, nous évaluons le critère de diversité et nous ajoutons l'exemple candidat $Exemple_i$ à l'ensemble si seulement $Exemple_i$ est assez différent de n'importe quel exemple précédemment inséré dans l'ensemble. Le seuil β est fixé à une valeur comprise entre la valeur maximale de similitude et la moyenne des similitudes par paires dans l'ensemble des exemples non annotés.

3 Validation expérimentale

Pour les expérimentations, nous avons utilisé un corpus de 600 résumés d'articles MEDLINE (Amrani et al, 2004) de biologie moléculaire. Ce corpus a été étiqueté par l'étiqueteur de Brill, puis par ETIQ. A partir de ce corpus, nous avons présenté à l'annotateur 4133 exemples où le

mot cible est étiqueté VBN et 3298 exemples où le mot cible est étiqueté NNS. Le nombre total d'exemples NNS était de 7708 dont 4410 sont des mots non-ambigus. L'annotateur a classé les mots cibles en VBN, JJ ou VBD pour le premier jeu d'exemples et NNS ou VBZ pour le deuxième jeu. Pour améliorer la précision, nous avons représenté les exemples comme suit : pour le cas des VBN, le mot cible est pris dans une fenêtre de 10 mots (5 mots à gauche et 5 mots à droite) et chaque mot du contexte est représenté par : son étiquette morphosyntaxique, le groupe auquel appartient son étiquette (verbal, nominal ou autre) et le mot est un verbe auxiliaire ou non. Pour le cas des NNS, le mot cible est pris dans une fenêtre de 6 mots : 3 mots à droite et 3 mots à gauche. En plus des attributs utilisés pour représenter les exemples des VBN, nous avons utilisé les suffixes et les préfixes les plus fréquents des mots. A partir de ces exemples, nous avons induit des règles avec les algorithmes PART (pour les VBN) et RIPPER (pour les NNS). Nous avons calculé les précisions de l'étiqueteur de Brill, d'ETIQ et d'ETIQ enrichi par les règles induites (voir Figure 1). La précision des règles induites a été calculée par la méthode «validation croisée 10 fois».

Confusion / %de précision	Brill	Brill+ ETIQ	Brill+ ETIQ+Règles induites
VBN→VBN-VBD-JJ (PART)	54	76	94
NNS→NNS-VBZ (RIPPER)	92	96	97,5

Figure 1 : Précisions obtenues sur deux jeux d'exemples de confusions d'étiquettes.

Nous avons appliqué la stratégie de l'apprentissage actif aux exemples correspondent à l'ambiguïté VBN-VBD-JJ. Parmi les 4133 exemples disponibles, nous avons pris 3100 exemples pour l'apprentissage actif, et 1033 exemples pour le test. Le modèle initial a été construit à partir de 423 exemples. A chaque itération, nous avons sélectionné 100 exemples. L'expérience a été répétée 5 fois. La courbe (figure 2) représente les valeurs moyennes obtenues. La précision obtenue avec tous les exemples (3100) est de 93,5.

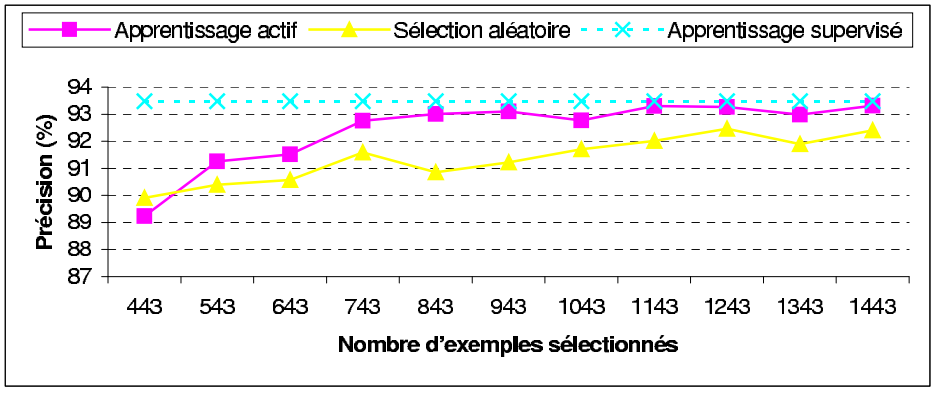


Figure 2 : Apprentissage actif *versus* sélection aléatoire.

4 Conclusions et perspectives

Dans le cadre d'une méthodologie globale pour l'étiquetage morphosyntaxique des corpus de spécialité, nous avons complété notre approche pour traiter efficacement les problèmes

d'étiquetage pointus. Après la détection des contextes ambigus et particuliers, les mots cibles sont annotés (exemples). A partir de ces exemples, nous avons induit des règles de correction. Nous avons obtenu une nette amélioration de la précision d'étiquetage. Pour réduire le nombre d'exemples à annoter, nous avons utilisé l'apprentissage actif avec une stratégie de sélection basée sur la confiance et la diversité. En annotant seulement un tiers des exemples, nous obtenons des performances équivalentes à celles obtenues en annotant tous les exemples. Nous étendrons cette approche à d'autres classes d'ambiguïtés. Nous envisageons également de considérer d'autres méthodes d'apprentissage, par exemple : la Programmation Logique Inductive. La combinaison optimale des règles obtenues par différents algorithmes pourrait améliorer les performances. Le critère de diversité, utilisé pour l'apprentissage actif, peut être amélioré en utilisant une valeur de similitudes (β) optimale.

Références

AMRANI, A., AZE, J., KODRATOFF, Y. (2005) ETIQ: Logiciel d'aide à l'étiquetage morpho-syntaxique de textes de spécialité. *Dans la revue RNTI, numéro spécial EGC'2005*.

AMRANI, A., KODRATOFF, Y., MATTE-TAILLIEZ, O. (2004) A Semi-automatic System for Tagging Specialized Corpora, *PAKDD 2004*, Sydney, LNAI, Vol. 3056, pp 670-681.

BRILL, E. (1994) Some Advances in Transformation-Based Part of Speech Tagging, *AAAI*, Vol. 1, pp 722-727.

COHEN, W. (1995) Fast Effective Rule Induction, *Proceedings of the 12th International Conference on Machine Learning*.

ENGELSON, S.A., DAGAN, I. (1999) Committee-Based Sample Selection for Probabilistic Classifiers. *Journal of Artifical Intel-ligence Research.*

FRANK, E., WITTEN, I.H. (1998) Generating Accurate Rule Sets Without Global Optimization, Shavlik, J. Eds., *Proceedings of the 15th ICML*, Madison, Wisconsin, pp 144-151.

JAIN, A. K., MURTY, M. N., AND FLYNN, P. J. Data clustering: a review. *ACM Computing Surveys*, 31(3):264-323.

PAROUBEK, P., RAJMAN, M. (2000) Chapitre 5: Etiquetage morpho-syntaxique, Ingénierie des Langues, sous la direction de Jean-Marie Pierrel, Collection "*Information Commande Communication*", aux Editions Hermes Science, 2000 pp 131-148.

QUINLAN, J.R (1993) *C4.5: Programs for Machine Learning*, Morgan Kaufmann San Mateo.

SHEN, D., ZHANG, J., SU, J., ZHOU, G., TAN, C-L. (2004) Multi-Criteria-based Active Learning for Named Entity Recognition. *Proceedings of ACL* 2004.

TANG, M., LUO, X., ROUKOS, S., 2002. Active Learning for Statistical Natural Language Parsing. *In Proceedings of the ACL* 2002.

VALLI, A., & VERONIS, J. (1999). Etiquetage grammatical de corpus oraux: problèmes et perspectives. *Revue Française de Linguistique Appliquée*, IV(2), 113-133.

Application du métalangage définitionnel de la BDéf au traitement formel de la polysémie

Lucie Barque (1), Alain Polguère (2)
(1) Lattice - Université Paris 7
UFRL, Case 7003
75251 Paris cedex 5, France
lbarque@linguist.jussieu.fr
(2) OLST - Université de Montréal
C.P.6128, succ. Centre-ville
Montréal (Qc), H3C 3J7, Canada
alain.polguere@umontreal.ca

Mots-clefs : Base de données lexicale, métalangage définitionnel, Lexicologie Explicative et Combinatoire, polysémie

Keywords: lexical database, definitionnal metalanguage, Explanatory Combinatorial Lexicology, polysemy

Résumé Cet article a pour objet le métalangage définitionnel de la base de données lexicale BDéf, plus précisément l'utilisation de ce métalangage dans la modélisation des structures polysémiques du français. La Bdéf encode sous forme de définitions lexicographiques les sens lexicaux d'un sous-ensemble représentatif du lexique du français parmi lequel on compte environ 500 unités polysémiques appartenant aux principales parties du discours. L'article comprend deux sections. La première présente le métalangage de la BDéf et le situe par rapport aux différents types de définitions lexicales, qu'elles soient ou non formelles, qu'elles visent ou non l'informatisation. La seconde section présente une application de la BDéf qui vise à terme à rendre compte de la polysémie régulière du français. On y présente, à partir d'un cas spécifique, la notion de patron de polysémie.

Abstract We present the defining metalanguage of the BDéf lexical database ; more specifically, we focus on how this metalanguage can be used to model relations of polysemy in French. The BDéf contains lexical definitions for a representative subset of the French lexicon : around 500 polysemic words belonging to all major parts of speech. This paper contains two sections. Firstly, the BDéf metalanguage is introduced and positioned relative to different types of existing lexical definitions : formal vs. non formal definitions, definitions that are tailored or not for implementation. Secondly, the paper shows how the BDéf approach is used in a research whose goal is the modeling of the regular polysemy of the French language. The notion of pattern of polysemy is introduced using a specific example.

1 Place de la définition BDéf dans une typologie de la définition en lexicographie et en sémantique formelle

La BDéf, base de données lexicale développée à l'Observatoire de Linguistique Sens-Texte de l'Université de Montréal (Altman et Polguère, 2003), encode les définitions du *Dictionnaire Explicatif et Combinatoire* (Mel'čuk *et al.* 1984, 1988, 1992, 1999) dans un formalisme qui explicite leur structure interne.

Nous classons les différents types de définitions lexicales connus en deux grandes familles, caractérisées non sur la base de leurs fondements théoriques, mais sur celle des formalismes adoptés : d'un côté les approches qui ont recours à la paraphrase pour expliciter le sens d'une unité lexicale, de l'autre celles qui représentent le sens lexical au moyen de structures de traits.

1. **Les formes de la paraphrase** La paraphrase définitionnelle est une décomposition sémantique d'une unité lexicale organisée de façon linéaire. Le caractère linéaire de ce type de représentation lexicale est le point commun des trois formes de paraphrases présentées ci-dessous. Elles divergent par leur degré de formalisation.

   (a) **Les définitions lexicographiques standard** utilisent un métalangage proche de la langue objet qu'il décrit. Cela leur confère un caractère naturel qui répond bien à l'exigence de substituabilité de la paraphrase.

   (b) **Les définitions analytiques** suivent les mêmes principes définitoires (Aristote, ed. 2004) que les dictionnaires de langue courants mais sont construites de manière à pouvoir identifier clairement les différents éléments de la décomposition (Wierzbicka, 1987), (Mel'čuk *et al.*, 1995).

   (c) **Les définitions logiques** comme celles de (Dowty, 1979) sont très différentes des deux autres types de paraphrases du point de vue de la théorie sémantique sous-jacentes. Les formules logiques se présentent cependant elles aussi sous une forme linéaire qui permet une appréhension relativement naturelle du sens lexical.

2. **Les structures de traits** Comme les paraphrases définitionnelles, les structures de traits représentent une décomposition du sens lexical. Ici, la décomposition n'est pas organisée de façon linéaire mais prend la forme d'un ensemble de traits. Les théories de sémantique computationnelle adoptent en général ce mode de représentation de la décomposition lexicale (Pottier 1974), (Rastier 1987), (Pustejovsky, 1995).

Le formalisme d'encodage des définitions lexicales de la BDéf emprunte ses caractéristiques à l'une et l'autre des deux grandes familles qui viennent d'être distinguées. D'une part la BDéf, parce qu'elle se veut un outil servant à réfléchir sur la nature du sens lexical, s'attache à mettre en évidence les différentes composantes de sens qui forment la définition ainsi que la façon dont ces composantes s'organisent. Le formalisme adopté permet de rétablir facilement la forme linéaire de manière à aider le lexicographe lors de la construction des définitions. D'autre part, la BDéf doit constituer une ressource au service de la communauté du TAL et utilise de ce fait un formalisme très proche de celui des réseaux sémantiques. Une définition Bdéf peut ainsi être exploitée dans des calculs, comme nous le verrons dans la seconde section de cet article.

Une présentation détaillée des principes sous-jacents à la base de données BDéf et du formalisme d'encodage des définitions adopté figure dans (Altman et Polguère, 2003). Nous en présentons ici les grandes lignes. La figure 1 ci-dessous reproduit la fiche BDéf modélisant le sens de la lexie MAISON**I.1** [*Elle cherche la maison où habite son cousin*]. Cet exemple va nous servir

à illustrer les différentes notions décrivant la structure des définitions : les propositions élémentaires sont organisées en blocs définitionnels élémentaires, eux-mêmes regroupés en blocs de second niveau, la composante centrale et les différences spécifiques.

MAISON **I.1**

Composante centrale :

```
1 : habitation de X
```

différences spécifiques :

```
  /*Dimensions*/
2 : *1 grand.relativement

  /*Structure*/
3 : *1 constitué de niveau

  /*Matériau*/
4 : *1 fabriqué avec matériau.résistant
```

Typage des actants :

```
X : individu
```

FIG. 1 – La définition BDéf de MAISON **I.1**

La proposition élémentaire C'est la composante sémantique minimale de la définition. Chaque proposition élémentaire est constituée d'un prédicat accompagné des positions actancielles qu'il requiert et est identifiée par un numéro. Une position actancielle peut ainsi être occupée par un pointeur (`*1`, `*2`, ...) vers une autre proposition. Par exemple la proposition n°2 de la définition présentée dans la figure 1 est constituée du prédicat `grand` qui a comme argument la proposition n°1.

Le bloc définitionnel élémentaire Les propositions élémentaires sont regroupées en blocs définitionnels qui représentent des composantes « autonomes » de la définition. Chaque bloc est introduit par un en-tête synthétisant l'information qui y est contenue. Dans la définition de MAISON **I.1**, la proposition n°2 spécifie les `dimensions` d'une maison.

La composante centrale et les différences spécifiques Les blocs définitionnels élémentaires sont organisés au niveau supérieur en deux grands blocs : la composante centrale, qui correspond au sens général de la lexie décrite – ce que la tradition appelle le *genre prochain* – et les différences spécifiques.

2 Modélisation des liens de polysémie

Le développement de la BDéf et de son métalangage a donné lieu à un projet de modélisation des structures polysémiques du français, que nous présentons dans cette section.

2.1 Définitions

Un **vocable polysémique** est constitué d'un ensemble de lexies partageant les mêmes formes et qui ont en commun une composante sémantique non triviale[1]. Cette dernière propriété permet de distinguer deux lexies en relation de polysémie de deux lexies en relation d'homonymie (Mel'čuk *et al.*, 1995), (Victorri et Fuchs, 1996), (Kleiber, 1999).

Un **lien de polysémie** se définit comme un lien sémantique entre deux lexies d'un vocable polysémique. Nous nous intéresserons ici à un type particulier de lien de polysémie, celui qui résulte d'un processus de dérivation sémantique. On dira qu'une lexie L2 est dérivée sémantiquement d'une lexie L1 si le sens de L2 est « construit à partir du » sens de L1. On appelera dans ce cas de figure L1 la **lexie source** et L2 la **lexie dérivée.**

Un **patron de polysémie** modélise un lien de polysémie au moyen d'une paire de structures définitionnelles sous spécifiées. Le nom du patron doit rendre compte d'une part du changement (ou non) de l'étiquette sémantique (Polguère, 2003) et d'autre part de la transformation à laquelle on a affaire (par exemple, quel type de métonymie, quel type de métaphore, etc). Un patron de polysémie doit être suffisamment général pour pouvoir s'appliquer à **au moins deux couples** de lexies.

2.2 Un exemple de modélisation : le cas du vocable MAISON

Notre choix s'est porté ici sur deux lexies du vocable MAISON :MAISON**I.1** [*Elle cherche la maison où habite son cousin.*] et MAISON**I.2** [*Elle a quitté la maison à l'âge de 18 ans.*][2]. Les définitions des lexies sont présentées ci-dessous sous forme linéaire[3] :

MAISONI.1 de X ≡ ' */*habitation*/* habitation de l'individu X, */*Dimensions*/* relativement grande, */*Structure*/* constituée d'un ou plusieurs étages et */*Matériau*/* fabriquée avec des matériaux solides '

MAISONI.2 de X ≡ ' */*lieu occupé*/* lieu occupé par un individu ou un groupe d'individus X */*Limites*/* situé à l'intérieur de l'habitation de X '

Notons que la lexie dérivée est beaucoup plus contrainte dans son fonctionnement dans la phrase que la lexie source puisqu'elle ne s'emploie qu'au défini singulier et n'accepte pas le déterminant possessif : dans la phrase *Elle a quittée **sa** maison à l'âge de 18 ans*, il s'agit nécessairement de MAISON**I.1** et non de MAISON**I.2.**

La figure 2 ci-dessous représente le patron de polysémie **Synecdoque faible : (partie d'une) construction→lieu occupé** s'appliquant au couple MAISON**I.1**~MAISON**I.2.** Le lien entre les deux lexies est une **synecdoque** dans la mesure où le sens de la lexie dérivée désigne une partie de ce que dénote la lexie source : comme on le voit dans le patron de la figure 2, le prédicat de synecdoque `être_intérieur_de` lie le sens général de la lexie dérivée à celui de la lexie source[4]. Le fait que le second argument du prédicat `être_intérieur_de`, dans la définition

[1]Concernant l'identification de la polysémie, nous adoptons les critères descriptifs et méthodologiques de la lexicologie explicative et combinatoire, explicités dans (Mel'čuk *et al.*, 1995).

[2]Ces deux lexies ont des correspondants en anglais qui n'appartiennent pas au même vocable, respectivement HOUSE et HOME.

[3]Nous avons introduit (en italique) dans les définitions linéaires les en-têtes de blocs des définitions BDéf correspondantes afin de faciliter par la suite la lecture du patron de polysémie auquel s'applique ce couple de lexies.

[4]le sens général d'un lexie est représenté dans le patron par l'étiquette du bloc de la composante centrale de sa définition, ou une étiquette mère de celle-ci. Par exemple, l'étiquette `construction` est l'étiquette mère de

de MAISON**I.2**, soit `habitation de X` et non MAISON**I.1** (*i.e* une inclusion complète) nous amène à qualifier cette synecdoque de **faible.** Cette particularité est marquée linguistiquement par le fait que l'on peut dire *elle a quittée la maison à l'âge de 18 ans*, que cette personne ait habité dans une maison, dans un appartement ou dans une tout autre habitation. À l'inverse, on dira que les lexies ASSIETTE**I.1** [*Il a cassé trois assiettes.*] et ASSIETTE**I.2** [*Il a mangé trois assiettes de riz.*] sont liées par une **synecdoque forte** car on ne peut pas dire *Il a mangé trois assiettes de riz* si le riz n'est pas contenu dans une assiette (ASSIETTE**I.1**). Comme il a été signalé plus haut, un patron de polysémie doit rendre compte du type de lien de polysémie (dans notre cas, il s'agit d'une synecdoque faible) mais également de l'éventuel changement du type sémantique entre la lexie source et la lexie cible. Cette information figure dans le patron au niveau de la valeur de l'attribut `Composante centrale`.

LEXIE SOURCE
Composante centrale [1] `(partie d'une) construction` **Composante exportée** [1] **Adresse de la composante exportée** `/*Construction*/`
LEXIE DÉRIVÉE SÉMANTIQUEMENT
Composante centrale [2] `lieu occupé` **Adresse de la composante importée** `/*Limites*/` **Prédicat de synecdoque** `être_intérieur_de([2], [1])`

FIG. 2 – Patron de polysémie **Synecdoque faible : (partie d'une) construction→lieu occupé**

Le patron de polysémie qui vient d'être identifié modélise une alternance récurrente en français, quoique plus marginale que celles de la famille des synecdoques fortes. Le patron **Synecdoque faible : (partie d'une) construction→lieu occupé** peut par exemple s'appliquer également au couple de lexies BUREAU**I.2** [*Les fenêtres du bureau donnent sur la cour.*]∼BUREAU**I.3** [*Quand on est au bureau, le temps passe lentement.*] dans la mesure où l'on peut dire que *Jean a passé la journée au bureau* même si Jean n'a pu mettre les pieds dans son bureau (BUREAU**I.2**), retenu toute la journée en salle de réunion. Le degré de décomposition des définitions et l'organisation explicite de cette décomposition nous permet ainsi distinguer de manière fine des liens de polysémie proches. Cela nous permet par exemple de préciser la distinction entre polysémie systématique et polysémie régulière (Apresjan, 1974) : toute lexie étiquetée `construction` ou `partie de construction` est susceptible de « générer » une lexie étiquetée `lieu occupé` (par exemple ÉCOLE, ÉGLISE, APPARTEMENT, ...) – en cela il s'agit d'une polysémie systématique – mais la définition de la lexie dérivée sera différente selon que les deux lexies seront reliées par un lien de synecdoque forte ou faible.

3 Conclusion

Le travail qui vient d'être présenté doit mener à terme à la constitution d'une typologie des liens de polysémie du français. La classification des différents types de structures polysémiques d'une langue a déjà fait l'objet de plusieurs travaux, notamment sur le français (Martin, 1972*a*, 1972*b*) et sur le russe (Apresjan, 1974), mais ces travaux ne reposaient pas sur une exploitation systématique de descriptions lexicales formelles du type de celles qui sont fournies par la BDéf.

`habitation`. L'opérateur optionnel `partie de` permet de rendre plus général le patron. Pour une présentation détaillée de la notion d'étiquette sémantique dans les lexiques Sens-texte, voir (Polguère, 2003).

Cette recherche présente l'intérêt de contribuer à la rationalisation de la base de données, le but étant d'obtenir des définitions suffisamment homogènes et structurées pour mettre en relation les informations de la BDéf avec celles du DiCo, qui est quant à elle déjà entièrement informatisée (Steinlin *et al.*, 2004). L'exploitation de la BDéf vise par ailleurs à produire un ensemble de patrons de polysémie qui pourront par exemple servir de base à l'élaboration de règles pour le développement de lexiques incrémentaux du type du Lexique Génératif (Copestake et Briscoe, 1995), (Rappaport et Levin, 1998).

Références

Altman J., Polguère A. (2003), La BDéf : base de définitions dérivée du Dictionnaire explicatif et combinatoire, *Proceedings of the First International Conference on Meaning-Text Theory*, Paris, p. 43-54.

Apresjan J. (1974), Regular Polysemy, *Linguistics*, 142, p. 5-32.

Aristote (2004), *Les Topiques*, Librairie Philosophique J.Vrin, Paris.

Bouillon P., Busa F. (2001), *Generativity in the Lexicon*, Cambridge University Press, Cambridge.

Copestake A., Briscoe E. (1995), Semi-Productive Polysemy and Sense extension, *Journal of Semantics*, 12(1), p. 15-67 .

Dowty D. (1979), *Word Meaning and Montague Grammar*, Reidel, Dordrecht.

Kleiber G. (1999), *Problèmes de sémantique : la polysémie en question*, Presses Universitaires du Septentrion, Villeneuve d'Ascq.

Martin R. (1972), Esquisse d'une analyse formelle de la polysémie, *Travaux de linguistique et de littérature*, 10, p. 125-136.

Martin R. (1972), La polysémie verbale, esquisse d'une typologie formelle, *Travaux de linguistique et de littérature*, 17, p. 261-256.

Mel'čuk I., Clas A., Polguère A. (1995), *Introduction à la lexicologie explicative et combinatoire*, AUPELF-UREF/Duculot, Louvain-la-Neuve.

Mel'čuk I. *et al.* (1984, 1988, 1992, 1999), *Dictionnaire explicatif et combinatoire du français contemporain*, vol I-IV, Les Presses de l'Université de Montréal, Montréal.

Polguère A. (2000), Towards a theoretically-motivated general public dictionnary of semantic derivations and collocations for French, *Proceedings of EURALEX'2000*, Stuttgart, p. 512-527.

Polguère A. (2003), Étiquetage sémantique des lexies dans la base de données DiCo, *TAL*, 44, 2, p. ?.

Pottier B. (1974), *Linguistique générale. Théorie et description*, Klincksieck, Paris.

Pustejovsky J. (1995), *The Generative Lexicon*, MIT Press, Cambridge, Mass.

Rappaport Hovav M., Levin B. (1998), Building Verb Meanings, in M. Butt and W. Geuder (eds) *The projection of arguments : Lexical and Compositionnal Factors*, CSLI Publications, Stanford, p. 97-134.

Rastier F. (1987), *Sémantique interprétative*, Presses Universitaires de France, Paris.

Steinlin J., Kahane S., Polguère A., El Ghali A. (2004), De l'article lexicographique à la modélisation objet du dictionnaire et des liens lexicaux, *Proceedings of EURALEX'2004*, Lorient, p. 177-186.

Victorri B., Fuchs C. (1996), *La polysémie, construction dynamique du sens*, Hermès, Paris.

Wierzbicka A. (1987), *English Speech Act Verbs : A semantic dictionnary*, Academic, Sydney.

Apprentissage de relations prédicat-argument pour l'extraction d'information à partir de textes conversationnels

Narjès Boufaden (1), Guy Lapalme (1)
(1) RALI - Université de Montréal
Département d'Informatique et de Recherche Opérationnelle
Université de Montréal
C.P. 6128, succ. Centre-Ville
Montréal, Québec, H3C 3J7 Canada
{boufaden,lapalme}@iro.umontreal.ca

Mots-clefs : Apprentissage de relations prédicat-argument, extraction d'information

Keywords: Learning predicat-argument relations, information extraction

Résumé Nous présentons les résultats de notre approche d'apprentissage de relations prédicat-argument dans le but de générer des patrons d'extraction pour des textes conversationnels. Notre approche s'effectue en trois étapes incluant la segmentation linguistique des textes pour définir des unités linguistiques à l'instar de la phrase pour les textes bien formés tels que les dépêches journalistiques. Cette étape prend en considération la dimension discursive importante dans ces types de textes. La deuxième étape effectue la résolution des anaphores pronominales en position de sujet. Cela tient compte d'une particularité importante des textes conversationnels : la pronominalisation du thème. Nous montrons que la résolution d'un sous ensemble d'anaphores pronominales améliore l'apprentissage des patrons d'extraction. La troisième utilise des modèles de Markov pour modéliser les séquences de classes de mots et leurs rôles pour un ensemble de relations données. Notre approche expérimentée sur des transcriptions de conversations téléphoniques dans le domaine de la recherche et sauvetage identifie les patrons d'extraction avec un F-score moyen de 73,75 %.

Abstract We present the results of our approach for the learning of patterns for information extraction from conversational texts. Our three step approach is based on a linguistic segmentation stage that defines units suitable for the pattern learning process. Anaphora resolution helps to identify more relevant relations hidden by the pronominalization of the topic. This stage precedes the pattern learning stage, which is based on Markov models that include *wild card* states designed to handle edited words and null transitions to handle omissions. We tested our approach on manually transcribed telephone conversations in the domain of maritime search and rescue, and succeeded in identifying extraction patterns with an F-score of 73.75 %.

1 Introduction

Nous présentons notre approche d'apprentissage de patrons dans le contexte de l'extraction d'information à partir de textes conversationnels spécialisés. Cette étape est la dernière de notre approche proposée pour l'extraction d'information à partir de ces textes que nous avons présentée dans nos travaux précédents (Boufaden *et al.*, 2002; Boufaden *et al.*, 2005). Nous proposons une modélisation utilisant des modèles de Markov pour apprendre des relations prédicat-argument (séquences de classes sémantiques étiquetant le verbe et ses arguments) et les rôles[1] des arguments à partir de textes étiquetés sémantiquement. Ces textes sont des transcriptions[2] manuelles de conversations téléphoniques portant sur des incidents survenus en mer. Ce sont des compte rendus où les locuteurs se communiquent des informations sur un incident, par exemple un bateau en difficulté, sur les conditions météorologiques lors d'une mission de recherche ou sur le lieu de l'incident. Un exemple de conversation est donné au tableau 1.

Le système repose sur trois étapes et prend en entrée des séquences de classes sémantiques étiquetant les mots clé des énoncés où les étiquettes sont définies dans une ontologie du domaine. La première étape segmente les conversations en unités linguistiques à l'instar de la phrase pour les textes bien formés tels que les dépêches journalistiques (section 2.1). Cette étape prend en considération la dimension discursive très importante dans ce types de textes (Levelt, 1989). La deuxième effectue la résolution des anaphores pronominales en position de sujet (section 2.2). Cette étape tient compte d'une particularité des textes conversationnels : la pronominalisation du thème. Nous montrons que la résolution d'un sous ensemble des anaphores pronominales améliore l'apprentissage des patrons d'extraction. La troisième utilise les modèles de Markov pour modéliser les séquences de classes de mots et leurs rôles pour un ensemble de relations données (section 3). La comparaison de notre approche avec celles développées pour les textes bien formés montrent la pertinence de notre approche (section 4).

2 Problématique de l'apprentissage des patrons d'extraction

Un patron d'extraction est une structure qui permet le repérage des informations que nous voulons extraire et établit une relation entre ces éléments d'information. Il se caractérise par des contraintes syntaxiques (position des arguments dans une relation **sujet-verbe-objet**) et sémantiques (type de classes sémantiques) permettant le filtrage d'un sous-ensemble d'énoncés qui contiennent des informations pertinentes au domaine d'application. Parmi les principales difficultés de l'apprentissage des patrons d'extraction à partir de textes bien formés mentionnés dans la littérature (Grishman, 1998; Surdeanu *et al.*, 2003), nous retenons: (1) la diversité des constructions phrastiques contenant l'information pertinente et (2) l'association de nouveaux éléments d'information à des objets référencés par une anaphore.

Dans le contexte des textes conversationnels, ces difficultés sont amplifiées. D'une part, les irrégularités langagières telles que les répétitions et les reprises modifient la structure syntaxique des énoncés, tandis que l'aspect conversationnel a pour effet de répartir l'information sur plus d'un énoncé, par exemple lors d'échanges de type question-réponse. D'autre part, la présence importante de pronoms notamment à l'intérieur des unités thématiques augmente le nombre de

[1] Un rôle est un nom de champ défi ni dans un formulaire.

[2] Ces textes ont été fournis par le Centre de Recherche de la défense Canadienne. Ils ne sont pas annotés prosodiquement et nous n'avions pas les enregistrements originaux pour reconstituer la prosodie.

No	Loc	Énoncé
4	b:	ha, Ha, I don't know if I was handled over to you at all, but we've got an overdue boat VESSEL on the South Coast of Newfoundland, just in LOCATION the area quite between Fortune Bay and Trepassey. LOCATION .....*Incident* .....
5	b:	it's on the south east coast of Newfoundland. LOCATION .....*Incident* .....
6	b:	this is been going on for, for 24 hours that the case has, TIME or almost anyway, and we had an DFO King Air up flying AIRCRAFT STATUS this morning. TIME .....*Search-unit* .....
7	b:	they did a radar search for us in that area. STATUS MEANSOFDETECTION LOCATION
8	a:	yes. .....*Search-unit* .....

Table 1: Exemple de conversation dans le domaine de Recherche et sauvetage. Les mots soulignés sont les informations que nous voulons extraire. Les étiquettes sous les barres en soulignés sont des classes de mots importants. Les pointillés sont les frontières des unités linguistiques que nous détectons dans la section (2.1). *Incident* et *Search-unit* sont des exemples de relations que nous voulons modéliser par des modèles de Markov.

relations partielles (par opposition à une relation complète où tous les arguments sont définis). L'approche que nous proposons tient compte de ses difficultés. Tout d'abord, nous effectuons une segmentation en paires d'adjacence[3] qui détecte, par exemple, les paires de type question-réponse pour repgrouper dans une seule unité linguistique les éléments d'information présents dans une question et sa réponse. Ensuite, nous procédons à la résolution des anaphores pronominales en position de sujet pour diminuer le nombre des relations partielles. Enfin, nous relaxons la contrainte de contiguité des arguments de la relation "sujet-verbe-objet", en apprenant les patrons à partir de séquences d'étiquettes sémantiques de longueur variable.

2.1 Segmentation en unités linguistiques

À l'instar des travaux en segmentation linguistique de conversations (Stolcke, 1997), nous avons utilisé un modèle de Markov d'ordre 1 pour modéliser des séquences de traits composés de marques lexicales telles que `ok`, `well` et `?` caractéristiques des paires d'adjacence, mais aussi la longueur d'un énoncé ainsi que l'identité de locuteur. Contrairement aux approches proposées, nous n'avons pas utilisé la prosodie car celle-ci est absente de nos textes. Le modèle contient deux états représentant la classe des énoncés indépendants (E) et la classe des énoncés complétant une paire d'adjacence (PA). Nous avons validé notre modèle en effectuant 10 validations croisées sur notre corpus contenant 64 conversations (3481 énoncés) avec 80 % réservé

[3]Les paires d'adjacence sont deux tours de parole, chacun venant d'un locuteur distinct où le premier tour nécessite un second tour de parole d'un certain type (source `http://www.sil.org/linguistics/GlossaryOfLinguisticTerms/`).

à l'entraînement. La moyenne des erreurs de classification obtenue à partir des 10 validations croisées est de 15,9 %. L'analyse des erreurs de classification a montré que la source principale des erreurs est due à l'absence de marques lexicales pour certains énoncés de la classe PA. Dans ces cas, l'information prosodique absente dans nos transcriptions permettrait de combler le manque d'information lexicale.

2.2 Résolution des anaphores pronominales

Nous nous intéressons aux anaphores pronominales `they`, `we`, `she`, `he` et `it` en position de sujet[4]. Notre approche se base sur la structure thématique des conversations et sur une liste des étiquettes sémantiques[5] extraites à partir de chaque énoncé d'une unité thématique. L'importance de la structure thématique a déjà été soulignée pour la résolution des coréférences dans les conversations (Grosz *et al.*, 1995).

Le choix d'un antécédent est dirigé par deux contraintes de compatibilité: **sémantique** et **thématique**. La première fixe des associations possibles entre les étiquettes sémantiques et les pronoms. Tandis que la seconde fournie un antécédent par défaut, lorsqu'aucun antécédent compatible avec l'anaphore n'a été détecté dans les énoncés précédents de l'unité thématique courante ou de la précédente portant sur le même thème. Les valeurs par défaut sont les étiquettes les plus fréquentes calculées sur 31 conversations du corpus.

L'évaluation de notre approche a été effectuée sur 31 conversations de notre corpus, soit 161 anaphores pronominales en position de sujet. Le taux moyen d'erreurs de résolution obtenu est de 79,5 %. Bien que le résultat soit encourageant, certains choix de notre approche ont contribué à augmenter le taux d'erreurs, en particulier, le choix d'une approche linéaire (non hiérarchique) de segmentation en unités thématiques (Boufaden *et al.*, 2002) dans la segmentation automatique et la simplicité de notre approche dans le calcul des antécédents par défaut qui se base sur les fréquences obtenues sur le corpus.

3 Apprentissage des patrons d'extraction

Le but de cette étape est d'exploiter les associations entre les étiquettes sémantiques afin d'apprendre des patrons d'extraction qui expriment une relation prédicat-argument où les arguments ont un rôle spécifique pour une relation donnée. Des exemples d'étiquettes sémantiques utilisées sont présentées dans l'extrait de conversation du tableau 1.

3.1 Approche

Nous avons considéré cinq relations dans nos expériences:

1. *Missing-object* qui décrit Le bateau en difficulté, c'est-à-dire sa description, le nom de son propriétaire.
2. *Incident* qui décrit le type d'incident, la cause, le type d'appel de détresse.

[4]Levelt (Levelt, 1989), montre que les pronoms position de sujet sont souvent le résultat de la pronominalisation du thème d'une unité thématique.

[5]La structure thématique et les étiquettes sémantiques sont générées de manière automatique par des systèmes développés dans nos travaux précédents (Boufaden *et al.*, 2005).

Schémas d'extraction	Modèle de Markov	Rappel	Précision	F-score
Incident (62 énoncés)	Ordre 1	59,0 %	79,6 %	
	Ordre 2	63,8 %	85,0 %	72,9 %
Search-mission (27 énoncés)	Ordre 1	79,0 %	89,5 %	83,9 %
	Ordre 2	70,7 %	81,4 %	
Search-unit (93 énoncés)	Ordre 1	53,3 %	75,2 %	
	Ordre 2	52,9 %	76,9 %	62.7 %
Missing-object (38 énoncés)	Ordre 1	54,4 %	71,7 %	
	Ordre 2	70,8 %	80,8 %	75,5 %

Table 2: Rappel, précision et F-score de l'apprentissage des patrons d'extraction pour les formulaires *Incident*, *Mission*, *Search-unit* et *Missing-object*. Le rappel et la précision sont obtenus par la méthode de validation croisée "Leaving one out" pour les deux modèles de Markov. Le F-score est la moyenne des F-scores du meilleur modèle.

3. *Search-unit* qui parle de la ressource utilisée dans une mission de recherche.

4. *Mission* qui décrit le lieu de la mission, les conditions météorologiques, la date.

Pour chaque type de relation, nous avons modélisé les séquences des étiquettes avec un modèle de Markov. Nous avons entraîné chaque modèle sur un sous-ensemble du corpus qui contient des exemples positifs du type de relation ciblée.

3.2 Expériences et résultats

Nous avons effectué deux expériences afin de déterminer l'ordre du modèle de Markov qui donne les meilleures performances pour chaque patron d'extraction. Nous avons testé un modèle de Markov d'ordre 1 et un modèle d'ordre 2. Étant donné la taille modeste des corpus d'entraînement (<100) pour les différents patrons d'extraction, nous avons opté pour une validation croisée avec l'approche "Leaving one out". Les rappels[6], précisions et F-scores des meilleures performances sont indiquées au tableau 2.

Nous constatons que le patron d'extraction associé à la relation *Search-mission* présente une meilleure performance avec le modèle de Markov d'ordre 1, tandis que les autres patrons d'extraction *Missing-object*, *Incident* et *Search-unit* montrent de meilleurs résultats avec les modèles d'ordre 2.

Le choix de l'ordre du modèle dépend du taux des étiquettes sémantiques ayant plusieurs rôles possibles. Par exemple, dans l'unité thématique *Mission*, l'étiquette la plus fréquente est WEATHER-CONDITIONS avec une fréquence relative de 37,7 %. Cette dernière a un seul rôle dans la relation *Mission*, contrairement à l'étiquette NUMBER qui peut avoir le rôle d'une date ou d'une position géographique (en degré par exemple). Le choix de l'ordre dépend également du bruit introduit par les irrégularités langagières, notamment les reprises, agrandit la taille du contexte nécessaire pour désambiguïser un rôle.

[6]Le rappel correspond au nombre de rôles corrects générés par le système sur le nombre de rôles dans le corpus de test, tandis que la précision est le nombre de rôles corrects générés par le système sur le nombre de rôles qu'il fournit.

4 Conclusion

Nous avons analysé la problématique de l'apprentissage des patrons d'extraction pour des textes complexes peu étudiés en EI: les transcriptions de conversations. Nous avons modélisé les patrons d'extraction par des modèles de Markov qui associent des rôles aux arguments des prédicats avec un F-score de 73,75 %. Bien que les modèles de Markov aient été utilisés pour l'apprentissage de patrons (Seymore *et al.*, 1999), peu de travaux les ont utilisés pour apprendre les rôles sémantiques. De ces travaux, nous retenons ceux de Gildea (Gildea & Palmer, 2002) effectués sur des textes journalistiques avec un F-score de 82 %. D'autres approches ont été utilisées, notamment les arbres de décisions sur des textes bien formés avec un F-score de 83,7 % (Surdeanu *et al.*, 2003). Cependant, cette approche ne permet pas de tenir compte des séquences de longueurs variables que l'on retrouve avec les textes conversationnels.

Nous avons ajouté une étape de résolution des anaphores pronominales en amont de l'étape d'apprentissage de patrons. Notre approche a permis un taux de résolution des anaphores de 79,5 % améliorant ainsi le F-score moyen pour l'apprentissage de patrons de 68,6 %. Quelques travaux Surdeanu (Surdeanu & Harabagiu, 2002) ont utilisé une approche similaire pour améliorer l'extraction des informations en résolvant les coréférences aux entités nommées.

Références

BOUFADEN N., LAPALME G. & BENGIO Y. (2002). Découpage thématique des conversations: un outil d'aide à l'extraction. In *Actes de la 9^e conférence annuelle sur le traitement automatique des langues naturelles (TALN 2002)*, volume I, p. 377–382, Nancy, France.

BOUFADEN N., LAPALME G. & BENGIO Y. (2005). Repérage de mots informatifs à partir de textes conversationnels. *Traitement Automatique de la Langue*, **45**(3).

GILDEA D. & PALMER M. (2002). The necessity of syntactic parsing for predicate argument recognition. In *Proceedings of the 40th Annual Conference of the Association for Computational Linguistics (ACL 2002)*, p. 239–246, Philadelphie, Pennsylvanie.

GRISHMAN R. (1998). Information extraction and speech recognition. In *Proceedings of the DARPA Broadcast Transcription and Understanding Workshop*, Lansdowne, Virginie: Morgan Kaufmann Publishers.

GROSZ B., JOSHI A. & WEINSTEIN S. (1995). Centering: A Framework for Modeling the local Coherence of Discourse. *Computational Linguistics*, **21**(2), 203–225.

LEVELT W. J. M. (1989). *Speaking: From Intention to Articulation*. ACL-MIT Press Series in Natural Language Processing. MIT Press.

SEYMORE K., MCCALLUM A. & ROSENFELD R. (1999). Learning hidden Markov structure for information extraction. In *Proceedings of the AAAI-99 Workshop on Machine Learning for Information Extraction*, p. 37–42, Orlando, Floride.

STOLCKE A. (1997). Modeling linguistic segment and turn boundaries for n-best rescoring of spontaneous speech. In *Proceedings of EUROSPEECH 1997*, volume 5, p. 2779–2782, Rhodes, Grèce.

SURDEANU M., HARABAGIU S., WILLIAMS J. & AARSETH P. (2003). Using predicate-argument structures for information extraction. In E. HINRICHS & D. ROTH, Eds., *Proceedings of ACL 2003*, p. 8–15.

SURDEANU M. & HARABAGIU S. M. (2002). Infrastructure for Open-Domain Information Extraction. In M. MITCHELL, Ed., *Proceedings of HLT 2002*, p. 325–330, San Diego, Californie.

TALN 2005, Dourdan, 6–10 juin 2005

Un analyseur LFG efficace : SXLFG

Pierre Boullier, Benoît Sagot, Lionel Clément
INRIA - Projet Atoll
Domaine de Voluceau, B.P. 105, 78153 Le Chesnay Cedex, France
{pierre.boullier, benoit.sagot}@inria.fr,
lionel.clement@lefff.net

Mots-clefs : syntaxe, analyseur, LFG, désambiguïsation, forêt partagée

Keywords: syntax, parser, LFG, disambiguation, shared forest

Résumé Dans cet article, nous proposons un nouvel analyseur syntaxique, qui repose sur une variante du modèle *Lexical-Functional Grammars* (Grammaires Lexicales Fonctionnelles) ou *LFG*. Cet analyseur LFG accepte en entrée un treillis de mots et calcule ses structures fonctionnelles sur une forêt partagée. Nous présentons également les différentes techniques de rattrapage d'erreurs que nous avons mises en œuvre. Puis nous évaluons cet analyseur sur une grammaire à large couverture du français dans le cadre d'une utilisation à grande échelle sur corpus variés. Nous montrons que cet analyseur est à la fois efficace et robuste.

Abstract In this paper, we introduce a new parser based on the *Lexical-Functional Grammars* formalism (*LFG*). This LFG parser accepts as input word lattices and computes functional structures on a shared forest. We also present various error recovery techniques we have implemented. Afterwards, we evaluate this parser on a large-coverage grammar for French in the framework of a large-scale use on various corpus. We show that our parser is both efficient and robust.

1 Introduction

Pour pallier les difficultés algorithmiques des analyseurs syntaxiques sur du texte tout venant, il est aujourd'hui habituel d'appliquer un mode opératoire robuste (méthodes markoviennes, automates finis, etc.). Ces méthodes sont très satisfaisantes pour un grand nombre d'applications qui ne dépendent pas d'une représentation complexe de la phrase, mais la finesse d'analyse en pâtit tellement qu'il est illusoire d'avoir une représentation du syntagme ou des dépendances non locales qui satisfassent une définition linguistique sérieuse. C'est pour cette raison que nous proposons de bâtir un analyseur syntaxique qui soit à la fois compatible avec une théorie linguistique (ici LFG) et qui soit robuste face à la variabilité des productions langagières.

Le développement d'un nouvel analyseur syntaxique pour le formalisme LFG (*Lexical-Functional Grammars*, cf. p. ex. (Kaplan, 1989)) n'est pas en soi très original. Il en existe déjà un certain nombre, comme ceux de (Kaplan & Maxwell, 1994), (Andrews, 1990), ou (Briffault *et al.*, 1997). Toutefois, ils n'utilisent pas toujours de la manière la plus complète possible les différentes techniques algorithmiques de partage de calcul et de représentation compacte de l'information qui permettent d'écrire un analyseur efficace bien que le formalisme LFG, comme de nombreux formalismes qui reposent sur l'unification, soit NP-complet.

Pour utiliser au maximum ces techniques, il nous a donc fallu adapter LFG sans pour autant diminuer son pouvoir d'expression. Associée à un analyseur non-contextuel (CF) tabulaire, cette variante de LFG nous permet d'effectuer efficacement l'analyse d'énoncés complexes. La construction des structures de constituance ne pose théoriquement[1] pas de problème particulier, car elles sont décrites par des grammaires non-contextuelles (CFG) sous-jacentes aux LFG et appelées ici grammaires *support*. Mais la construction efficace des structures fonctionnelles est beaucoup plus problématique. Nous avons développé un module de calcul de ces structures qui partage les sous-structures communes à plusieurs analyses. De plus, des mécanismes de rattrapage d'erreur à tous les niveaux en font un analyseur robuste. Cet analyseur, appelé SXLFG, est décrit ci-dessous puis évalué avec une grammaire du français sur des corpus variés.

2 L'analyseur SXLFG : analyse standard

2.1 L'analyseur Earley

Le moteur de SXLFG est un analyseur CF général qui traite la grammaire support de la LFG. L'ensemble des analyses qu'il produit est représenté sous la forme d'une forêt partagée[2]. L'évaluation fonctionnelle se fait dans une seconde phase au cours d'un parcours bas-haut de cette forêt[3]. L'entrée de l'analyseur est un treillis de mots transformé par le *lexeur* en un treillis de lexèmes (terminaux de la CFG et structures fonctionnelles sous-spécifiées associées). Un post-traitement (facultatif) permet alors de désambiguïser.

L'analyse de la grammaire support est réalisée par une évolution de l'analyseur Earley décrit dans (Boullier, 2003) : il prend en entrée des treillis de mots et permet de récupérer les erreurs syntaxiques (cf. section 3.1). Traiter un treillis en entrée ne nécessite pas, d'un point de vue théorique, des changements considérables à l'algorithme Earley, même aidé d'un guide régulier.

[1] Même si, en pratique, la disponibilité d'un *bon* analyseur est déjà plus délicate.

[2] Rappelons que cette structure permet de représenter en une taille polynomiale en n, nombre de mots du source, l'ensemble potentiellement non borné des arbres d'analyse.

[3] Ce parcours assure que si un symbole se trouve en partie droite d'une production reconnue, toutes les structures fonctionnelles associées à ce symbole ont déjà été calculées (nos forêts partagées sont non cycliques).

2.2 Calcul des structures fonctionnelles

Disposant d'une forêt partagée en sortie de l'analyse CFG, nous devons maintenant calculer les structures fonctionnelles. Bien entendu, la méthode qui consiste à *déplier* la forêt pour en extraire chaque arbre sur lequel on évalue les structures fonctionnelles est impratiquable en termes de temps de calcul. En revanche, l'autre possibilité, une évaluation des structures fonctionnelles directement sur la forêt partagée, est toujours un sujet de recherche. Le problème se simplifie cependant si l'on suppose, comme c'est le cas dans SXLFG, que l'évaluation des équations fonctionnelles associées à une production CFG ne modifie pas les structures fonctionnelles associées aux symboles de sa partie droite. Cette légère restriction dans l'écriture des équations fonctionnelles ne diminue pas pour autant le pouvoir d'expression.

La conséquence directe de cette évaluation *bottom-up* des structures fonctionnelles est que toute sous-forêt n'est évaluée qu'une seule fois et son calcul partagé entre tous ses parents. L'autre conséquence est qu'à chaque nœud de la forêt est associée non pas une structure fonctionnelle unique mais une disjonction de structures fonctionnelles. Très souvent, le résultat de cette évaluation est donc un grand nombre de structures fonctionnelles associées à la racine de la forêt.

2.3 Désambiguïsation

La sortie de l'étape précédente (sauf échec, voir partie suivante) est une forêt partagée de structures de constituants associée à un ensemble de structures fonctionnelles avec partage de structures communes. Ces informations peuvent être la description d'une ou de plusieurs analyses. Il faut donc pouvoir désambiguïser, c'est-à-dire choisir parmi ces analyses celle qui est la plus vraisemblable. Deux familles de techniques sont envisageables : les techniques probabilistes et les techniques à règles. Suivant sur ce point (Clément & Kinyon, 2001), nous utilisons un ensemble de règles qui est une refonte et une extension des trois principes simples qu'ils énoncent et qui s'applique sur les structures fonctionnelles[4]. Chacune de nos règles est appliquée successivement (on peut en changer l'ordre, voire ne pas toutes les appliquer). L'application d'une règle consiste à éliminer les analyses qui ne sont pas optimales au sens de cette règle[5].

À l'issue de ce mécanisme de désambiguïsation sur les structures fonctionnelles, la forêt d'analyse (qui représente les structures en constituants) est filtrée afin qu'elle corresponde exactement aux structures fonctionnelles retenues. En particulier, si la désambiguïsation est complète, ce filtrage rend en général une structure en constituants unique (un arbre).

[4]Cf. (Kinyon, 2000) pour une argumentation sur l'importance de désambiguïsation en se fondant sur des structures comme les arbres de dérivation TAG ou les structures fonctionnelles LFG et non sur celles en constituants.

[5]Nos règles, dans leur ordre d'application par défaut, sont :

règle 1 : *Préférer les analyses maximisant la somme des poids des lexèmes utilisés* ; parmi les entrées lexicales de poids supérieur à la moyenne se trouvent les multi-mots, qui sont ainsi favorisés.
règle 2 : *Préférer les noms communs avec déterminant.*
règle 3 : *Préférer les arguments aux modifieurs, et les relations auxiliaire-participe aux arguments* (le calcul se fait récursivement sur toutes les (sous-)structures).
règle 4 : *Préférer les arguments les plus proches* (même remarque).
règle 5 : *Préférer les structures les plus enchâssées.*
règle 6 : *Trier les structures selon le mode des verbes* (on préfère récursivement les structures à l'indicatif à celles au subjonctif, et ainsi de suite).
règle 7 : *Trier selon les catégories des gouverneurs d'adverbes.*
règle 8 : *Choisir une analyse au hasard* (pour garantir qu'on rende une analyse et une seule).

3 Mécanismes pour l'analyse robuste

3.1 Rattrapage d'erreur pendant l'analyse

La détection d'une erreur dans l'analyseur Earley peut être la manifestation de deux phénomènes : la CFG support n'est pas assez couvrante ou l'énoncé n'est pas du français. Bien entendu, même si l'analyseur ne distingue pas ces deux situations, le concepteur de la grammaire doit y réagir différemment. Le traitement des erreurs dans les analyseurs est un sujet de recherche qui a surtout été abordé dans le cas déterministe et très peu dans le cas des analyseurs CF généraux. Pour des raisons de place, nous ne pouvons décrire ici le mécanisme général de rattrapage CFG que nous avons développé. Il fera l'objet d'une publication ultérieure.

Le calcul des structures fonctionnelles échoue si et seulement si aucune structure fonctionnelle n'est associée à la racine de la forêt partagée. Cette situation d'erreur provient du fait que les contraintes (d'unification) spécifiées par les équations fonctionnelles n'ont pas pu toutes être vérifiées ou que les structures fonctionnelles résultantes sont incohérentes.

Un premier échec déclenche une deuxième évaluation des structures fonctionnelles sur la forêt partagée, au cours de laquelle les vérifications de cohérence sont supprimées. En cas de succès, on obtient à la racine un certain nombre de structures fonctionnelles incohérentes. Si cette seconde tentative échoue, on recherche dans la forêt partagée tous les nœuds *maximaux* qui ont des structures fonctionnelles et dont aucun des pères n'a de structure fonctionnelle. Ils correspondent donc à des analyses partielles disjointes éventuellement incohérentes[6].

3.2 Sur-segmentation des énoncés inanalysables

Malgré les mécanismes exposés précédemment, il arrive que l'analyseur SxLfg ne rende aucune analyse. Ceci peut être dû à l'expiration d'un délai maximum que l'on peut donner en paramètre (*time-out*), ou au fait que le rattrapage d'erreur de l'analyseur Earley n'a pas été capable de produire une analyse raisonnable. La cause peut en être l'insuffisance de la couverture de la grammaire ou un énoncé d'entrée par trop déraisonnable.

Pour cette raison, nous avons réalisé une surcouche à SxLfg qui permet une *sur-segmentation* des énoncés agrammaticaux. L'idée est qu'il arrive fréquemment que des portions de l'énoncé d'entrée soient analysables en tant que phrases, alors même que l'énoncé d'entrée dans son ensemble ne l'est pas. Nous découpons donc en *segments* les énoncés inanalysables (découpage de niveau 1), puis, le cas échéant, redécoupons en segments les segments de niveau 1 inanalysables[7] (découpage de niveau 2) et ainsi de suite. Les niveaux de découpage correspondent successivement aux frontières probables de phrases, aux ponctuations fortes, aux ponctuations faibles, aux coordonnants, et enfin aux frontières de mots.

La qualité de l'analyse décroît évidemment avec le niveau de découpage. Si le découpage de niveau 1 ne pose aucun problème, des difficultés apparaissent au niveau 2. Les niveaux 3 et 4 sont véritablement du rattrapage. Et le niveau 5 n'est là que pour analyser toutes les phrases possibles, et en particulier celles dont on sait analyser certains morceaux de niveau 1 ou 2.

[6] Le processus de désambiguïsation présenté à la section 2.3 s'applique alors à tous les nœuds maximaux.

[7] Un énoncé peut être découpé en deux segments de niveau 1 dont le premier est analysable. Seul le second sera alors sur-segmenté au niveau 2. Et seuls les segments de niveau 2 inanalysables seront sur-segmentés, etc.

4 Mise en œuvre et évaluation

4.1 Mise en œuvre

Nous avons utilisé SXLFG à grande échelle pendant la campagne EASy d'évaluation des analyseurs syntaxiques. Nous l'avons couplé avec une grammaire LFG développée pour XLFG (Clément & Kinyon, 2001), étendue et adaptée aux contraintes liées à ce que SXLFG calcule de manière *bottom-up* les structures fonctionnelles sur la forêt d'analyse CFG. Le lexique et la chaîne de traitement pré-syntaxique mis en œuvre sont décrits dans (Boullier *et al.*, 2005).

4.2 Évaluation

Dans cette section, nous n'évaluerons pas la *qualité* d'une analyse qui dépend pour l'essentiel de la grammaire et qui nécessiterait de disposer d'un corpus de référence annoté manuellement[8]. Nous nous concentrons ici sur l'efficacité de notre système en présentant les résultats obtenus pendant la campagne EASy et sur les corpus EUROTRA et TSNLP.

Corpus	#phrases	couverture (sans vérif. de coh.)	couverture (avec vérif. de coh.)	temps d'analyse			
				moy.	méd.	$\geq$ 0.1s	$\geq$ 1s
EUROTRA	334	94.61%	84.43%	0.33s	0.02s	22.2%	6.0%
TSNLP	1661	98.50%	79.12%	0.03s	0.00s	2.8%	0.6%
EASy	40859	66.62%	41.95%	n.d.[10]			

TAB. 1 – Évaluation de SXLFG, avec un *time-out* de 15 secondes[9].

		Corpus complet	Phrases valides pour la CFG support	
		Analyse CFG	Analyse CFG	Analyse complète
Données sur les corpus[11]	#phrases	40859	35756	
	n_{moy} / n_{max}	20.95 / 541	19.06 / 173	
	UW_{moy} / UW_{max}	0.79 / 97	0.75 / 65	
Temps d'analyse	moy	0.05s	0.01s	3.35s
	med	0.00s	0.00s	0.03s
	$<$ 0.1s	98.2%	98.8%	57.8%
	$<$ 1s	99.8%	99.9%	71.0%
Nombre d'analyses	max	3.10^{73}	5.10^{52}	1^{12}
	med	32 028	29 582	1
	$\geq 10^6$	36.13%	35.28%	0%
	$\geq 10^{12}$	8.86%	7.84%	0%

TAB. 2 – Données sur le EASy corpus, les temps et les nombres d'analyses, avant application de l'heuristique de sur-segmentation (*time-out* de 15 secondes[9]).

[8]Cette qualité dépend aussi des heuristiques de désambiguïsation utilisées et du traitement de la robustesse.

[9]Un *time-out* plus élevé aurait augmenté les taux de couverture mais également les temps d'analyse.

[10]Nous n'avons pas conservé les informations permettant de donner les résultats sur l'ensemble du corpus. Toutefois, la table 2 donne les temps d'analyse pour les 87.51% de phrases reconnues par la CFG support.

[11]n désigne un nombre de mots, et UW un nombre de mots inconnus.

[12]Dans 14.34% des cas, aucune analyse n'a été trouvée en moins de 15s. C'est dans ces cas-là que sont alors appliquées les heuristiques de sur-segmentation.

5 Conclusion

Dans cet article, nous avons introduit l'analyseur SXLFG. À notre connaissance, c'est la première fois qu'un système d'analyse fondé sur le modèle LFG traite du texte tout venant de façon efficace et robuste sans que le pouvoir expressif du formalisme ne soit dégradé. Il est en outre possible de décrire des phénomènes complexes dans SXLFG en accord avec les nombreux travaux linguistiques qui s'y rapportent.

Les expériences relatées utilisent une grammaire du français et un lexique morpho-syntaxique que nous avons également réalisés. Les résultats extrêmement encourageants obtenus ne doivent bien entendu pas masquer qu'il s'agit d'une première tentative qui doit se poursuivre et qui peut être améliorée. Les perfectionnements possibles concernent le formalisme, la grammaire du français et l'analyseur SXLFG lui-même.

Nous avons quelques idées pour étendre notre variante de LFG qui pourraient faciliter certains traitements, en particulier celui des coordonnées. La grammaire doit être étendue et améliorée. En effet, certaines constructions comme les clivées, les comparatives, les coordinations à ellipse, et d'autres, ne sont pas couvertes. D'autre part, la grammaire support (CFG) doit être affinée car son ambiguïté actuelle est déraisonnable (voir section 4.2). Même si notre analyseur Earley y est relativement peu sensible, elle peut rendre prohibitif le temps d'évaluation des structures fonctionnelles associées. Les autres pistes de recherche sur l'analyseur proprement dit concernent essentiellement l'amélioration de la robustesse et du temps de calcul des structures fonctionnelles.

Références

ANDREWS A. (1990). *Functional closure in LFG.* Rapport interne, The Australian National University.

BOULLIER P. (2003). Guided Earley parsing. In *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT'03)*, p. 43–54, Nancy, France.

BOULLIER P., CLÉMENT L., SAGOT B. & ÉRIC VILLEMONTE DE LA CLERGERIE (2005). Chaînes de traitement syntaxique. In *Actes de TALN 05*, Dourdan, France.

BRIFFAULT X., CHIBOUT K., SABAH G. & VAPILLON J. (1997). An object-oriented linguistic engineering environment using LFG (Lexical-Functional Grammar) and CG (Conceptual Graphs). In *Proceedings of Computational Environments for Grammar Development and Linguistic Engineering, ACL'97 Workshop.*

CLÉMENT L. & KINYON A. (2001). XLFG – an LFG parsing scheme for French. In *Proceedings of LFG'01*, Hong Kong.

KAPLAN R. (1989). The formal architecture of lexical functionnal grammar. *Journal of Informations Science and Engineering.*

KAPLAN R. M. & MAXWELL J. T. (1994). *Grammar Writer's Workbench, Version 2.0.* Rapport interne, Xerox Corporation.

KINYON A. (2000). Are structural principles useful for automatic disambiguation ? In *Proceedings of in COGSCI'00*, Philadelphia, Pennsylvania, United States.

Etiquetage morpho-syntaxique du français à base d'apprentissage supervisé

Julien Bourdaillet, Jean-Gabriel Ganascia
LIP6 - Univesité Paris VI
8 rue du capitaine Scott - 75015 Paris
{Julien.Bourdaillet, Jean-Gabriel.Ganascia}@lip6.fr

Mots-clefs : étiquetage morpho-syntaxique, apprentissage supervisé, modèle de Markov caché, évaluation, homographes

Keywords: part-of-speech tagging, supervised learning, hidden Markov model, evaluation, homographs

Résumé Nous présentons un étiqueteur morpho-syntaxique du français. Celui-ci utilise l'apprentissage supervisé à travers un modèle de Markov caché. Le modèle de langage est appris à partir d'un corpus étiqueté. Nous décrivons son fonctionnement et la méthode d'apprentissage. L'étiqueteur atteint un score de précision de 89 % avec un jeu d'étiquettes très riche. Nous présentons ensuite des résultats détaillés pour chaque classe grammaticale et étudions en particulier la reconnaissance des homographes.

Abstract A french part-of-speech tagger is described. It is based on supervised learning: hidden Markov model and trained using a corpus of tagged text. We describe the way the model is learnt. A 89 % precision rate is achieved with a rich tagset. Detailed results are presented for each grammatical class. We specially pay attention to homographs recognition.

1 Introduction

L'étiquetage morpho-syntaxique consiste à assigner la bonne classe grammaticale, définie suivant un certain niveau de granularité, à chaque mot d'un texte en entrée. Nous présentons ici un étiqueteur (ou tagger) du français basé sur un modèle d'apprentissage supervisé. Celui-ci est une adaptation du tagger de l'analyseur syntaxique RASP de la langue anglaise. Cet étiqueteur apprend un modèle de langage à partir d'un corpus préalablement étiqueté.

Plusieurs approches ont été présentées pour l'étiquetage morpho-syntaxique du français. Le Brill Tagger (Brill, 92) apprend des règles à partir d'un corpus étiqueté et a été adapté pour le français avec WinBrill. (Giguet, 97) et (Chanod, 95) se basent sur des propriétés de la langue comme les mots noyaux ou des méthodes à base de contraintes. (Chanod, 95) présente un autre étiqueteur à base d'apprentissage non-supervisé, qui apprend un modèle de langage à partir d'un corpus non-étiqueté via une variante de l'algorithme Estimation-Maximisation (EM).

(Chanod, 95) présente les limites de ce modèle qui est fortement dépendant des conditions initiales et peut rester bloqué sur un optimum local. (Stein, 95) présente une adapatation au français du TreeTagger. Celui-ci est basé sur un modèle de Markov caché (Hidden Markov Model ou HMM) modélisé par un arbre de décisions. Nous nous situons dans la lignée de ces deux derniers travaux et utilisons également un HMM. Toutefois puisqu'il a été prouvé dans (Elworthy, 94) que l'apprentissage d'un modèle à partir d'un corpus manuellement étiqueté produit de meilleurs résultats que la procédure d'apprentissage d'EM et que nous disposions d'un tel corpus, nous avons choisi cette alternative.

2 Présentation de l'étiqueteur

Nous utilisons l'étiqueteur morpho-syntaxique de l'anglais du projet RASP (Briscoe, 02). Cet étiqueteur, détaillé dans (Elworthy, 94), est basé sur HMM du premier ordre (bigramme). Celui-ci est représenté par un lexique des formes fléchies qui associe à chaque mot ses tags potentiels et par une matrice de transition entre états. Nous l'avons adapté au traitement de la langue française.

Nous avons utilisé le corpus GRACE qui est annoté morpho-syntaxiquement. Il comporte environ 800.000 mots et est constitué pour moitié d'articles du Monde et pour moitié d'oeuvres et d'essais littéraires. Ce corpus est étiqueté avec un jeu de 312 étiquettes qui correspondent à un étiquetage très fin. Les mots sont tout d'abord regroupés en 12 grandes classes très générales : adjectif, conjonction, déterminant, mot-phrase, nom, pronom, adverbe, préposition, verbe, résidu, ponctuation et extra-lexical. Ces classes sont ensuite affinées, comme par exemple conjonction en conjonction de coordination et conjonction de subordination, verbe en verbe auxiliaire et verbe principal. On aboutit ainsi à un jeu de 36 étiquettes. Enfin, ont été adjoints à ces classes grammaticales des traits proprement morphologiques, tels que le genre, le nombre, la personne, le mode ou encore le temps qui donnent le jeu de 312 étiquettes. Nous avons ajouté, pour des facilités d'implémentation et sans dénaturer la cohérence de l'ensemble, huit étiquettes composées (qui existaient dans le corpus sous la forme d'une composition de deux étiquettes) et sommes ainsi arrivés à un jeu de 320 étiquettes.

Nous avons choisi d'effectuer l'apprentissage sur le corpus GRACE avec le jeu de 320 étiquettes. En effet, l'intérêt de ce jeu est que les classes très fines apportent beaucoup d'informations sur les mots et permettent de se passer d'analyseur morphologique en vue d'une étape ultérieure d'analyse en constituants.

Ainsi la procédure d'apprentissage permet d'apprendre un HMM basé sur 320 états. Nous avons gardé la majeure partie du corpus comme données d'apprentissage et utilisé le reste comme données de test, soit environ 26.000 mots. L'entraînement du modèle a permis d'obtenir un dictionnaire d'environ 44.000 formes fléchies.

Nous avons modifié l'algorithme d'étiquetage proprement dit pour que celui-ci prenne en compte les locutions. Pour cela, nous avons appliqué l'heuristique du motif le plus long. Au moment de lire le texte en entrée, l'étiqueteur va chercher, grâce au lexique, si la combinaison du mot suivant au mot courant forme une locution. Si tel est le cas, on applique ce principe itérativement avec le mot suivant jusqu'à ce que l'application ne soit plus possible, auquel cas, on garde la dernière locution trouvée. L'étiquetage de la phrase par Viterbi est ensuite effectué avec celle-ci.

3 Evaluation

Pour évaluer notre travail, nous utilisons la précision (proportion d'étiquetages corrects parmi les étiquetages stricts) et la décision (proportion d'étiquetages stricts parmi l'ensemble de tous les étiquetages). Dans un premier temps, nous avons développé un script Perl chargé de cette évaluation, qui sera appelé par la suite EVAL. Dans un second temps nous avons réutilisé la boîte à outils d'évaluation des analyseurs du projet ELSE[1]. Notons que EVAL comporte 150 lignes de code alors que l'évaluateur ELSE en comporte plusieurs milliers, même si ce dernier se veut plus ambitieux et traite, par exemple, divers formats de fichiers en entrée.

3.1 Résultats

Lors de tous nos tests, notre analyseur atteint un score de décision de 100% et ceci pour deux raisons. Tout d'abord, il ne renvoie qu'une seule étiquette par mot, ce qui ne génère pas d'étiquetage ambigu. Et ensuite, notre jeu d'étiquettes est identique à celui de GRACE. Il n'y a donc pas de problèmes de projection d'un jeu dans l'autre (phénomène qui entraîne des regroupements de plusieurs étiquettes en une seule ou inversement) lors de l'utilisation de l'évaluateur ELSE.

Nous avons effectué des tests avec trois niveaux de granularité de l'étiquetage. Le premier niveau correspond au jeu complet des 320 étiquettes. Le second est ce même jeu mais sans les traits morphologiques, ce qui donne 36 étiquettes. Le dernier niveau est l'étiquetage le plus général en 12 grandes classes grammaticales. Dans le tableau 1 nous présentons les scores de précision dans différents cas. La première colonne correspond à l'étiquetage du corpus de test produit par notre analyseur morpho-syntaxique avec EVAL. La deuxième fait référence au même étiquetage mais avec l'évaluateur ELSE. Et la dernière présente l'évaluation par ELSE de ce même corpus de test mais étiqueté par l'analyseur Cordial. En effet, ce dernier cas nous a semblé intéressant comme élément de comparaison puisque Cordial semble actuellement être le meilleur analyseur du français.

	EVAL	ELSE	ELSE / Cordial
320 tags	87.65 %	89.07 %	94.44 %
36 tags	92.24 %	93.52 %	94.63 %
12 tags	94.39 %	95.69 %	97.52 %

Table 1: Scores de précision

A partir de ces chiffres nous pouvons tirer plusieurs constations. Tout d'abord remarquons qu'entre l'évaluation de notre analyseur avec EVAL et celle avec ELSE, on a pour les trois niveaux environ 1.5 point d'écart. Ceci vient de la différence des méthodes de segmentation utilisées. Avec ELSE, l'alignement entre la sortie de l'étiqueteur et le corpus correctement étiqueté n'est pas très bon car ELSE utilise un segmenteur générique et non un segmenteur adapté à l'analyseur comme EVAL. Ainsi plus de tokens ne sont pas correctement réalignés avec ELSE, comme certains mots composés ou contenant une apostrophe. Ceux-ci ne sont donc pas soumis à évaluation, ce qui tend à améliorer mécaniquement la précision. D'autre part, cela souligne l'importance de la question de la segmentation et la difficulté d'y apporter une réponse qui soit valide à travers plusieurs formalismes. Avec EVAL, les tokens incorrectement réalignés sont des mots contenant des tirets, des mots composés et des expressions non présentes dans le dictionnaire. On en compte environ 0.8 % du nombre de tokens contenus dans le corpus de test.

[1] http://www.limsi.fr/TLP/ELSE

Ensuite, avec le jeu d'étiquettes complet (320 tags) et l'évaluateur ELSE, notre analyseur est moins performant que Cordial. Cependant si l'on compare ce score avec ceux présentés dans (Adda, 99), nous nous situons dans la moyenne des analyseurs évalués qui obtiennent de 82 à 96 % de précision. De plus, notons que lors de cette campagne les meilleurs résultats ont été obtenus non par des étiqueteurs mais par des analyseurs syntaxiques complets. Ceux-ci ayant un avantage certain sur les premiers pour désambiguïser les cas les plus difficiles. En effet, leur résolution peut être reportée au niveau syntaxique, ce qui améliore de quelques (mais précieux) points la précision. Au vu de nos résultats et du fait que nous nous limitons au niveau de l'étiquetage, notre approche est intéressante vis-à-vis des autres.

Enfin, nous avons effectué l'évaluation avec des jeux d'étiquettes plus concis. En effet, ceux-ci se rapprochant plus des jeux utilisés par les analyseurs de l'anglais, la comparaison est rendue possible. Notre score se situe également dans la moyenne de ceux présentés dans la littérature. Notons que dans (Briscoe, 02), donc pour RASP sur l'anglais, les auteurs atteignent une précision de 97 %. Nous nous situons en dessous, ce qui laisse entrevoir des possibilités d'amélioration. Toutefois un tel score semble difficilement atteignable par un étiqueteur seul pour le français. En effet, du fait de sa plus grande richesse morphologique, le français est plus dur à étiqueter que l'anglais. Ceci est confirmé, au vu de la littérature, par le fait que globalement les scores des étiqueteurs de l'anglais sont plus élevés, de quelques (et toujours précieux) points, que ceux du français.

3.2 Evaluation par classes

Nous cherchons dans cette partie à évaluer les points forts et les points faibles de notre étiqueteur de façon plus précise. Dans un premier temps, nous présentons dans cette section les résultats pour chaque classe grammaticale. Dans un second temps, nous présentons dans la section suivante le taux de reconnaissance des homographes.

Dans le tableau 2, nous présentons pour chaque classe grammaticale, le nombre d'occurrences dans le corpus de test, le pourcentage d'étiquetage correct et les trois types d'erreurs les plus fréquentes par ordre décroissant.

	Occurrences	% Correct	Erreur 1	Erreur 2	Erreur 3
Nom	6506	95.2 %	Adj / 2 %	V / 0.8 %	D / 0.7 %
Verbe	3184	96 %	Adj / 2.4 %	N / 1.1 %	Prep / 0.3 %
Verbe auxiliaire	669	78.5 %	Vp / 21 %	N / 0.1 %	Adv / 0.1 %
Verbe principal	2515	92.8 %	Adj / 3 %	Va / 2.3 %	N / 1.3 %
Adjectif	1773	85.4 %	V / 6.6 %	N / 5.1 %	Adv / 1.2 %
Pronom	1456	89.3 %	C / 4.5 %	D / 2 %	Adv / 1.7 %
Déterminant	3162	94.0 %	Prep / 1.8 %	Pron / 1.7 %	N / 1.6 %
Adverbe	1170	94.9 %	C / 2.3 %	Prep / 1 %	N / 0.8 %
Conjonction	856	97.9 %	Adv / 1%	Prep / 1 %	Pron / 0.6 %
Préposition	3863	81.8 %	D / 16.4 %	Pron / 0.6 %	N / 0.3 %
Ponctuation	3424	98.8 %	Adv / 0.8 %	Pron / 0.3 %	N / 0.1 %

Table 2: Scores et types d'erreurs par classes

Nous considérons que les classes présentant un taux supérieur à 94 % sont bien reconnues et que les efforts d'amélioration de l'étiqueteur devraient plutôt se porter sur la reconnaissance

des autre classes. En étudiant en particulier les deux composantes de la classe Verbe (auxiliaire et principal), on constate que les auxiliaires sont les plus mal reconnus. Néanmoins le tagger hésite essentiellement entre Verbe auxiliaire et Verbe principal, ce qui est satisfaisant pour une étape ultérieure d'analyse syntaxique. La ponctuation présente un score élevé mais toutefois insuffisant, ce qui est dû aux points de suspension mal étiquetés.

Les prépositions semblent particulièrement mal reconnues. En analysant les erreurs, on constate que ce sont les mots du type "de la, de l', des" qui posent la plupart des problèmes, c'est-à-dire les mots homographes qui sont soit des articles partitifs, soit des prépositions contractées. Cette difficulté est due à la forte ambiguïté entre les deux cas, élément caractéristique de la grammaire française. De plus, les adjectifs et les pronoms sont un point faible de notre étiqueteur.

3.3 Evaluation des homographes

En nous inspirant de (Vergne, 99) qui présente différents types d'homographes du français, nous avons cherché à affiner ces résultats. Le tableau 3 présente ci-dessous les taux de reconnaissance de différents types d'homographes. Dans la première colonne (H1) sont présentés les homographes déterminant/pronom (le, l', la, les); en H2, les homographes préposition contractée/article partitif (de, d', du, des); en H3, les homographes conjonction/pronom relatif (que, qu'); en H4, les homographes auxiliaire/nom (est, être, avoir); en H5, les homographes adverbe/nom (bien, mal, moins, plus, pas, point...); en H6, les homographes nom/adjectif et en H7, les homographes nom/verbe principal. La première ligne présente le nombre d'occurrences dans le corpus de test de la classe majoritaire (en effet, pour les cas d'homographes, une classe est largement majoritaire par rapport à l'autre, de l'ordre de 70 à 95 % des cas); la seconde, le nombre d'occurrences de la classe minoritaire; la troisième, le taux de reconnaissance de la classe majoritaire, et la dernière le taux de reconnaissance de la classe minoritaire.

	H1	H2	H3	H4	H5	H6	H7
Occ. Maj	D: 1767	Prep: 1601	C: 200	Aux: 267	Adv: 223	Adj: 290	N: 141
Occ. min	Pro: 59	Part: 132	Pro: 47	N: 6	N: 27	N: 171	VP: 96
% OK Maj	97.7 %	82.5 %	98 %	99.2 %	97.3 %	91.7 %	92.9 %
% OK min	69.5 %	62.1 %	17 %	85.7 %	85.2 %	87.7 %	83.3 %

Table 3: Reconnaissance des homographes

Au vu de ces résultats, nous constatons que les classes majoritaires sont bien reconnues (sauf en H2, où on retrouve la difficulté précédente). Pour les classes minoritaires, les résultats sont réellement encourageants pour H4, H5, H6 et H7, or ce sont ces cas qui sont les plus intéressants pour l'étape ultérieure d'analyse syntaxique. En effet, nous pensons que discriminer un nom d'un verbe (H7) apporte plus d'information qu'une préposition d'un partitif (H2), même si cela semble difficilement quantifiable. Nous avons déjà discuté de la difficulté du cas H2. Pour ce qui est de H1 et H3, dans les deux cas, la classe minoritaire est le pronom. Nous pensons atteindre ici certaines limites de l'étiquetage pour ce qui est de la reconnaissance des pronoms. En effet, celle-ci serait probablement plus pertinente au niveau de l'analyse en constituants .

4 Conclusion

A travers ce travail, nous avons cherché à développer un étiqueteur morpho-syntaxique du français se fondant sur une méthode d'apprentissage supervisé. Pour ce faire, en nous basant sur un tagger de l'anglais, nous avons adapté la procédure d'apprentissage aux exigences du traitement automatique du français. Nous obtenons un score de précision de 89 %, score inférieur à ceux des meilleurs étiqueteurs mais comme, d'une part ceux-ci sont des analyseurs syntaxiques complets et d'autre part nous n'avons effectué aucune optimisation sur notre étiqueteur, ce résultat est intéressant. Après avoir analysé en détail les erreurs d'étiquetage, il ressort que les points faibles se situent au niveau des adjectifs, pronoms et prépositions et en particulier sur ceux qui sont homographes avec une autre classe. Toutefois nous avons identifié certains cas d'homographes comme étant les plus intéressants et ceux-ci s'avèrent bien reconnus.

Dans une perspective de poursuite de ce travail, nous avons effectué des tests prospectifs avec le dictionnaire appris à partir de tout le corpus (mais avec les mêmes transitions), la précision s'est améliorée de 1.5 à 2 points pour les trois niveaux. Cela laisse à penser que l'augmentation de la taille du dictionnaire présenterait des gains significatifs de précision et, plus généralement, l'acquisition d'un modèle de langage à partir d'un corpus plus conséquent. D'autre part, nous avons réutilisé tel quel le guesser qui est optimisé pour l'anglais. Puisqu'il repose sur des règles statistiques, il fonctionne également ici mais mériterait d'être étudié plus en détail et adapté.

Remerciements

Nous tenons à remercier John Carroll pour son concours et Emmanuel Giguet pour ses commentaires sur ce travail.

Références

Adda G., Mariani J., Paroubek P., Rajman M., Lecomte J. (1999), Métrique et premiers résultats de l'évaluation GRACE des étiqueteurs morphosyntaxiques pour le français, Actes de *la Sixième Conférence sur le Traitement Automatique des Langues Naturelles.*

Brill E. (1992), A simple rule-based part of speech tagger, Actes de *Third Conference of Applied Natural Language Processing.*

Briscoe T., Carroll J.(2002), Robust accurate statistical annotation of general text. Actes de *Third International Conference on Language Resources and Evaluation.* p. 1499-1504.

Chanod J-P., Tapanainen P. (1995), Tagging French - comparing a statistical and a constraint-based method, Actes de *Seventh Conference of the European Chapter of the ACL.*

Elworthy D. (1994), Does Baum-Welch re-estimation help taggers ?, Actes de *Fourth ACL Conference on Applied NLP.*

Giguet E., Vergne J. (1997), From part of speech tagging to memory-based deep syntactic analysis, Actes de *Fifth Iternational Workshop on Parsing Technologies.*

Stein A., Schmid H. (1995), Etiquetage morphologique de textes français avec un arbre de décisions, *Revue T.A.L*, Vol. 36.

Vergne J. (1999), *Habilitation à diriger des recherches*, p. 36.

TALN 2005, Dourdan, 6-10 juin 2005

Contextes multilingues alignés pour la désambiguïsation sémantique : une étude expérimentale

Boxing Chen, Meriam Haddara, Olivier Kraif (1)
Grégoire Moreau de Montcheuil, Marc El-Bèze (2)

(1) LIDILEM, Université Stendhal Grenoble 3
Meriam.Haddara@laposte.net, Olivier.Kraif@u-grenoble3.fr

(2) LIA – CNRS, Université d'Avignon et des Pays du Vaucluse
{Gregoire.Moreau-de-Montcheuil, Marc.El-Beze}@univ-avignon.fr

Mots-clés : Désambiguïsation sémantique, alignement multilingue, lexique sémantique
Keywords: Word sense disambiguation, multilingual aligning, semantic lexicon

Résumé Cet article s'intéresse à la désambiguïsation sémantique d'unités lexicales alignées à travers un corpus multilingue. Nous appliquons une méthode automatique non supervisée basée sur la comparaison de réseaux sémantiques, et nous dégageons un critère permettant de déterminer *a priori* si 2 unités alignées ont une chance de se désambiguïser mutuellement. Enfin, nous développons une méthode fondée sur un apprentissage à partir de contextes bilingues. En appliquant ce critère afin de déterminer pour quelles unités l'information traductionnelle doit être prise en compte, nous obtenons une amélioration des résultats.

Abstract This paper addresses the sense disambiguation of aligned words through a multilingual corpus. We apply an unsupervised disambiguation method using inter word net comparison. We study a criterion that allows to identify the cases for which disambiguation can take advantage of alignment. Finally, we implement a method based on a training stage using both monolingual and bilingual context, and we apply the previous criterion in order to select between monolingual or bilingual clues, showing some improvement of the results.

1 Introduction

Les corpus multilingues alignés présentent un intérêt particulier pour la désambiguïsation sémantique au niveau lexical. Dans la mesure où les distributions sémantiques des unités lexicales sont différentes d'une langue à l'autre, les unités alignées à travers des textes parallèles peuvent jouer un rôle réciproque de révélateur sémantique (Ide *et al.*, 2002). Dès les premiers développements des corpus multilingues alignés, il y a une quinzaine d'années, des méthodes de désambiguïsation multilingue ont été proposées (Gale *et al.* 1992, Dagan & Itai, 1994). Le principe est simple : les résultats obtenus grâce aux méthodes de désambiguïsation monolingue, qui se basent en général sur un examen du contexte proche du mot cible, devraient *a priori* être améliorés par l'apport d'information pertinente consistant en la prise en

compte des unités et contextes alignés. Fondé sur cette hypothèse de travail, le projet Carmel, financé par le réseau RIAM[1] vise au développement d'un corpus multilingue aligné concernant quatre langues européennes : l'anglais, l'espagnol, le français et l'italien. Outre son intérêt culturel (récits de voyage s'échelonnant de la fin du XIX^{e} au début du XX^{e} siècle), ce corpus quadrilingue doit nous permettre de coupler trois types de traitement : désambiguïsation sémantique, identification thématique et alignement. Nous décrivons dans un premier temps les résultats de quelques expériences de désambiguïsation non supervisée. Les sections 3 et 4 sont consacrées à la description d'une méthode avec apprentissage, basée sur un mélange d'indices monolingues et bilingues où nous montrons comment la triangulation des méthodes et des langues permet d'améliorer et de consolider les résultats.

2 Désambiguïsation multilingue non supervisée

Un des problèmes posés par les méthodes classiques de désambiguïsation sémantique est la nécessité de passer par une phase d'apprentissage, nécessitant le recours à des corpus étiquetés manuellement dont la production est longue et coûteuse. Récemment, des travaux originaux (Diab & Resnik, 2002, Tufis *et al.*, 2004) ont montré qu'il était possible d'utiliser des corpus alignés pour effectuer une désambiguïsation *non supervisée*. Comme Tufis *et al.* (2004), nous avons utilisé 2 réseaux sémantiques (les lexiques français et anglais livrés avec EuroWordNet), afin de comparer les unités par le biais d'index interlingue (ILI), qui permettent d'établir des équivalences de sens entre des unités. La désambiguïsation de 2 unités U_s et U_c est alors effectuée en cherchant le couple de sens (s_s,s_c) qui maximise une certaine mesure de similarité $Sim(s_s,s_c)$: $D(U_s,U_c) = \text{argmax\_}\{(s_s,s_c) \in S_s \times S_c\}\ Sim(s_s,s_c)$
$Sim(s_s,s_c)$ peut être calculée à partir du nombre de liens séparant chacun des ILI de leur plus proche parent commun dans la hiérarchie. Dans l'expérience décrite ci-après, afin de privilégier la précision (au détriment du rappel) et de s'approcher de la désambiguïsation manuelle effectuée précédemment, nous avons utilisé une définition maximaliste de la similarité, basée sur l'identité des ILI.

La première étape consiste à aligner automatiquement les textes. À partir des alignements phrastiques, nous avons procédé à l'extraction des correspondances lexicales, en utilisant une combinaison d'indices : fréquence des occurrences et des cooccurrences au sein des phrases alignées, positions dans les phrases, ressemblance graphique, identité des parties du discours. Ces indices, combinés de façon appropriée, nous ont permis d'obtenir une F-mesure voisine de 90% sur un corpus étiqueté (*Madame Bovary*, de Flaubert, et sa traduction en anglais) pour l'appariement des mots pleins (Kraif & Chen, 2004). Le corpus étudié est *The voyage of the Beagle* de Darwin, comportant environ 200 000 mots dans chaque langue, préalablement segmenté et étiqueté avec les parties du discours. Pour qu'un couple d'unités soit partiellement ou complètement désambiguïsé, il faut que les 2 unités alignées apparaissent chacune dans leurs réseaux respectifs. Seulement 21 133 couples ont satisfait cette condition.

Les couples totalement désambiguïsés représentent environ 4,3% de la totalité des mots, et 42% des couples pour lesquels les 2 réseaux n'étaient pas silencieux. Pour estimer la précision

[1] Le réseau RIAM, créé en 2001, dépend du Centre national de la cinématographie. Les partenaires du Projet sont l'association ACCE, la société SINEQUA et les laboratoires LIA (Avignon) et LIDILEM (Grenoble).

des résultats, nous avons soumis à l'évaluation manuelle d'un seul annotateur 100 couples totalement désambiguïsés, prélevés aléatoirement. Seul l'anglais a pour le moment été évalué. Ces résultats doivent être pris avec précaution du fait de l'incomplétude et du déséquilibre des réseaux (le réseau français contient 22 745 sens contre 91 600 pour Wordnet 1.5). Notons néanmoins que la précision est plutôt bonne pour les couples totalement désambiguïsés.

	Anglais	Français
Proportion moyenne de sens éliminés	63 %	46 %
Unités totalement désambiguïsées	34,6 % (7 316 / 21 133)	22,7 % (4 804 / 21 133)
Précision estimée	79 %	

Tableau 1 : Réduction des sens pour la désambiguïsation automatique

Nos observations manuelles indiquent que la désambiguïsation multilingue fonctionne très bien pour certains couples d'unités, pas du tout pour d'autres (quand les différentes acceptions sont trop similaires). Pour départager automatiquement les cas favorables des autres, on peut se tourner vers une troisième langue, qui fournira vraisemblablement des indices quant à cette similarité. Par exemple, pour (en-*scarcely*, fr-*presque*) si l'on considère les unités espagnoles alignées avec chaque mot du couple, indépendamment, dans notre corpus, on trouve :

en-*scarcely* -> es-*tampoco* es-*asegurar* es-*casi* es-*apenas* fr-*presque* -> es-*casi*

Un simple filtrage des fréquences nous permet d'éliminer les alignements erronés tel que es-*asegurar*. On constate alors que les 3 sens de *scarcely* indiqués par le dictionnaire, que l'on pourrait gloser par /presque pas/, /difficilement/, /à peine (sens temporel)/, se manifestent par des équivalents espagnols plus variés. La différence sémantique entre *presque* et *scarcely*, qui aboutit en fait à une désambiguïsation correcte de l'anglais, est donc rendue manifeste par cette projection dans une tierce langue. Pour prédire s'il est judicieux ou non d'employer, pour un couple donné, la désambiguïsation multilingue, nous proposons de recourir à un critère numérique, comme l'indice de DICE : $s = \frac{2 \cdot |ES(e) \cap ES(f)|}{|ES(e)| + |ES(f)|}$, où *ES(e)* et *ES(f)* représentent les ensembles d'équivalents espagnols dérivés de l'alignement pour les unités *e* et *f*. Calculé sur un corpus trilingue suffisamment important, *s* peut être un bon indicateur de la similarité sémantique de 2 unités : une valeur faible devrait indiquer de meilleures chances de désambiguïsation multilingue. Cette hypothèse semble confirmée par une certaine corrélation entre la similarité *s* obtenue par projection sur l'espagnol et la proportion de sens éliminés.

% sens éliminés	$0 \leq s < 0{,}25$	$0{,}25 \leq s < 0{,}5$	$0{,}5 \leq s < 0{,}75$	$0{,}75 \leq s \leq 1$
Anglais	75 %	65 %	62 %	60 %
Français	60 %	49 %	43 %	40 %

Tableau 2 : Corrélation entre *s* et la proportion des sens éliminés

3 Méthode de désambiguïsation supervisée

Puisque l'alignement des unités fournit, dans certains cas, une information exploitable pour la désambiguïsation, il paraît naturel d'intégrer cette information dans une méthode "classique", basée sur l'observation des contextes (de Loupy, 2000). Dans cette nouvelle tâche, nous disposons d'un ensemble d'exemples d'apprentissage, noté *Tr*, dont chaque élément représente un (ou plusieurs) sens pour un mot ambigu ; et nous recherchons pour un contexte particulier, noté C^0, les sens correspondants. La méthode employée est décrite en section 3.2.

3.1 Prétraitements

Chaque mot du contexte est lemmatisé, à l'exception du mot à désambiguïser, puis une série de réductions est effectuée : remplacement des nombres par CD, des mois par MONTH et des jours de la semaine par DAY, filtrage pour ne garder que les substantifs, adjectifs, verbes, adverbes, noms propres, prépositions et nombres. Parmi les différentes possibilités de contextes alignés, nous avons choisi de travailler sur la traduction *mot à mot* du contexte source : les termes alignés avec chacun des mots du contexte d'origine. Les contextes projetés (cf. Tableau 3) conservent l'ordre des mots du voisinage de la langue source. L'alignement vers une langue est total, même si quelques mots n'ont pas d'image.

En	strew:Verb	with:Prep	fine:Adj	sand:Noun	street:Noun	3
Pfr	poudrer:Verb	de:Prep	fin:Adj	sable:Noun	Rue:Noun	3

Tableau 3 - Contexte "projeté"

3.2 Algorithme des K plus proches voisins (KPPV)

L'algorithme des KPPV consiste à rechercher parmi les données d'apprentissage les k exemples qui ressemblent le plus au contexte C^0. Nous utilisons pour cela une mesure de similarité entre 2 contextes, *Simil(C,C')* qui permet de classer les différents exemples d'apprentissage et de ne retenir que les k plus proches (ensemble *KNN*) Chacun des exemples de l'ensemble des *KNN* vote pour le (ou les) sens qu'il caractérise, proportionnellement à sa similarité avec C^0, ce qui donne pour chaque sens s, le score : $Sc_{KPPV}(s) = p(s) + \sum_{C \in KNN \cap Tr(s)} Simil(C, C^0)$, où *p(s)* est la probabilité du sens s et *Tr(s)* est l'ensemble des exemples d'apprentissage pour le sens s. Pour comparer 2 contextes dans une même langue, nous avons décidé de comptabiliser le nombre de termes identiques placés à la même position (par rapport au mot ambigu) ou légèrement décalés (pour prendre en compte des phénomènes de décalage comme l'incise d'un adverbe). Si on note $C=(w_i)_{-g \le i \le +d}$ et $C'=(w'_i)_{-g' \le i \le +d'}$ deux contextes « centrés » autour d'un lemme ambigu (w_0 et w'_0 sont les 2 occurrences de ce lemme), la formule de calcul de similarité entre C et C' s'écrit :

$$Simil(C,C') = \left(\sum_{i=-G}^{+D} f_i \right)^2 \text{ avec } f_i = \begin{cases} 1, \text{ si } w_i \equiv w'_i \\ \frac{1}{2}, \text{ si } w_i \neq w'_i \wedge (w_i \equiv w'_{i-1} \vee w_i \equiv w'_{i+1}) \\ 0, \text{ sinon} \end{cases} \text{ et } \begin{matrix} G = \min(g, g') \\ D = \min(d, d') \end{matrix} .$$

Enfin, pour comparer 2 contextes multilingues, nous avons choisi d'additionner les similarités monolingues (norme 1). Soit, si les contextes sont $C=\{C_l\}_{l \in Lang}$ et $C'=\{C'_l\}_{l \in Lang}$, la similarité multilingue est : $Simil_{Mu}(C,C') = \sum_{l \in Lang} Simil(C_l, C'_l)$.

4 Expériences et stratégie de décision

Dans cette série d'expériences, pour 15 lemmes anglais (6 verbes, 5 substantifs et 4 adjectifs), nous disposions de 2 886 contextes désambiguïsés manuellement extraits de 14 récits de voyage d'auteurs anglophones et francophones du XIX° siècle. Par un tirage au sort respectant la distribution des sens, nous en avons écarté environ 80% pour l'apprentissage, ce qui fait un

jeu de 659 tests. Dans l'apprentissage comme dans le test, pour plus de 95% des contextes nous disposons de l'alignement anglais-français.

Lemme	#App	#Test	en	%	en-Pfr	%	#Couple	%	Mel.Fin(0,7)	%	Opt.Fin	%
Total	2227	659	410	62,2	411	62,4	229	35	421	63,9	435	66,0

Tableau 4 - Résultats monolingues, bilingues et avec sélection fine

La première observation qu'il convient de faire, sur les expériences monolingues, est que le résultat global (62,2%) cache une disparité de comportement entre les 15 termes retenus. Il n'est pas étonnant, au regard de la littérature, de voir les noms (maximum atteint par *lady* à 87,9%) à un meilleur niveau de performance que les verbes (minimum atteint par *strike* à 38,3%). Le nombre de sens (21) de *strike* rapporté à une taille de corpus d'apprentissage assez faible (147 exemples) explique en partie le mauvais résultat obtenu par ce verbe.

Comme le montre la colonne *en-Pfr* du tableau 4, le recours systématique au français entraîne une toute petite amélioration des performances au niveau global (1 test). En fait, il n'est pas possible de faire état d'un apport quelconque à mettre au crédit de l'alignement, la différence n'étant pas suffisante pour pouvoir statuer quoi que ce soit. En outre, un examen détaillé des résultats montre que, pour certains mots, non seulement il n'y a pas de gain, mais il y a des pertes. C'est le cas de *carry, child, curious, nature, strike* et *use*. Ce comportement hétérogène ouvre la perspective d'une amélioration plus franche.

Il est naturel d'imaginer une sélection permettant de décider, en fonction de la traduction du mot à étiqueter, s'il convient ou non de recourir à l'alignement. Dans cette expérience nous avons fait une sélection fine, test par test, en fonction du coefficient de Dice du couple anglais-français. Cependant, cela n'est possible que dans environ un tiers des cas (colonne *#Couple* du Tableau 4). Dans les autres cas, nous utilisons la moyenne pondérée des coefficients de Dice pour les différents couples du mot. Dans l'idéal, un grain fin aurait pu permettre d'atteindre un résultat de 66%. Force est de constater que nous en sommes loin avec 63,9%. Ceci s'explique par le fait que la décision fine n'est prise qu'une fois sur trois. Cela constitue cependant une progression par rapport à l'usage systématique de l'alignement et laisse la porte ouverte à de nouvelles améliorations.

5 Conclusion et perspectives

Une étude manuelle a permis de dégager le potentiel important du recours aux unités alignées pour la désambiguïsation lexicale. Ainsi, environ 30% des noms ont pu être désambiguïsés totalement, sans que l'annotateur ne détecte de sens incorrect par rapport au contexte. Pour exploiter au mieux cette information par des méthodes automatisées, nous avons étudié 2 approches différentes. La première, basée sur la comparaison de lexiques sémantiques, a confirmé qu'il était possible d'obtenir une désambiguïsation complète avec une très bonne précision, mais pour un nombre faible d'unités (dépendant de la complétude des lexiques en question) ; la deuxième consistait à intégrer le contexte aligné comme si c'était un élément du contexte monolingue, et à appliquer une méthode de WSD supervisée basée sur les KPPV. De cette manière, nous avons observé que les contextes alignés jouaient un rôle parfois positif, parfois négatif, les traductions inopérantes semblant dans certains cas bruiter des indices

monolingues plus pertinents. Sur la base de nos observations préliminaires, nous avons dégagé un critère permettant d'estimer grossièrement la pertinence du contexte aligné : pour 2 unités, on peut comparer les équivalents proposés par l'alignement avec une tierce langue. Ainsi, 2 unités aux différents sens voisins, *a priori* peu intéressantes pour la désambiguïsation mutuelle, devraient avoir des équivalents similaires dans la langue tierce. Ce critère, appliqué au choix, selon les unités, entre WSD monolingue stricte et WSD mono et bilingue, a permis d'obtenir une légère amélioration des résultats. Nous pensons que ce critère de triangulation, s'il implique plus que 3 langues, et un plus grand volume de textes alignés peut permettre un gain important sur le plan de la précision, et ouvrir la voie à la constitution automatique de corpus étiqueté de bonne qualité, pouvant servir de corpus d'apprentissage peu coûteux, pour des méthodes monolingues sur des corpus non-alignés.

Remerciements

Nous remercions le réseau RIAM, qui finance le projet Carmel ainsi que nos partenaires ACCE et la société SINEQUA.

Références

DAGAN I., ITAI A. (1994), Word Sense Disambiguation Using a Second Language Monolingual Corpus. *Computational Linguistics*, 20(4):563-596.

DE LOUPY CL. (2000), Évaluation de l'Apport de Connaissances Linguistiques en Désambiguïsation Sémantique et Recherche Documentaire. Thèse de Doctorat, Université d'Avignon et des Pays du Vaucluse.

DIAB M., RESNIK P. (2002), An Unsupervised Method for Word Sense Tagging using Parallel Corpora, *Proc. of 40th Anniversary Meeting of the Association for Computational Linguistics* (ACL-02), Philadelphia.

DIAB M. (2003), Word Sense Disambiguation within a Multilingual Framework. Ph.D. thesis,University of Maryland.

GALE W. A., CHURCH K. W., YAROWSKY D. (1992). Using Bilingual Materials to Develop Word Sense Disambiguation Methods. In *Proc.of the Fourth Inter-national Conference on Theoretical and Methodological Issues in Machine Translation*, pages 101-112, Montreal.

IDE N., ERJAVEC T., TUFIS D. (2002). Sense Discrimination with Parallel Corpora. *Proceedings of ACL'02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Philadelphia, 54-60.

KRAIF O., BOXING, CHEN (2004) Combining clues for lexical level aligning using the Null hypothesis approach, in *Proceedings of Coling 2004*, Geneva, August 2004, pp. 1261-1264.

TUFIS D., ION R., IDE N. (2004), Fine-Grained Word Sense Disambiguation Based on Parallel Corpora, Word Alignment, Word Clustering and Aligned Wordnets. In *Proceedings of the 20th International Conference on Computational Linguistics, COLING2004*, Geneva.

Naviguer dans les textes pour apprendre

Javier Couto(1), Lita Lundquist (2), Jean-Luc Minel(1)

(1)LaLICC – CNRS, Université Paris-Sorbonne
Maison de la Recherche, 28 rue Serpente, 75 006 Paris
{Prenom.nom}@paris4.sorbonne.fr
(2) Département de français Institut F.I.R.S.T
Handelshøjskolen i København, CBS, Denmark
ll.first@cbs.dk

Mots-clés : Navigation textuelle, apprentissage en lingusitique textuelle
Keywords: Textual navigation, teaching texts and text linguistics

Résumé Dans cet article nous présentons un langage de navigation textuelle et son implantation dans la plate-forme Navitexte. Nous décrivons une application de ces principes de navigation dans un cadre d'apprentissage de la bonne formation des textes, destinée à des dans un cadre d'apprentissage de la bonne formation des textes, destinée à des étudiants apprenant le français langue étrangère.

Abstract In this article we present a declarative language, which models ways of visualising and navigating in texts, together with its implementation in a workstation, *NaviTexte*. First, we show how the language can be used to build an application in text linguistics. This application aims to teach foreign language students to identify different coherence creating units in a text and to navigate between them. Then we detail the declarative language.

1 Introduction

L'activité consistant à sélectionner des informations à l'intérieur d'un texte donné est une démarche qui mobilise de multiples fonctions cognitives et qui varie en fonction des sujets, de leur expertise du domaine traité ainsi que de leur degré d'attention au moment où ils en prennent connaissance. Si la modélisation de ces fonctions n'est pas actuellement possible, du moins dans toute leur complexité, en revanche, un certain nombre de travaux en linguistique textuelle nous fournissent des concepts et des modèles sur lesquels il est possible de s'appuyer, notamment dans un contexte pédagogique. En dépassant le palier de la phrase, la linguistique textuelle a en effet prouvé sa portée dans l'enseignement, que ce soit dans l'enseignement de la langue maternelle, dans l'enseignement des langues étrangères dans l'enseignement de la littérature, dans l'enseignement des langues de spécialité et de la traduction [Lundquist 1999], pour ne mentionner que quelques disciplines pour lesquelles la linguistique textuelle présente un intérêt pédagogique. De plus, la linguistique textuelle permet de focaliser l'attention de l'apprenant sur des structures textuelles et des réalisations linguistiques de la cohérence qui sont propres à des types de textes particuliers (texte narratif, texte argumentatif, etc.), voire propres à des langues différentes. En effet, il s'avère, en adoptant une perspective linguistique contrastive, que les textes ne s'organisent pas de la même manière, même dans des langues apparemment proches, comme le français et l'anglais, ou comme le français et le danois. Ainsi, les langues romanes réalisent souvent la reprise anaphorique par des SNs pleins et variés (les soi-disant « anaphores infidèles »), [Lundquist à paraître], comparées aux langues germaniques comme le danois, qui préfèrent l'anaphorisation pronominale ; un signe parmi d'autres de stratégies totalement différentes pour présenter et organiser l'information dans les textes.

Si la cohérence textuelle constitue bien une « heuristique générale » d'interprétation textuelle [Charolles 1981], il est possible de l'exploiter pour montrer, et enseigner, comment la cohérence se manifeste linguistiquement dans les textes en général, et dans des types de textes différents en particulier, de même que dans des langues différentes. Il nous semble donc que l'enseignement de la linguistique textuelle contribue, dans de multiples contextes, à aiguiser l'attention des apprenants vers la réalisation de la « bonne formation » des textes, et à stimuler leur propre production de textes bien formés. Un outil qui permette à l'élève/l'étudiant de **voir** et de **naviguer** dans le texte entre des unités textuelles assurant la cohérence, sera un instrument didactique très performant et motivant, tant pour l'apprentissage de la lecture que pour l'apprentissage de la production écrite de textes. Dans cette perspective, les dispositifs de visualisation et de navigation sont considérés comme essentiels. Par conséquent une plate-forme logicielle qui permette d'une part, de spécifier des connaissances de visualisation et de navigation, et d'autre part de les exécuter dans un environnement interactif, constitue un dispositif d'apprentissage extrêmement utile.

2 Représentation des unités textuelles nécessaires à la situation d'apprentissage

En étudiant le procédé, par lequel le lecteur apprend à naviguer dans un texte en suivant ses différentes pistes de cohérence, basées sur la référence, sur la prédication et sur les connecteurs – nous attaquons des problèmes cognitifs cruciaux pour lire, comprendre et interpréter correctement un texte, ainsi que pour apprendre par les textes. Le premier problème consiste à identifier les référents discursifs d'un texte et d'établir les relations correctes entre les SN qui y réfèrent. En d'autres mots, de décider s'il s'agit d'une relation de

coréférence, cas où il y a identité et donc question d'un même référent ou s'il y a disjonction référentielle, c'est-à-dire question de deux (ou plusieurs) référents différents. Cette compétence cognitive est primordiale pour arriver à établir une représentation mentale cohérente et correcte du texte en question, qui est, de son côté à la base de toute compréhension par les textes : « learning from text requires that the learner construct a coherent mental representation of the text » [Kintsch 1998: 307] ; voir aussi la question de la *Anaphora resolution*, ibid. : 144 ss).

Le second problème cognitif consiste à identifier le « où veut en venir l'émetteur » du texte. Cette orientation – expressive, argumentative, et d'autre – a été qualifiée de « programme d'interprétation » par [Lundquist 1990], étant donné qu'il s'agit d'une orientation marquée dès le début du texte, qui agit tel un « programme » qui fonctionne *du général au particulier*, et qui permet d'identifier des marques suivantes dans le texte qui « vont dans le même sens », c'est-à-dire *du spécifique au générique*, (voir *macrostructure* et *microstructure*, [Kintsch 1998: 50 ss.]. Cette identification de l'orientation, apportée entre autres par les prédications, est primordiale pour un déchiffrage correct de la cohérence sémantique et pragmatique du texte.

Pour naviguer dans l'objet texte, nous isolons des unités textuelles qui permettent de spécifier des opérations de navigation, ce qui équivaut à établir des liens de cohérence entre des unités de même nature. Comme les éléments textuels appartiennent à des types différents, la navigation permet d'une part de suivre des pistes de cohérence différentes dans un même texte, et d'autre part d'en identifier les réalisations linguistiques dans une langue donnée (ici et pour le moment, le français). Plutôt que de manipuler des structures textuelles hiérarchiques [Couto, Minel 2004], nous distinguons ici des pistes parallèles de marques textuelles qui chacune contribue à un type particulier de cohérence. Ces types de cohérence sont fondés, *grosso modo*, sur les principes exposés dans [Lundquist 1980], selon lesquels on peut distinguer dans les textes une cohérence référentielle, une cohérence prédicative et une cohérence pragmatique, fondée respectivement sur les trois actes de langage : la référence, la prédication et l'illocution qui entrent dans l'énonciation de chaque phrase [Searle 1969]. En fait, chaque phrase contient, en règle générale, un ou des SN qui réfèrent à des entités extra-textuelles et instaurent des référents discursifs ; une ou plusieurs prédications qui d'une part prédiquent des propriétés et établissent des relations sémantiques entre les référents discursifs et d'autre part peuvent porter des traces énonciatives, indicateurs avec d'autres, comme par exemple les connecteurs, de l'acte d'illocution : décrire, convaincre, etc., et de la cohérence pragmatique.

A partir de ces unités, il importe d'identifier les pistes de cohérence qui sont propres au texte en question. Ainsi, dans l'exemple 1, le titre *L'amnistie fiscale n'est pas immorale !* installe un programme d'interprétation du type argumentatif, avec deux voix, celle du protagoniste et celle de l'antagoniste : *L'amnistie fiscale est immorale.* Ces deux voix ouvrent deux pistes de cohérence, qui sont à identifier et à relier de par le texte par les procédés de navigation. Elles sont identifiables, entre autres, par des marqueurs argumentatifs, tels que les modes (l'indicatif et le conditionnel) et les connecteurs *(il est vrai, mais faut-il pour autant…).*

Nous allons illustrer l'exploitation de ces marques sur le texte suivant :

EXEMPLE (1) : « **L'amnistie fiscale n'est pas immorale !** ***PAR FLORIN AFTALION***

« Le gouvernement de Jean-Pierre Raffarin étudie l'opportunité d'instituer une taxe sur les fonds placés à l'étranger et rapatriés en France. Le produit de cette taxe financerait le plan

que vient de dévoiler son ministre de la Cohésion sociale. De nombreux responsables politiques ont manifesté leur hostilité à une telle mesure. Elle serait inefficace et, surtout, immorale car elle blanchirait les «criminels en col blanc». [...] Une telle somme ne contribuerait, il est vrai, que modestement au financement du plan de cohésion sociale dont le coût, en cinq ans, se monterait à 13 milliards d'euros. Mais faut-il pour autant déclarer la mesure envisagée par M. Raffarin inefficace ? Non, car d'autres avantages en résulteraient. Des capitaux importés, qu'ils soient prêtés ou investis, créent des emplois. » [Le Figaro, le 16 juillet 2004].

De même, il est possible de naviguer entre des unités fondées sur des référents discursifs, tel *la taxe*, qui est reprise par des anaphores du type *cette taxe*, *une telle mesure*, *elle*, *la mesure envisagée par M. Raffarin*, des types de reprises anaphoriques qui prêtent amplement à des commentaires linguistiques contrastives.

3 Modélisation et description des connaissances de visualisation et de navigation

Avant de décrire le langage qui permet d'encoder les connaissances de visualisation et de navigation, il convient de rappeler brièvement le modèle de description du texte utilisé. En effet, le texte traité par *NaviTexte* doit être préalablement annoté[1] ce qui implique un choix dans les unités textuelles susceptibles d'être annotées, et par conséquent un modèle de description du texte le plus flexible possible. Notre associe une description hiérarchique, il est possibilité d'emboîter des unités textuelles de type différent, et une description ouverte grâce aux mécanismes prédéfinis de création de nouveaux éléments. Il faut néanmoins souligner que la sémantique de ces relations est implicite, ce qui implique qu'elle doit être prise en compte par le concepteur des modules de navigation. La représentation du texte se décrit dans un format standard XML et se divise en deux parties : le *Corps*, où les unités textuelles, significatives pour la tâche sont délimitées, et la *Tête*, où s'expriment les relations non hiérarchiques entre ces mêmes unités. Dans le *Corps*, l'élément de base de notre modèle est l'*Unité Textuelle* (UT) typée, ce qui permet d'incorporer de nouveaux éléments textuels de manière simple. Une UT peut avoir un titre associé et un nombre non limité d'attributs typés et valués. Les unités textuelles les plus internes possèdent un élément nommé *chaîne* qui représente la chaîne de caractères typographiques. Dans la *Tête*, il est possible de définir de nouveaux éléments composés d'unités textuelles du *Corps* du texte en les associant avec sous la forme d'un *Ensemble*, d'une *Séquence*, d'une *Reference* ou d'un *Graphe*. Un *Ensemble* définit un ensemble d'UT ; Une Séquence définit un ensemble ordonné d'UT ; Une *Reference* définit une relation entre une UT et une séquence finie d'UT ; un *Graphe*, comme dans la TEI, permet de relier des unités textuelles, les nœuds du graphe, par des arcs étiquetés. Enfn, un nouvel élément décrit dans la Tête peut posséder des attributs propres.

[1] Le processus d'annotation du texte peut être manuel ou réalisé par une plate-forme dédiée comme ContextO [Minel & al. 2001], ou Lingstream [Bilhaut & al. 2003].

3.1 Présentation du langage de description des connaissances

Ce langage a pour finalité d'offrir des fonctionnalités à la fois suffisamment génériques tout en proposant une sémantique qui se focalise sur l'essentiel du processus de visualisation et de navigation, à l'inverse de langages de transformation ou de programmation comme, par exemple, XSLT ou XPATH. Notre langage est donc de type déclaratif et s'appuie sur des opérations prédéfinies, mais qui pourront être enrichies en fonction du développement de la plate-forme. Techniquement, l'ensemble des connaissances est décrit dans des modules sous la forme d'expressions symboliques qui doivent respecter une DTD XML. Ces modules sont propres à chaque tâche et correspondent à des pratiques de lecture spécifiques qui peuvent être différentes pour un même texte.

La navigation est conceptualisée comme une opération qui relie une unité textuelle source avec une unité textuelle cible. Une opération de navigation comprend une *source*, une *cible*, une ou plusieurs *conditions*, et un *empan*. L'exécution d'une opération est ainsi soumise à des conditions qui contraignent les attributs des unités textuelles considérées. L'*empan* de texte peut être spécifié de manière à restreindre l'espace de recherche des unités textuelles cibles. Chaque opération est typée avec une valeur qui appartient à l'ensemble {*Premier, Dernier, Suivant[i], Précédent[i]*}. Ces valeurs spécifient d'une part l'orientation, c'est-à-dire dans quel sens (avant ou après l'unité textuelle source) doit être effectué la recherche de l'unité textuelle cible, et d'autre part le référentiel, absolu (*Premier, Dernier),* ou relatif (*Suivant[i], Précédent[i]*), par rapport à la source. Dans le cas d'un référencement relatif, l'index *i* permet de spécifier le rang de la cible recherchée. Les conditions expriment des contraintes sur les valeurs des attributs des unités textuelles source et/ou cible. Ces conditions simples peuvent être combinées entre elles avec les opérateurs logiques (ET, OU, NON). En ce qui concerne *l'empan*, les valeurs possibles doivent être choisies parmi les valeurs déclarées dans la partie *Corps* du modèle du texte (cf. 3). Formellement, la syntaxe générique d'une opération de navigation est la suivante : Opération_Navigation : (Texte, Conditions, Type, Empan) → Unité Textuelle

4 Construction des parcours de lecture

Une première version de la plate-forme *Navitexte*, développée en Java, est opérationnelle [Couto & Minel 2004] et son utilisation pour l'apprentissage, par des étudiants étrangers, des manifestations langagières de la cohérence textuelle constitue une expérimentation inédite et une mise à l'épreuve des hypothèses fondatrices [Minel 2002]. Cette expérimentation nous conduit par ailleurs à réfléchir à une méthodologie d'acquisition, d'organisation et d'exploitation des connaissances mises en œuvre dans la navigation textuelle. En effet, les nombreuses expériences réalisées avec l'hypertexte ont mis en évidence le phénomène de désorientation cognitive [Cotte 2004] engendré par une trop grande richesse dans le graphe des lectures possibles. Comme le montre l'expérimentation décrite précédemment, le concepteur d'un module doit faire face à deux difficultés. Premièrement, les parcours de navigation qu'il conçoit sont en général construits à partir d'exemples de textes annotés, alors que les modules vont s'appliquer sur toute une « famille » de textes qui présentent le même type d'annotations. Il s'agit alors pour le concepteur d'anticiper les effets que peuvent impliquer l'utilisation de certaines variantes par un scripteur qui ne se conformerait pas aux standards, plus ou moins normatifs, des pratiques d'écriture. Dans le cas de l'apprentissage, ce problème peut être résolu par un contrôle des textes proposés, mais cette situation est spécifique et nous visons une utilisation de *Navitexte* beaucoup plus flexible.

La deuxième difficulté découle de notre conceptualisation d'une opération de navigation qui est une relation, de type 1-1, entre une source et une cible. Ce qui implique qu'une unité cible pour une opération Op_1 peut être une source pour une opération Op_p et qu'une plusieurs opérations Op_i peuvent avoir la même source. Un module de navigation construit ainsi un graphe de parcours orienté entre des unités textuelles, et c'est la maîtrise de ce graphe, dans son aspect combinaoire qu'il convient de maitriser. Notre solution repose sur le développement d'un environnement d'aide à la conception qui propose au concepteur des aides visuelles à la fois génériques, indépendantes d'un texte, et spécifiques, par l'exécution d'un module sur un texte éventuellement fabriqué pour la validation[2].

5 Conclusion

Nous avons présenté une expérimentation d'aide à l'apprentissage des procédés de marquage de la cohérence textuelle qui s'appuie sur l'utilisation d'un outil de navigation dans les textes, *Navitexte*. Concue à l'origine à partir des résultats obtenus dans le cadre de travaux menés dans le domaine du filtrage sémantique [Minel & al. 2001] et de la segmentation thématique [Couto & al. 2004], les concepts de navigation et de visualisation proposés démontrent ainsi leur plasticité et leur généricité.

Références

BILHAUT, F., HO-DAC, M., BORILLO, A., CHARNOIS, T., ENJALBERT, P., LE DRAOULEC, A., MATHET, Y., MIGUET, H., PERY-WOODLEY, M .P., SARDA, L., (2003), « Indexation discursive pour la navigation intradocumentaire : cadres temporels et spatiaux dans l'information géographique », *Actes de TALN*, Batz-sur Mer, , p. 315-320.

CHAROLLES, M. (1981), «Coherence as a principle in the interpretation of discourse ». *Text* 3, p. 71-99.

CHAROLLES, M. (1997), « L'encadrement du discours : univers, champs, domaines et espaces », *Cahier de Recherche Linguistique, LANDISCO*, Université Nancy 2, p. 1-73.

COTTE, D. (2004), « Leurres, ruses, désorientation dans les écrits de réseau : la métis à l'écran. », *Communication & langages*, n° 139, Avril 2004, p. 63-74.

COUTO, J., FERRET, O., GRAU, B., HERNANDEZ, N., JACKIEWICZ, A., MINEL, J.-L., PORHIEL, S. (2004), « RÉGAL, un système pour la visualisation sélective de documents. », *Revue d'Intelligence Artificielle*, Hermès, p. 481-514.

COUTO, J., MINEL, J.-L. (2004), « Outils dynamiques de fouilles textuelles », *Actes de RIAO,* Avignon, p. 420-430.

KINTSCH, W. (1998), *Comprehension. A Paradigm for Cognition*, Cambridge, Cambridge University Press.

LUNDQUIST, L. (1980), *La cohérence textuelle: syntaxe, sémantique, pragmatique*. Copenhagen : Nordisk Forlag.

LUNDQUIST, L. (1999), « Le factum textus. Fait de grammaire, fait de linguistique ou fait de cognition? » *Langue française*. 56-75.

LUNDQUIST, L. (a paraître) « Noms, verbes et anaphores (in)fidèles. Pourquoi les Danois sont plus fidèles que les Français. », *Langue française*.

MINEL, J.-L., CARTIER, E., CRISPINO, G., DESCLÉS, J.-P., BEN HAZEZ, S., JACKIEWICZ, A. (2001), « Résumé automatique par filtrage sémantique d'informations dans des textes, Présentation de la plate-forme FilText », *Technique et Science Informatiques,* n° 3, Hermès, Paris, p. 369-396

SEARLE, J. (1969), *Speech Acts, An Essay in the Philosophy of Language*, Cambridge, Cambridge University Press.

[2] Ce travail a été réalisé par deux étudiants, Isabelle Ranque et Benoit Lamey, du DESS ILSI de l'Université Paris-Sorbonne.

Projection et monotonie dans un langage de représentation lexico-grammatical

Benoit Crabbé
LORIA - Université Nancy 2
BP 239, 54506 Vandoeuvre-lès-Nancy Cedex
crabbe@loria.fr

Mots-clefs : Syntaxe, Lexique, Liage, Interface syntaxe sémantique, TAG

Keywords: Syntax, Lexicon, Linking, Syntax Semantics interface, TAG

Résumé Cet article apporte une méthode de développement grammatical pour la réalisation de grammaires d'arbres adjoints (TAG) de taille importante augmentées d'une dimension sémantique. La méthode que nous présentons s'exprime dans un langage informatique de représentation grammatical qui est déclaratif et monotone. Pour arriver au résultat, nous montrons comment tirer parti de la théorie de la projection dans le langage de représentation que nous utilisons. Par conséquent cet article justifie l'utilisation d'un langage monotone pour la représentation lexico-grammaticale.

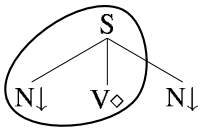
Abstract This papers provides a methodology for the grammatical development of large sized tree adjoining grammars (TAG) augmented with a semantic dimension. The provided methodology is expressed in a monotnic and declarative language designed for the compact representation of grammatical descriptions. To achieve the result, we show how to express a linking theory in the language used. Therefore this paper justifies the use of a monotonic language for lexico-grammatical representation.

1 Introduction

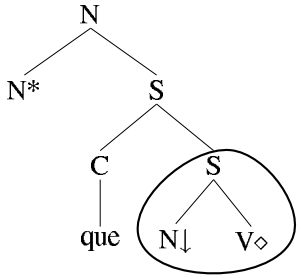
Dans cet article, nous donnons une méthode générale qui permet représenter de manière compacte un fragment significatif de grammaire TAG pour le français comportant une interface syntaxe-sémantique. TAG n'étant pas en tant que tel un formalisme d'analyse syntaxique comportant une spécification explicite de la représentation sémantique, nous utilisons ici la variante introduite par (GK03). Celle-ci utilise des unités élémentaires qui sont des structures à deux dimensions. Une dimension syntaxique qui est l'arbre TAG habituel, et une dimension sémantique qui représente l'information sémantique associée à cet arbre. La dimension sémantique est constituée de formules du langage de représentation sémantique utilisé par (GK03). Celui-ci est de la même famille que la *minimal recursive semantics*.

2 Enjeux méthodologiques

Pour représenter une grammaire TAG de manière non redondante, nous utilisons un langage informatique destiné à exprimer la grammaire de manière compacte (CD04). Celui-ci repose sur l'utilisation de macros (ou classes) réutilisables. Ce langage est destiné à servir de source à un interpréte effectivement implémenté qui réalise l'expansion de la description compacte (DLP04). Le langage utilisé pose que dans un système de description grammatical, il est souhaitable de reconnaître deux types de généralisations : les généralisations de structure d'une part et les généralisations d'alternatives d'autre part.



Jean *mange une pomme*



La pomme que **Jean** *mange*

On identifie dans la description grammaticale un certain nombre de constructions récurrentes que l'on souhaiterait pouvoir réutiliser. Par exemple dans les deux arbres suivants, nous avons encerclé l'information qui représente un sujet nominal réalisé en position canonique. D'autre part, et indépendamment du partage de structure, on souhaite capturer la notion d'alternative. Par exemple ou souhaitera identifier qu'une construction passive est une alternative d'une construction active. Les alternatives que nous identifions ont un statut particulier. En effet, ces généralisations contribuent à décrire des ensembles d'unités grammaticales mises en relation. Dans l'exemple donné ci-dessus, l'alternative entre un contexte transitif actif (a) et un contexte transitif passivisé (b) contribuent à décrire un ensemble de deux arbres qui partagent une sémantique commune (prédicat binaire transitif). La reconnaissance explicite des alternatives constitue un point important du langage que nous utilisons. En effet, une famille d'arbres (Abe02), n'est rien d'autre qu'un ensemble de réalisations alternatives d'une même structure prédicat-argument. Or la représentation des alernatives a souvent été négligée dans les propositions récentes pour la représentation grammaticale de TAG (Can99; Xia01; GCR02)[1].

Rejet des règles lexicales Le cadre que nous proposons ici se veut opérationnel. Dans ce contexte, la méthode que nous proposons se caractérise par le rejet des règles lexicales. La raison principale est que l'utilisation d'une mécanique procédurale dans le développement de grammaires d'arbres adjoints de tailles conséquentes pose en pratique de très sérieux problèmes d'ordonnancement de règles (Pro02). En lieu et place de règles lexicales, nous utilisons un langage purement déclaratif muni de deux opérateurs qui permettent de combiner des descriptions grammaticales fragmentaires : la conjonction et la disjonction.

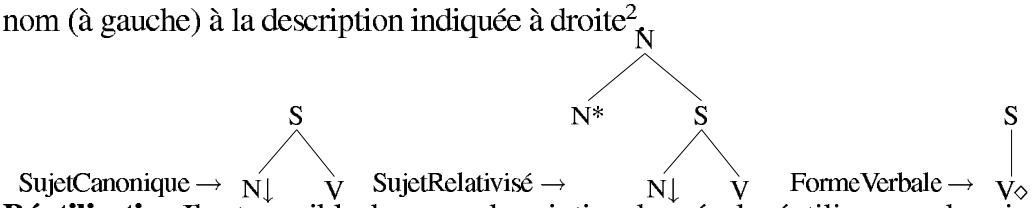
3 Langage utilisé

Le langage que nous utilisons (noté $\mathcal{L}_C$) permet de tirer parti des deux types de généralisations que nous avons identifiés. Celui-ci repose crucialement sur la notion de *classe*. Une classe contient une *description* de structure(s) grammaticale(s) partielle(s). Pour une TAG, une description est une description partielle d'arbre potentiellement augmentée de structures de traits associées

[1] En particulier, mentionnons que (Can99; GCR02) permettent d'exprimer les alternatives de manière détournée à l'aide d'un algorithme de « croisement ». En pratique nous avons observé que cet algorithme complique considérablement le travail de conception de grammaires.

aux noeuds. Une classe a pour fonction de nommer la description D à laquelle elle est associée de telle sorte qu'il est possible de réutiliser D par ailleurs dans une description D' quelconque. Ce sont les classes qui nous permettent dans le langage de capturer les généralisations grammaticales. Ce langage détaillé par (CD04) comporte quatre aspects.

Nommage Le langage permet de nommer une description. Associer un nom à une description permet de la réutiliser par après. Ceci est illustré par les exemples suivants où on associe un nom (à gauche) à la description indiquée à droite[2].



Réutilisation Il est possible dans une description donnée de réutiliser une description déjà associée à un nom par ailleurs. Dans l'exemple suivant, on réutilise ainsi la « macro » FormeVerbale dans la description de VerbeActif.

VerbeActif → *FormeVerbale*

Alternative On peut exprimer dans le langage la notion d'alternative (ou de choix) en utilisant le symbole $\vee$. Ainsi, l'exemple suivant montre comment exprimer que la notion de sujet recouvre aussi bien la notion de sujet canonique que de sujet relativisé.

Sujet → *SujetCanonique* $\vee$ *SujetRelativisé*

Composition Il est également possible de combiner deux descriptions. Ainsi, on peut exprimer qu'une famille de constructions intransitives est faite d'un sujet et d'une forme verbale à l'actif de la manière suivante :

VerbeIntransitif → *Sujet* $\wedge$ *VerbeActif*

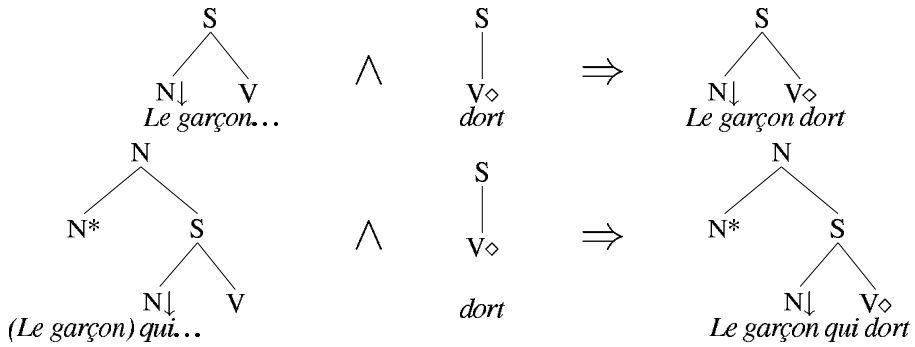


Figure 1: Deux arbres à contexte intransitif

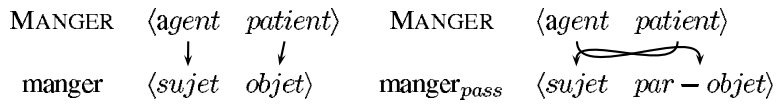
Interprétation du langage Formellement, comme l'indiquent (CD04), l'interprétation d'une description grammaticale d formulée en $\mathcal{L}_C$ se ramène à un interpréter un programme logique de la famille *Definite Clause Grammar* (DCG). Cela se comprend en considérant que d est une grammaire dont les terminaux sont des descriptions arborescentes, que l'opération de concaténation est ici la conjonction et en imposant que la DCG sous-jacente à d soit non récursive (CD04). L'interprétation d'une description d engendre de manière indéterministe toutes les phrases, prises comme des conjonctions de descriptions, de la DCG sous-jacente à d. Dans l'exemple présenté jusqu'ici, en considérant la classe *VerbeIntransitif* comme axiome de la DCG, l'interprète engendre deux arbres représentant deux contextes de verbe intransitif, comme indiqué en figure 1. Où on illustre sur la gauche les deux conjonctions de descriptions engendrées par l'interprète et sur la droite le résultat de la composition des descriptions[3].

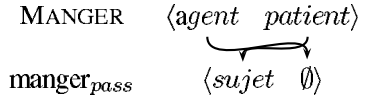
[2]Nous utilisons dans ce document une syntaxe abstraite pour illustrer notre propos. Un interprète concret pour ce langage a été implanté par (DLP04).

[3]Le mécanisme de composition est détaillé par (CD04)

4 Théorie de la projection lexicale

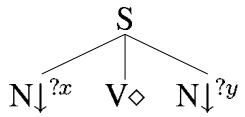
Le langage étant posé, nous donnons dans cette section une méthode générale qui permet d'exprimer de manière déclarative une grammaire à l'aide de ce langage. Cette méthode repose sur la théorie de la projection (*linking theory*). La fonction de projection associe chaque élément de la structure argumentale à un élément de la structure syntaxique, comme indiqué ici :





Approche monotone Utiliser la projection dans un langage tel que celui que nous proposons consiste à considérer que l'on génère un ensemble d'unités grammaticales complètement spécifiées par surspécifiation de la structure argumentale. Cette méthode permet de garantir un mécanisme monotone pour exprimer l'interface syntaxe sémantique. Pour illustrer ce point, prenons le cas classique du passif court qui, dans la littérature est souvent utilisé pour justifier l'utilisation d'une mécanique procédurale (Bec93; Can99; Xia01). Dans ce cas, *Marie est vue* est considéré comme variante de *Jean voit Marie* dans laquelle l'agent n'est pas exprimé. Le traitement par règle lexicale du passif court consiste à effacer l'agent (i.e. *Jean*). Dans le cas du modèle projectif, rien n'est effacé. Au contraire l'agent non exprimé est projeté sur une réalisation vide en syntaxe, comme illustré à gauche. L'approche projective ne fait qu'ajouter de l'information, et éventuellement de l'information vide à la structure argumentale.

Projection pour une grammaire TAG L'approche projective présentée jusqu'à présent s'adapte facilement à une grammaire semi-lexicalisée du type LFG. Ainsi dans une grammaire LFG, une structure syntaxique comme $manger\langle sujet \quad objet\rangle$ correspond directement à la f-structure de l'entrée lexicale d'un verbe.



◇(?x,?y)

En pratique, dans le cas d'une grammaire TAG à dimension sémantique, les entrées lexicales à décrire sont des schémas d'arbres comme illustré à droite. La dimension sémantique est prise comme contrepartie de la structure argumentale et la structure arborescente est prise comme contrepartie de la dimension syntaxique. La fonction de projection est représentée par le partage de valeurs des traits sémantiques associés aux noeuds de l'arbre et les variables d'individus de la prédication élémentaire (indiquée en gras).

La méthodologie que nous proposons repose sur l'idée de base que les arbres élémentaires TAG sont décrits par assemblage de fragments de structure correspondant à la représentation des arguments et du prédicat. La méthodologie qui suit consiste ensuite à produire des familles d'arbres en assemblant ces différents fragments. Ce qui généralise l'exemple donné en figure 1. La méthode comporte quatre étapes : (1) Description des fragments (2) Description de fonctions syntaxiques (3) Description de changements de diathèse (4) Description de familles. Pour réaliser l'interface syntaxe-sémantique, l'idée consiste à établir le lien entre la structure prédicative et les valeurs de traits dans les arbres à l'aide de classes paramétrées[4]

La première étape de la description consiste à définir des fragments représentant différentes

[4]On se rappelle que le langage de contrôle $\mathcal{L}_C$ est vu comme la contrepartie d'une DCG ce qui nous autorise à utiliser des paramètres.

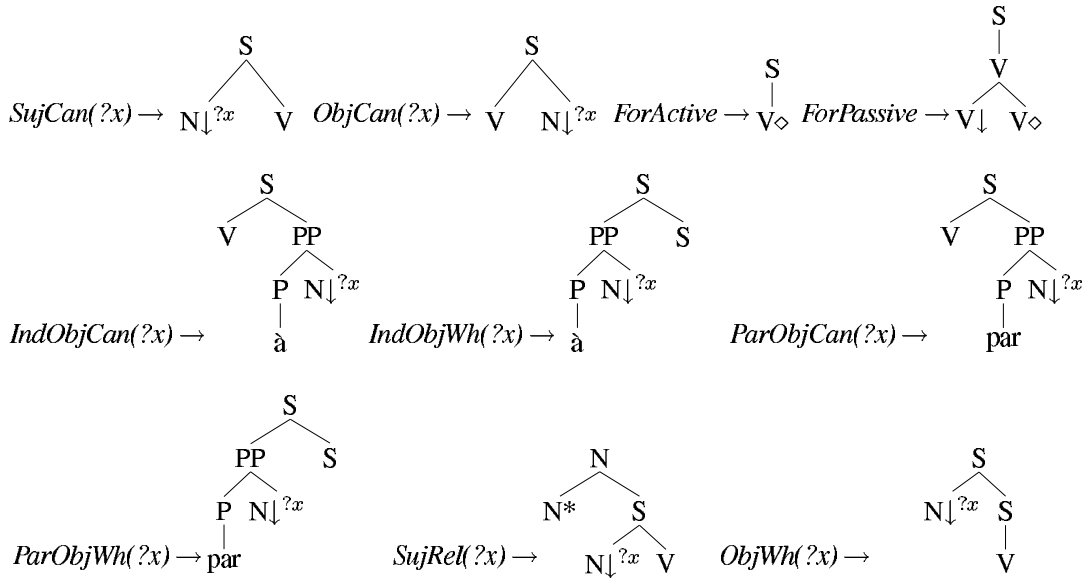


Figure 2: Fragments d'arbres utilisés comme « blocs de construction » de la grammaire

formes du prédicat et des arguments (Figure 2). Chacun de ces fragments est réutilisé pour la description des arbres. Par exemple, toute description d'arbre qui représente un contexte où le sujet est réalisé en position canonique utilisera le fragment *SujCan*. De plus, pour implémenter le liage syntaxe/sémantique, chaque classe décrivant un fragment arborescent coindexe la valeur du trait sémantique dans la structure arborescente avec un paramètre de classe approprié. La seconde étape regroupe les classes qui décrivent les fonctions syntaxiques. Celles-ci capturent le fait qu'une fonction syntaxique est une notion qui permet de s'abstraire de la question de l'ordre des mots.

(1) Sujet(?x) → SujetCan(?x) ∨ SujRel(?x)
Objet(?x) → ObjCan(?x) ∨ ObjWh(?x)
ParObjet(?x) → ParObjCan(?x) ∨ ParObjWh(?x)
ObjetIndirect(?x) → IndObjCan(?x) ∨ IndObjWh(?x)

Par exemple, la classe *Sujet* quoi que simplifiée ici capture le fait que le sujet recouvre alternativement une réalisation nominale devant le verbe (*SujCan*) ou en position relativisée (*SujRel*). En outre, les classes décrivant les variantes fonctionnelles contribuent à établir le liage entre syntaxe et sémantique en propageant la coindexation de valeurs par utilisation de paramètres.

La troisième étape où l'on décrit les changements de diathèse constitue le point important. Pour capturer l'alternance actif/passif, nous utilisons la définition suivante :

(2) AlternancePassive(?x, ?y) →
(Sujet(?x) ∧ FormeActive ∧ Objet(?y))
∨(Sujet(?y) ∧ FormePassive ∧ ParObjet(?x))
∨(Sujet(?y) ∧ FormePassive)

Ce qui indique que les réalisations alternatives de l'actif et du passif sont soit faites d'un sujet, d'une forme verbale active et d'un objet; soit d'un sujet, d'une forme verbale passive et optionellement d'un complément d'agent. Cette classe définit deux paramètres $?x$ et $?y$ qui sont coindexés avec les paramètres des classes de fonctions syntaxiques. Les valeurs des paramètres $?x$ et $?y$ représentent les valeurs associées aux arguments dans la structure argumentale. Ainsi la première ligne, qui représente l'actif, a pour effet de lier le premier argument ($?x$) au sujet et le second ($?y$) à l'objet. Par contre, la seconde ligne, qui représente le passif, lie le premier argument au complément d'agent et le second au sujet. Finalement la dernière ligne, qui représente le passif sans agent, lie uniquement le second argument au sujet. Le premier argument n'est tout simplement pas lié.

En dernier lieu, la définition des familles (Abe02) demande d'introduire explicitement la représentation sémantique et de lier les variables de la prédication élémentaire avec les paramètres de classe appropriés. On définira une famille ditransitive de la manière suivante :

(3) FamilleDitransitive →
AlternancePassive(?x, ?y) ∧ ObjetIndirect(?z)
∧ <sem> $\diamond(?x, ?y, ?z)$

où <sem> représente la description d'un littéral sémantique dont les variables d'individu sont coindexées avec les paramètres de classes appropriés Le littéral sémantique ainsi décrit est la contrepartie dans notre langage de ce que la théorie de la projection appelle la structure argumentale. À chacun des niveaux de description les variables sont coindexées de manière appropriée. La génération de l'ensemble des arbres de la famille produit bien un ensemble de couples (arbre, représentation sémantique) dans lesquels la projection est bien réalisée.

5 Conclusion

Dans cet article, nous avons proposé une méthode de développement de grammaires d'arbres adjoints qui s'exprime dans un cadre monotone. Cette méthode permet de justifier l'usage d'un langage monotone pour la représentation lexico-grammaticale. En second lieu elle ouvre la porte au développement de grammaires d'arbres adjoints à large couverture comportant une interface syntaxe-sémantique. Ces grammaires doivent être interfacées avec les analyseurs TAG existants (GP05). Le langage ainsi que la méthode proposés ont permis de générer une grammaire d'arbre adjoints de taille importante (environ 4000 arbres). De plus le cadre proposé ici s'adapte facilement au développement de grammaires pour d'autres formalismes fortement lexicalisés. Ainsi, G. Perrier a pu réutiliser cette méthode pour produire une grammaire d'interaction. Nous avons également mené quelques expériences encourageantes qui ont permis de créer une petite grammaire XDG du français.

Références

Anne Abeillé. (Abe02) *Une grammaire d'arbres adjoints pour le français.* CNRS, Paris, 2002.

Tilman Becker. (Bec93) *HyTAG: A new Type of Tree Adjoining Grammars for Hybrid Syntactic Representation of Free Word Order Language.* PhD thesis, Universität des Saarlandes, 1993.

Marie-Hélène Candito. (Can99) *Organisation Modulaire et Paramétrable de Grammaires Electroniques Lexicalisées.* PhD thesis, Université de Paris 7, 1999.

Benoit Crabbé and Denys Duchier. (CD04) Metagrammar redux. In *Constraint Solving and Language Processing*, Copenhagen, 2004.

Denys Duchier, Joseph Leroux, and Yannick Parmentier. (DLP04) The metagrammar compiler: An nlp application with a multi-paradigm architecture. In *Mozart 2004*, Charleroi, 2004.

Bertrand Gaiffe, Benoit Crabbé, and Azim Roussanaly. (GCR02) A new metagrammar compiler. In *Proc. TAG+6*, Venise, 2002.

Claire Gardent and Laura Kallmeyer. (GK03) Semantic construction in feature-based tree adjoining grammar. In *Proc. EACL*, 2003.

Claire Gardent and Yannick Parmentier. (GP05) Large scale semantic construction for tree adjoining grammar. In *Proc. LACL 2005*, Bordeaux, 2005.

Carlos Prolo. (Pro02) Systematic grammar development in the XTAG project. In *COLING'02*, 2002.

Fei Xia. (Xia01) *Automatic Grammar Generation from two Different Perspectives.* PhD thesis, University of Pennsylvania, 2001.

Dialogue automatique et personnalité : Méthodologie pour l'Incarnation de Traits Humains

Florence Duclaye, Franck Panaget
France Télécom R&D
2, avenue Pierre Marzin 22307 Lannion Cedex
{florence.duclaye,franck.panaget}@francetelecom.com

Mots-clefs : dialogue incarné, personnages virtuels, personnalité, génération automatique

Keywords: embodied dialogue, virtual characters, personality, natural language generation

Résumé Cet article introduit une méthodologie d'intégration de la personnalité dans un système de dialogue automatique, en vue de l'incarnation de personnages virtuels. Notion complexe non encore épuisée dans la littérature, la personnalité d'un individu peut s'illustrer de multiples manières possibles. Notre objectif consiste à présenter une méthode générique de prise en compte de la personnalité dans un système de dialogue par modélisation et exploitation des connaissances relatives à la personnalité de l'individu à incarner. Cet article présente les avantages et inconvénients de cette méthode en l'illustrant au travers de la stylistique des énoncés générés par le système.

Abstract This article introduces a methodology to integrate personality into a dialogue system. This work constitutes a step towards the embodiment of virtual characters. The personality of an individual, which is a complex and non-exhausted concept in literature, can show through multiple possible ways. Our purpose is to describe a generic methodology of personality integration into a dialogue system based on the modelling and use of knowledge relative to the personality of a given individual. This article describes the advantages and drawbacks of the proposed methodology by illustrating it through the linguistic style of the generated messages.

1 Introduction

La personnalité constitue un champ de recherche prometteur dans le monde des systèmes d'interaction homme/machine. Inscrite dans la perspective d'une meilleure acceptabilité des systèmes de dialogue par les utilisateurs, la présente étude propose une méthodologie d'introduction d'une personnalité dans ces systèmes. Cet article se présente comme suit. La partie 2 introduit le concept de personnalité et son intégration dans les systèmes existants à travers les travaux de la littérature. La partie 3 développe la manière dont la personnalité est modélisée et exploitée pour être rendue à travers notre système de dialogue. Quelques résultats sont présentés dans cette partie et nous amènerons à mettre en évidence les perspectives d'évolution possibles de notre système.

2 Personnalité et dialogue automatique : analyse de l'existant

2.1 La personnalité : caractéristiques et moyens d'expression

La littérature nous fournit plus de cinquante définitions du concept de personnalité (Allport, 1937) mais toutes les théories s'accordent à affirmer que la personnalité concerne les caractéristiques uniques et distinctives des individus et s'intéresse à la nature fondamentale de l'être humain. (Cattell, 1965) affirme l'idée suivante : "la personnalité est ce qui permet de prédire ce qu'une personne fera dans une situation donnée". Notre objectif est d'analyser par quels moyens la personnalité peut s'exprimer afin de les reproduire. La théorie développée par Cattell repose sur quelque seize paramètres, tels que le tempérament extraverti/réservé, le niveau d'intelligence, le tempérament expérimentateur/traditionnel, téméraire/timide, etc. Chacun de ces traits varie pour constituer la personnalité d'un individu. (McCrae, John, 1992) proposent une version distillée de la théorie de Cattell et considérent seulement cinq traits qui se combinent pour constituer la personnalité ("Five Factor Model"). Ces traits sont l'ouverture (fantaisie, libéralisme, créativité...), le tempérament consciencieux (ordre, auto-discipline, volonté de réussir...), l'extraversion (agitation, calme, timidité...), le tempérament agréable (altruisme, indifférence, égo-centrisme...) et la névrose (anxiété, dépression, impulsivité...).

2.2 Intégration de la personnalité dans les systèmes de dialogue existants

De multiples moyens sont envisageables pour faire véhiculer des traits de personnalité par un système de dialogue : l'apparence physique (visage, corps, vêtements...), l'intonation de la voix, les traits prosodiques, la vitesse d'élocution, ou encore la manière de s'exprimer du personnage (ex : choix lexicaux, syntaxiques, blancs, hésitations...). Les actions menées par un personnage constituent une autre manière efficace de transmettre sa personnalité (ex : tempérament téméraire/timide), tout comme le contenu de ses actes de langage (ex : degré de coopération). Nous n'abordons ici que les travaux de recherche portant sur l'expression de la personnalité d'un individu au travers de ses productions linguistiques.

La littérature traditionnelle introduit souvent directement les paramètres stylistiques dans le générateur. Par exemple, (Biber, 1988) analyse la variation systématique entre énoncés en ter-

mes de dimensions sous-jacentes. L'auteur considère pour cela un espace multi-dimensionnel, dans lequel chaque dimension comporte un ensemble de traits syntaxiques et lexicaux qui cooccurrent fréquemment dans les énoncés (ex: temps employé, etc). Cette méthode de paramétrage du générateur par des traits sémantiques, surfaciques, etc, est reprise par (Stede, 1993) et (Reiter, 2004). (Hovy, 1990) constitue l'un des travaux-phares en matière de variation stylistique d'énoncés générés. Les paramètres de son générateur concernent des aspects comme l'atmosphère conversationnelle de l'interaction, les caractéristiques personnelles des interlocuteurs, ou encore le but du locuteur sur son auditeur.

3 Modélisation et expression de la personnalité dans notre système de dialogue

3.1 Modélisation de la personnalité

Notre système de dialogue est constitué de trois composants principaux. Une unité rationnelle, qui confère au système ses capacités de dialogue, implémente une Théorie de l'Interaction formalisant des attitudes mentales telles que la croyance ou l'intention avec un modèle d'actions communicatives (Bretier, Sadek, 1997). En contexte de dialogue homme-machine, le système nécessite une unité de traitement des langues naturelles qui produit l'interprétation sémantique des énoncés de l'utilisateur d'une part, et d'autre part les énoncés verbalisant les actes communicatifs planifiés par l'unité rationnelle. Les connaissances nécessaires pour comprendre, raisonner et s'exprimer sont présentes dans le composant de gestion des connaissances, qui est notamment constitué d'un réseau sémantique spécifique à l'application visée (domaine fermé). La personnalité est modélisée dans notre système de dialogue, au niveau de ce composant de gestion des connaissances, sous forme d'un modèle sémantique unique. La figure 1 présente le sous-modèle décrivant un individu. On y aperçoit au centre la classe $Personne$, d'où partent trois relations ($Biographie$, $LangueParlee$, $Personnalite$). $LangueParlee$ désigne les langues connues par la $Personne$ et la manière dont elle les parle (voir 3.2). Un ensemble de classes est lié à la $Personnalite$, inspiré des traits de personnalité majeurs mis en évidence par (McCrae, John, 1992) (voir 2.1). La figure 1 n'inclut pas l'intégralité des caractéristiques du personnage mais illustre néanmoins que les connaissances sont modélisées à un seul endroit dans le système de dialogue et peuvent être exploitées de divers moyens possibles. Dans cet article, nous ne détaillerons que l'influence de ces traits sur la génération (voir partie 3.2). Notons que nous ne modélisons pas la personnalité de l'allocutaire humain avec lequel l'interaction se fait. L'originalité de cette approche réside dans le fait que les paramètres permettant au système de dialogue d'incarner un personnage sont modélisés et entrés au coeur du système.

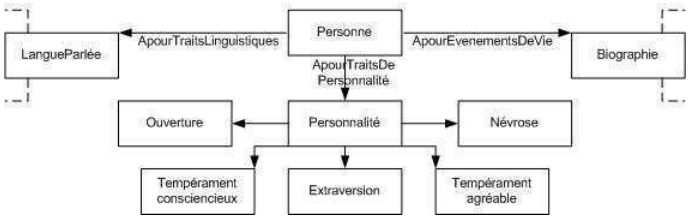


Figure 1: Extrait de la base de connaissances

3.2 La personnalité vue à travers l'expression des messages

Afin de restituer les traits de personnalité, le générateur prend en compte les connaissances relatives au profil linguistique du personnage à incarner au moment des choix lexicaux (entre pseudo-synonymes) et syntaxiques (insertion d'hésitations, de tics de langage...). La figure 2 illustre (réseau incomplet) notre modélisation d'un profil linguistique. Pour une $LangueParlee$, de $Nature$ donnée (ex : français), un ensemble de paramètres associés décrivent la manière dont le personnage $Personne$ s'exprime dans cette langue (niveau de langage, la couverture lexicale ou $Etendue$, ou encore variantes parlées).

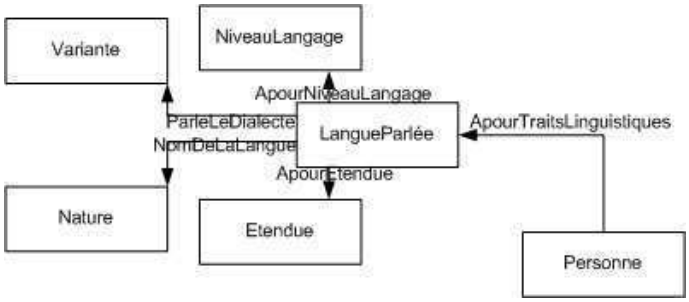


Figure 2: Extrait de la base de connaissances - Profil linguistique d'un individu

Le générateur de notre système de dialogue (Panaget, 1996) respecte le schéma fonctionnel classique à trois composants (Levelt, 1989) : un macro-planificateur, un micro-planificateur et un formulateur. Le macro-planificateur gère la structure organisationnelle du message à générer en transformant les actes communicatifs produits par l'unité rationnelle en une expression de sémantique idéationnelle. Le micro-planificateur sélectionne le meilleur ensemble de ressources linguistiques (lexicales et grammaticales) verbalisant l'expression à communiquer. Finalement, le formulateur utilise une grammaire d'arbres adjoints (Joshi, 1987) pour ordonner les éléments de l'énoncé, les accorder et appliquer un traitement morphologique. Les traits de personnalité qui peuvent produire des effets sur la forme des messages produits sont utilisés pour impacter la macro-planification (production d'expressions idéationnelles différentes), et la micro-planification (sélection de pseudo-synonymes). Au niveau de la macro-planification, le générateur peut chercher à réduire une expression idéationnelle à verbaliser dans le but de produire un énoncé minimal. Par exemple, à la question "*Où es-tu né ?*" , le système pourra répondre "*Je suis né à Ulm*" (personnage posé, rigoureux) ou uniquement "*à Ulm*" (personnage stressé, pressé). La seconde tâche est l'introduction d'interjections telles que l'exclamation positive ("*ah*") ou négative ("*oh*") (personnage extraverti) ou encore l'hésitation ("*euh*") (personnage timide). Un mécanisme d'ajout de tics de langage a aussi été mis en oeuvre. Au niveau de la micro-planification, le générateur met en oeuvre un mécanisme de sélection de ressources linguistiques verbalisant l'expression idéationnelle produite. Les ressources linguistiques possèdent des informations de style qui les différencient. Pour un concept donné, le choix entre les pseudo-synonymes disponibles est fondé sur l'exploitation d'un vecteur de style associé aux ressources linguistiques. Dans la version actuelle, les vecteurs de style sont limités à trois traits : (1) formalité (vulgaire, familier, courant, formel, soutenu), (2) fioriture (réaliste, imagé, figure de style), (3) temporel (archaïque, vieilli, actuel, néologisme). Le pseudo-synonyme retenu est celui ayant la distance la plus faible selon la formule ci-dessous. Cette formule est constituée de trois composantes. La première composante est la distance euclidienne au carré entre le vecteur de style de la ressource linguistique candidate RL ($Vect_{RL}$) et le vecteur de

référence ($Vect_{Ref}$), défini par le modèle sémantique de l'individu. Cette distance est semblable à celle décrite dans (Stede, 1993). La seconde composante, qui est la distance entre la ressource candidate et la dernière ressource linguistique sélectionnée ($Vect_{RLprec}$), exprime le fait qu'il faut éviter au maximum les écarts de styles entre les ressources sélectionnées. La troisième composante modélise l'étendue des compétences linguistiques de l'individu en pénalisant les ressources qui ont été utilisées dernièrement (Ph_c est le numéro de la phrase en cours de génération, Ph_{RL} est celui de la dernière phrase construite à partir de la ressource linguistique RL, Ph_c=Ph_{RL} si cette ressouce n'a pas encore été retenue). Le paramètre $Etendue$ représente l'étendue des compétences langagières de l'individu (diversification des ressources linguistiques choisies).

$$d(Vect_{RL}, Vect_{Ref}, Vect_{RLprec}, Ph_c, Ph_{RL}) = \quad (1)$$

$$si \quad Ph_c = Ph_{RL} \quad (2)$$

$$\alpha_1 \sum_{i=1}^{n}(Vect^i_{Ref} - Vect^i_{RL})^2 + \alpha_2 \sum_{i=1}^{n}(Vect^i_{RLprec} - Vect^i_{RL})^2 + \alpha_3 Etendue \quad (3)$$

$$sinon \quad (4)$$

$$\alpha_1 \sum_{i=1}^{n}(Vect^i_{Ref} - Vect^i_{RL})^2 + \alpha_2 \sum_{i=1}^{n}(Vect^i_{RLprec} - Vect^i_{RL})^2 + \alpha_3 \frac{Etendue}{Ph_c - Ph_{RL}} \quad (5)$$

Cette méthode est illustrée dans une application de dialogue avec deux personnages virtuels aux personnalités très différentes. Charles est un brillant scientifique ayant vécu au début du 20ème siècle et ayant un style linguistique riche et soutenu. Fred est une pop star de l'époque actuelle et s'exprime d'une manière quasi-opposée à celle de Charles :

Charles	:	*Mes hommages, Madame ! Mes hommages, Monsieur !*
		Que puis-je faire pour vous être agréable ?
Utilisateur	:	*En quelle année êtes vous né, Charles ?*
Charles	:	*J'ai vu le jour en 1879.*
Utilisateur	:	*Et dans quelle ville avez-vous vu le jour ?*
Charles	:	*Je suis venu au monde à Ulm.*
Utilisateur	:	*Quand êtes-vous mort ?*
Charles	:	*J'ai vécu jusqu'en 1955.*
Utilisateur	:	*Merci Charles. Au revoir !*
Charles	:	*Dois-je comprendre que vous souhaitez mettre fin à notre conversation ?*
Utilisateur	:	*Oui, je voudrais parler avec Fred, maintenant.*
Fred	:	*Ah que coucou ! Qu'est-ce que tu veux ?*
Utilisateur	:	*Quand es-tu né, Fred ?*
Fred	:	*A que j'ai pointé le bout du nez en 1943.*
Utilisateur	:	*Où tu es né ?*
Fred	:	*Que je suis né à la capitale.*
Utilisateur	:	*OK. Merci Fred. A bientôt !*
Fred	:	*Tu veux vraiment qu'on arrête de discuter tous les deux ?*
Utilisateur		*Oui. A plus tard.*
Fred	:	*Tchao, à plus.*

4 Conclusion et perspectives

Cet article présente une méthode générique de prise en compte de la personnalité d'un individu en vue de son incarnation par un système de dialogue automatique. La méthode décrite permet

d'exprimer la personnalité au travers de variations stylistiques des messages générés. Notre méthode présente l'avantage d'intégrer les données relatives à la personnalité au niveau-même du composant de gestion des connaissances et ainsi d'exploiter ces paramètres dans différents composants, et donc par plusieurs moyens possibles (génération, raisonnement...). De plus, par le paramétrage de chaque trait indépendamment des autres traits, cette méthode permet de véhiculer de nombreuses personnalités existantes, ou inexistantes. Illustrés sur deux personnages virtuels, les résultats obtenus s'avèrent d'ors et déjà prometteurs, mais tracent également la voie des prochaines améliorations. Une modélisation plus fine de la personnalité permettrait en effet d'atteindre plus de granularité et d'atténuer la caricature comportementale et stylistique obtenue. De plus, des axiomes de comportement adaptés restent à mettre en oeuvre dans le système. Enfin, les prochains efforts se concentreront sur une reproduction stylistique plus précise et pourront par exemple s'appuyer sur des travaux tels que ceux de (Paiva, Evans, 2004), qui propose une méthode d'apprentissage de règles de production stylistique à partir de corpus de textes illustrant un éventail de variations stylistiques.

Références

Allport G.W. (1937), *Personality - A psychological interpretation*, New York, Henry Holt.

Biber D. (1988), Variation across speech and writing, *Cambridge University Press*, Cambridge.

Bretier P., Sadek D. (1997), A rational agent as the kernel of a cooperative spoken dialogue system: implementing a logical theory of interaction. *Proc. of ECAI'96 workshop on Agent theories, Architectures and Languages*, 189-204, Berlin:Springer.

Cattell R.B. (1965), The Scientific Analysis of Personality, Penguin.

Hovy E.H. (1990), Pragmatics and Natural Language Generation, *Artificial Intelligence*, Vol. 43, 153-197.

Joshi A. (1987), The relevance of Tree Adjoining Grammar, *Natural Language Generation: News Results in Artificial Intelligence, Psychology and Linguistics*, G. Kempen ed., Nato Asi Series 135, Martinus Nijhoff Publishers, Boston.

Levelt W. (1989), *Speaking: From Intention to Articulation*, Cambridge MIT Press.

McCrae R., John O. (1992), An introduction to the five-factor model and its applications, *Journal of Personality*, Vol. 60, 175-216.

Paiva D., Evans R. (2004), A Framework for Stylistically Controlled Generation, *International Conference on Natural Language Generation*, 120-129.

Panaget F. (1996), D'un système générique de génération d'énoncés en contexte de dialogue oral à la formalisation logique des capacités linguistiques d'un agent rationnel dialoguant. *Thèse de doctorat, Université de Rennes I, France.*

Reiter E., Sripada S. (2004), Contextual Influences on Near-Synonyme Choice, *Proceedings of the Third International Conference on Natural Language Generation*, 161-170.

Searle, J.R. (1969), *Speech Acts*, Cambridge University Press.

Stede M. (1993), Lexical Choice Criteria in Language Generation, *Proceedings of European Chapter of the Association for Computational Linguistics*, 454-459.

RITEL : dialogue homme-machine à domaine ouvert

Olivier Galibert, Gabriel Illouz, Sophie Rosset
LIMSI - CNRS
F-91403 Orsay Cedex
{galibert,gabrieli,rosset}@limsi.fr

Mots-clefs : dialogue homme machine, recherche d'information précise, corpus

Keywords: human machine dialog, question answering, information retrieval, corpus

Résumé L'objectif du projet RITEL est de réaliser un système de dialogue homme-machine permettant à un utilisateur de poser oralement des questions, et de dialoguer avec un système de recherche d'information généraliste (par exemple, chercher sur l'Internet "Qui est le Président du Sénat ?") et d'en étudier les potentialités. Actuellement, la plateforme RITEL permet de collecter des corpus de dialogue homme-machine. Les utilisateurs peuvent parfois obtenir une réponse, de type factuel (**Q :** qui est le président de la France ; **R :** Jacques Chirac.). Cet article présente brièvement la plateforme développée, le corpus collecté ainsi que les questions que soulèvent un tel système et quelques unes des premières solutions envisagées.

Abstract The project RITEL aims at integrating a spoken language dialog system and an open-domain question answering system to allow a human to ask a general question (f.i. "Who is currently presiding the Senate?") and refine his research interactively. As this point in time the RITEL platform is used to collect a new human-computer dialog corpus. The user can sometimes recieve factual answers (**Q :** who is the president of France ; **R :** Jacques Chirac). This paper briefly presents the current system, the collected corpus, the problems encountered by such a system and our first answers to these problems.

Introduction

Les progrès réalisés ces dernières années tant en reconnaissance de la parole qu'en recherche d'information permettent d'envisager de nouvelles études. L'objectif du projet RITEL est de réaliser un système de dialogue homme-machine permettant à un utilisateur de poser oralement des questions, et de dialoguer avec un système de recherche d'information généraliste (par exemple, chercher sur l'Internet "Qui est le Président du Sénat ?") et d'en étudier les potentialités. Le terme de **système de dialogue** indique généralement un système permettant une interaction entre un humain et un système dans un cadre restreint (Glass J. R. et al., 2000). Toutefois, notamment dans le cadre des travaux sur les systèmes de question-réponse, le cadre tend à s'élargir. Un dialogue est une suite d'échanges entre interlocuteurs dans un contexte donné. Un système de dialogue homme-machine interprète les requêtes de l'utilisateur en fonction de la tâche à accomplir, de l'histoire du dialogue et du comportement de l'utilisateur. Son

objectif est de donner à l'utilisateur les informations recherchées tout en assurant une interaction efficace et naturelle. Actuellement, Les systèmes concernent des domaines restreints tels que l'information horaire de moyens de transports (trains, avions, cinéma) ou les informations touristiques, par exemple le projet européen Le3-Arise (horaires de train), le projet américain ATIS du DARPA Communicator (voyages en avions), et le projet français Technolangue MEDIA (informations touristiques). Ces projets ont donné lieu à des évaluations et ont permis d'en asseoir la faisabilité. Des modèles de gestion dynamique du dialogue et de génération adaptée ont été proposés. Même si ce qu'on entend par interaction naturelle est très variable d'un système à un autre (Villaneau J., 2003), ces systèmes permettent une interaction (orale) relativement naturelle : l'utilisateur peut à tout moment changer d'avis et revenir sur des choix exprimés, interrompre la réponse du système en prenant la parole, le système peut lui aussi changer de stratégie d'interaction en fonction des réactions de l'utilisateur. Un système de dialogue utilise des sources de connaissances diverses et complexes : connaissances acoustiques, phonétiques, lexicales, morphologiques, syntaxiques et sémantiques, pragmatiques, ainsi que des connaissances sur le dialogue, sur la tâche à réaliser et sur l'interlocuteur. En recherche d'information et extraction d'information, des progrès (Harabagiu S. et al., 2001) ont été motivés par des campagnes d'évaluations (américaine TREC, 98-2003, européenne CLEF, et nationale Equer/Technolangue, 2004). Ces systèmes sont limités à une question et une réponse. Des tentatives ont été faites pour des questions enchaînées, portant sur un même thème. Mais il ne s'agissait pas de dialogues, il n'y avait pas réellement d'interaction, il n'y avait pas de négociations possibles. Le projet RITEL s'appuie sur des progrès récents en reconnaissance de la parole conversationnelle et multi-locuteurs (Gauvain J.L., Lamel L., 2002) Un projet riche et complexe doit faire face à plusieurs points épineux. Les plus évidents sont : la reconnaissance de la parole qui doit être à grand vocabulaire et sur laquelle une contrainte temps réel s'applique, la gestion d'un dialogue en domaine ouvert, la communication et l'échange d'informations entre un système de question-réponse et le dialogue, la génération de la réponse. Cet article ne répond pas à toutes ces questions, seuls certains problèmes rencontrés et les réponses apportées seront présentés. Nous présentons un état des lieux du projet RITEL et décrivons le corpus collecté jusqu'à présent avec la plateforme. Nous concluons sur les perspectives de cette étude.

1 Dialogue oral et recherche d'information : intégration

Un système de dialogue oral homme-machine a pour objectif de donner à l'utilisateur l'information qu'il recherche en s'aidant de diverses sources de connaissances (statiques : connaissance du domaine, dynamique...), et le système de question-réponse de rechercher une réponse précise à une question. L'objectif du projet RITEL est l'intégration de ces différents systèmes en une plateforme unique.

1.1 Reconnaissance de la parole

Intégrer un système de reconnaissance de parole dans une telle plateforme suppose de mettre en place un système de reconnaissance à grand vocabulaire (taille de lexique de 65000 à 300000 mots), temps réel, multi-locuteur et fonctionnant sur un signal téléphonique. Dans de telles conditions aucun système de reconnaissance, au niveau de l'état de l'art, ne permet d'obtenir des performances nécessaires pour un système de dialogue (aux alentours de 20% d'erreur). Il faut donc envisager des techniques d'adaptations dynamiques des différents modèles. L'adaptation

Table 1: Exemples avec une question : *qui est le président des États-Unis*?

Réponses	score	thème	contexte	*Réponses*	score	thème	contexte
G.W. Bush	0.99	pol	2000-	élu au suffrage indirect	0.6	droit	-
Bill Clinton	0.7	pol	93-2000	né sur le sol américain	0.6	droit	-
Satan	0.1	pol_opi	2005	un pantin	0.2	pol_opi	-
Mère Theresa	0.2	pol_opi	2002-04				
Dumbo Bush	0.3	pol_opi	2004				
Bartlet	0.6	fic_série	1999-				
H. Ford	0.5	fic_film	1997-				

dynamiques des modèles de langage est rendue possible par l'indexation en thème des énoncés utilisateur et des réponses du système de recherche d'information. L'adaptation des modèles acoustiques s'effectue dynamiquement et de manière de plus en plus poussée au fur et à mesure des échanges entre l'utilisateur et le système. De plus le gestionnaire de dialogue peut le cas échéant demander à l'utilisateur d'épeler certains mots de sa demande.

1.2 Recherche d'information

Le système d'information prend en entrée la sortie du système de reconnaissance. Il s'agit de parole libre et potentiellement erronée. Il est donc nécessaire d'adapter la communication par rapport à un moteur question-réponse classique, notamment l'analyse de la question ne peut se faire à l'aide de contraintes morpho-syntaxiques fortes mais plutôt d'une analyse syntaxico-sémantique plus lâche. Les étapes suivantes restent sensiblement les mêmes que dans un moteur question-réponse classique. Par contre, pour ce qui est du retour de l'information, celle-ci doit être adapté au dialogue. Il ne s'agit plus seulement de mettre les "meilleures" réponses en premier mais bien de permettre au dialogue d'aider l'utilisateur à choisir celles qui lui conviennent le mieux. Pour ce faire, les réponses sont constituées de listes indicées par un score de confiance ce qui aide le système de dialogue à prendre sa décision pour générer une réponse (ou non) à l'utilisateur (cf. tableau 1). Selon ce score la réponse pourra comporter une information informant l'utilisateur du degré de confiance qu'a le système dans sa réponse (ex. je crois que,...) Le nombre de document est lui aussi associé à une réponse. Pour un grand nombre de document retourné, deux stratégies sont possibles.
Le **Regroupement par thèmes** permet au dialogue de proposer différentes possibilités à l'utilisateur pour que celui-ci oriente sa recherche. Pour chacun de ces topics, un score de confiance sera attribué de façon à aider le dialogue à orienter au mieux l'utilisateur. (**R:** *J'ai plusieurs réponses possibles, 2 dans le domaine de la politique, 3 qui s'apparentent à des opinions et 2 concernant des fictions. Quelle est la thématique de votre recherche ?*)
La **Demande de précision** a lieu si le regroupement par thème n'est pas possible. Le système soit demande à l'utilisateur de préciser sa requête soit lui propose quelques exemples pour qu'il y réagisse. (**R:** *J'ai des réponses avec des noms de personnes et d'autres sans. Par exemple, G.W. Bush est le ... ou une réponse de type définition comme élu au suffrage universel indirect. Que recherchez-vous précisément ?*) Ainsi, contrairement à l'augmentation de la précision en utilisant le retour d'information en aveugle (*Blind relevance feedback*), nous avons ici un retour d'information éclairé par l'utilisateur.

2 Architecture générale du système actuel

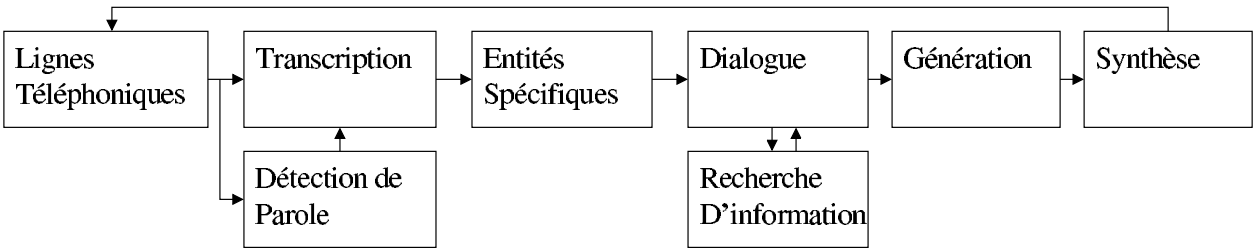


Figure 1: Organisation des composants de RITEL

Le système est composé d'un ensemble simple de composants communicants (figure 1). Le module de gestion de lignes téléphonique envoie le signal audio au détecteur de parole et au système de transcription. Le détecteur de parole contrôle l'activité de la transcription. La suite de mots obtenue est envoyée à la détection d'entités spécifiques puis au gestionnaire de dialogue. Celui-ci communique avec le composant de recherche d'informations pour décider d'une réponse à l'utilisateur. Le schéma sémantique obtenu est envoyé à la génération qui en fait une phrase que la synthèse transforme en un signal audio qui revient finalement au module téléphonique. Nous présentons brièvement le système de reconnaissance, le système d'analyse et le gestionnaire de dialogue dans les sections suivantes.

2.1 Détection de parole et système de reconnaissance

La détection de parole est basée sur notre segmenteur de conversations téléphoniques spontanées : des modèles multi-gaussiens modélisent le silence et la parole. Un décodage de viterbi classique segmente le signal. Le système de reconnaissance est un décodeur monopasse utilisant des modèles acoustiques triphones et un modèle de langage trigramme suivi d'un rescoring en consensus avec un modèle quadrigramme. Les modèles acoustiques sont ceux du système de dialogue Arise (Lamel et al., 2000). Vocabulaire (65000 mots, 1.7% OOV) et modèles de langage ont été obtenus par interpolation de modèles de diverses sources sur le développement. Elle sont le corpus d'entraînement RITEL, les textes de journaux de 1988 à nos jours, les transcriptions d'émissions de nouvelles radio- et télé-diffusées, des questions venant d'internet (ex.: Madwin.fr). La prochaine version intègrera les nouvelles transcriptions du corpus ainsi que des transcriptions de conversations téléphoniques spontanées.

2.2 Détection d'entités spécifiques

Un système de détection d'entités spécifiques est utilisé. Il prend en entrée la suite de mots envoyée par le système de reconnaissance. Les entités spécifiques se décomposent en trois classes: **Entités nommées** telles que les personnes (<pers>), les productions (<prod>), les marqueurs temporels (<time>), les organisations (<org>), les unités lexicales (<ul>) et les lieux (<loc>), les **Marqueurs syntaxiques de requête** (qui, où, quand, comment, combien...), les **Marqueurs sémantiques de requête** qui correspondent à des thèmes (littérature, géographie, histoire, société, cinéma, politique, économie...) et sous-thèmes (auteur, dirigeant, orthographe,

Table 2: Génération de réponses : à gauche, une réponse factuelle, à droite une incitation

S: Bonjour et bienvenue sur le service Ritel du Limsi. Souhaitez-vous écouter le guide ?	**S:** Bonjour et bienvenue sur le service Ritel du Limsi. Souhaitez-vous écouter le guide ?
U: euh non merci je voudrais la capitale de la France	**U:** je voudrais des informations sur l' ONU
S: La capitale de la France est Paris. Quelle autre informations désirez-vous ?	**S:** Vous cherchez une information sur ONU. Pouvez-vous être plus précis ?
U: euh quel est son président	**U:** quels pays sont impliqués
S: Jacques-Chirac est le dirigeant de la France. Quelle autre informations désirez-vous ?	**S:** Votre question porte sur la géographie. Mais plus précisément ? Un pays ? Un ville ?

population...). Toutes ces entités sont extraites et normalisées via un ensemble de règles écrites sous forme d'expressions régulières de mots. Des macros (règles locales) et des listes peuvent être utilisées pour la définitions des règles. Ces macros et listes, ainsi que des règles elles-mêmes, peuvent servir dans la définition de contexte. Les analyses sont séparées selon leur fonction : la première étape consiste à traiter les nombres. La deuxième étape consiste en un traitement lexical c'est à dire une normalisation des mots marqueurs syntaxiques de questions. La troisième étape enfin annote les thèmes, sous-thèmes et entités nommées. Ainsi un énoncé comme *qui a écrit le rouge et le noir en mille huit cent trente* a pour représentation *<subsubject> <Tauteur> <Tqui> qui </Tqui> a écrit </Tauteur> </subsubject> <prod> le rouge et le noir </prod> <time> <1830> mille huit cent trente </1830> </time>* .

2.3 Gestionnaire de dialogue

La première version du gestionnaire de dialogue a consisté exclusivement à inciter les sujets à parler le plus possible tout en permettant de conserver une interaction raisonnablement naturelle. La seconde version permet d'interroger une base de données. Le rôle du gestionnaire de dialogue est d'interpréter contextuellement le schéma sémantique que le système de détection d'entités spécifiques lui envoie ; récupérer les informations nécessaires à l'interrogation de la base de données ; générer les schémas sémantiques pour la génération ; choisir une stratégie de génération ; Le gestionnaire de dialogue commence par mettre à jour les différents marqueurs caractérisant le dialogue en cours (fonctionnalités en cours, générations précédentes, nombre de nouveaux éléments...) L'énoncé est alors traité. **l'interprétation contextuelle** génère un schéma sémantique en adéquation avec l'état du dialogue. Le **module de décision** réinterprète ce schéma selon un historique plus large en se fondant sur le modèle de la tâche et le modèle de dialogue. Si ce module "décide" que l'énoncé correspond à une possible recherche dans une base de données, celle-ci est effectuée. Sinon, la requête est traitée par le **module d'incitation** (cf. tableau 2). Ces deux modules génèrent enfin un schéma sémantique qui est envoyé au **module de génération** en langue naturelle.

Table 3: Corpus RITEL

dialogues	369	mots distincts U	1993	Thèmes	767
durée totale U	3h40	moyen énoncés U / dial.	9	Sous-Thèmes	701
énoncés U	3300	durée moyenne U / dial.	34s.	marqueurs syntaxiques	3220
mots U	32634	Entités Nommées	8174		

3 Corpus

Le corpus (Table 3) a été collecté entre septembre 2004 et janvier 2005. 13 personnes ont appelé le serveur. Ces personnes ont reçu chacune une liste différente d'environ 300 questions. Il leur était demandé de chercher à obtenir une information et d'utiliser pour cela les moyens qu'elles souhaitaient. Il leur était précisé qu'elles ne devaient pas lire les questions qu'elles avaient en exemple et qu'elles pouvaient en choisir d'autres plus proches de leurs intérêts.

Conclusion - Perspectives

Actuellement la plateforme permet une interaction naturelle, quoique limitée, entre un utilisateur qui recherche des informations et le système. Elle a permis de collecter un corpus réaliste et riche d'un point de vue linguistique. Une recherche d'information (minimaliste) est possible, puisque une partie des dialogues a abouti à une réponse. Cette première étude nous permet de dégager les points sur lesquels nos travaux futurs vont porter : reconnaissance vocale temps-réel en flux sur vocabulaire large dynamiquement extensible et en domaine ouvert avec adaptation des modèles à l'interlocuteur tout au long de l'interaction, gestion de dialogue en domaine ouvert et gestion de l'information multi-niveau retourné par le système de recherche d'information, recherche d'information(classification/présentation des réponses suivant l'état du dialogue), types d'information nécessaire pour permettre au dialogue et aux autres modules d'évaluer leurs analyses et réponses suivant l'état du dialogue et des résultats de la recherche d'information, étude du coût des différentes stratégies pour le lancement de la recherche d'information (en continu ou sur décision du gestionnaire de dialogue), fonctionnement en parallèle de la recherche d'information et du dialogue, génération de la réponse, résumé automatique, etc...

Références

EQueR, ELDA (2003) - Evaluation de systèmes de Question-Réponse, http://www.elda.org/article118.html

Gauvain J. L. et Lamel L. F. (2002), Systèmes de reconnaissance, de compréhension et de dialogue, *Reconnaissance de la parole Traitement automatique du langage parlé*, Hermes Lavoisier, J. Mariani.

Glass J. R. et al. (2000), Data collection and performance evaluation of spoken dialogue systems : the MIT experience, Actes de *ICSLP'00*, Pekin, Chine.

Grau B. (2005), Les systèmes de question-réponse, *M*éthodes avancées pour les systèmes de recherche d'informations, sous la direction de Madjid Ihadjadene, collection Traité des sciences et techniques de l'information, Hermes-science.

S. Harabagiu, et al., (2001), The Role of Lexico-Semantic Feedback in Open-Domain Textual Question-Answering, Actes de *Association for Computational Linguistics*

L. Lamel et al., (2000), The LIMSI ARISE System, *Speech Communication* Vol. 31(4):339-354.

Devillers L. et al. (2004), The French MEDIA/EVALDA Project: the Evaluation of the Understanding Capability of Spoken Language Dialogue Systems, *L*REC'04

J. Villaneau (2003), Contribution au traitement syntactico-pragmatique de la langue naturelle parlée : approche logique pour la compréhension de la parole, Thèse de Doctorat, Université de Bretagne Sud.

Un système de génération automatique de dictionnaires linguistiques de l'arabe

Ahmed HADDAD (1), Mounir ZRIGUI (2), Mohamed Ben AHMED (3)

(1) Laboratoire RIADI (unité de Monastir), Faculté des Sciences de Monastir

ahmed.haddad@ensi.rnu.tn

(2) Laboratoire RIADI (unité de Monastir), Faculté des Sciences de Monastir

mounir.zrigui@fsm.rnu.tn

(3) Laboratoire RIADI, Ecole Nationale des Sciences Informatiques

mohamed.benahmed@riadi.rnu.tn

Mots-clés : dictionnaires électroniques, Conditions de Structures Morphématiques, matrices lexicales, Restrictions combinatoires, Restrictions séquentielles.

Keywords: electronic dictionaries, Conditions of Morphemic Structures , lexical matrix, Combinative circumscriptions, Sequential circumscriptions.

Résumé L'objectif de cet article est la présentation d'un système de génération automatique de dictionnaires électroniques de la langue arabe classique, développé au sein du laboratoire RIADI (unité de Monastir). Ce système entre dans le cadre du projet "oreillodule": un système embarqué de synthèse, traduction et reconnaissance de la parole arabe.

Dans cet article, nous présenterons, les différentes étapes de réalisation, et notamment la génération automatique de ces dictionnaires se basant sur une théorie originale : les Conditions de Structures Morphématiques (CSM), et les matrices lexicales.

Abstract the objective of this article is the presentation of a system of automatic generation of electronic dictionaries of the classic Arabian language, developed within RIADI laboratory (unit of Monastir). This system enters in the setting of project "oreillodule": an embedded system of synthesis, translation and recognition of the Arabian word.

In this article, we will present the different stages of realization, and notably the automatic generation of these dictionaries basing on an original theory: Conditions of Morphemic Structure (CSM), and the lexical matrixes.

1 Introduction

Plusieurs linguistes ont poussé l'idée de génération de lexique à partir des conditions de structures morphématiques(CSM), comme Cantinau et Greenberg (HABAILI, 1976). Cette idée est à la base de ce travail où nous présenterons un système de génération automatique des dictionnaires arabes, en se basant sur les CSM et les matrices lexicales (ML), leurs structures et leurs modes d'accès (ZAAFRANI, 2004) (SILBERZTEIN, 1993). Nous focalisons sur le

dictionnaire des racines admissibles et attestées. Nous appliquerons des procédures autant que possible automatisées, pour engendrer un maximum d'entrées et d'informations et éliminer les bruits : l'intervention manuelle des spécialistes de la langue s'avère nécessaire dans certains cas bien déterminés et limités pour éliminer ces derniers.

2 Base théorique :

2.1 Les conditions de structures morphématiques (CSM)

Les phonèmes de l'arabe sont liés à des restrictions combinatoires et des restrictions séquentielles très strictes qui sont énoncées sous la forme de CSM. Ces conditions sont des règles qui régissent la génération des mots dans la langue arabe : un mot qui enfreint une condition ne peut pas appartenir à l'arabe (HABAILI, 1976).

Cadre théorique:

Soit x l'ensemble des traits possibles définis par la théorie linguistique.

Soit C l'ensemble des 28 consonnes de la langue arabe. Soit $C_1C_2C_3$ une racine trilitère, avec C_1, C_2 et C_3 $\in$ C. Soit MP[j] [k] la matrice phonologique (avec $1 \leq j \leq 14$ et $1 \leq k \leq 28$) cette matrice représente l'ensemble des traits des consonnes de l'arabe. Soit V l'ensemble des 6 voyelles de la langue arabe. Soit $C_1V_1C_2V_2C_3V_3$ une racine trilitère voyellée, avec V_1, V_2 et V_3 $\in$ V. Soit MPv[j][k] la matrice phonologique des voyelles (avec $1 \leq j \leq 14$: l'ensemble des traits des voyelles de l'arabe et $1 \leq k \leq 6$)

Les linguistes dénombrent cinq CSM qui régissent la formation des mots arabes. Ces conditions sont classées en deux types: les restrictions combinatoires et les restrictions séquentielles.

2.1.1 Restrictions combinatoires :

Ces restrictions régissent les spécifications des traits correspondant aux phonèmes de la langue arabes. Dans ce cas trois règles sont à énoncer :

1) CSM1 : tous les phonèmes sont [-aspirés]

Tout phonème de l'arabe est une colonne de x spécifications correspondant à ces x traits, les (x-quatorze) spécifications qui ne sont pas représentées découlent automatiquement des quatorze présentes en vertu de conditions propres à l'arabe classique. La condition CSM1 distingue l'arabe classique de nombreuses langues naturelles qui opposent phonèmes aspirés et non aspirés. C'est l'existence de telles restrictions valables pour tous les phonèmes de l'arabe classique, qui a permis de ne faire figurer que quatorze traits (HABAILI, 1976), parmi x traits possibles définis par la théorie linguistique.

$$\text{Si } ci \in C \text{ et } ci \subset C_1C_2C_3 \text{ (avec } 1 \leq i \leq 28) \text{ alors MP[aspiré][i] = [0] .} \quad (1)$$

2) CSM2 : tous les phonèmes vocaliques sont [-nasal]

La condition CSM2 exclut les voyelles nasales de l'inventaire des phonèmes de l'arabe classique.

Si vi $\in$ V et vi $\subset$ $C_1V_1C_2V_2C_3V_3$ (avec $1 \leq i \leq 6$) alors MPv[nasale][i] = [0] . **(2)**

3) CSM3 : tous les phonèmes qui sont [+consonantiques] sont aussi [-syllabiques]

La condition CSM3 exclut les consonnes [+syllabiques]. Cette règle est formulée de la manière suivante:

Si MP[consonante][i] = [-] alors MP[syllabique][k] = [0] . **(3)**

Outre les restrictions combinatoires entre les valeurs des traits appartenant à un même segment, il existe aussi des restrictions séquentielles.

2.1.2 Restrictions séquentielles

Ce sont des restrictions qui lient les spécifications de traits appartenant à des segments successifs de la matrice de l'arabe classique, ces restrictions reflètent le fait que n'importe quelle séquence de phonèmes de l'arabe n'est pas un morphème-racine ou un allomorphe possible (variante combinatoire d'un phonème). Par exemple مدّ et كجب sont des séquences de consonnes permises par la structure de la langue, mais pas " خخذ " (SAIDANE, 2004).

Le fait qu'il n'existe aucun morphème-racine dont la représentation phonologique soit "كجب" n'est la séquence d'aucune contrainte structurelle, il s'agit seulement d'une lacune accidentelle : Il s'agit d'une combinaison admissible par la structure de la langue, mais qui est absente du lexique. En revanche, des séquences telles que "خخد" ou "نبذ" ne sont pas des morphèmes-racines possibles en arabe classique. La première enfreint la restriction qui est exprimée par la condition CSM4 et la seconde celle qui est exprimée par CSM5 :

CSM4 : La condition CSM4 exclut de l'ensemble des morphèmes-racines possibles en arabe classique toute séquence de phonèmes formée de deux segments identiques, en première et en deuxième consonne radicale.

Si c,d $\in$ C et c,d $\subset$ $\{C_1C_2C_3\}$ (tel que $c = C_1$ et $d = C_2$) alors ($c \neq d$). **(4)**

CSM5 : La condition CSM5 interdit des consonnes identiques qui sont [+continu, +voisé] en première et troisième consonnes radicales.

Si [(MP[continu][i] = [+], MP[voisé][i] = [+]) et (MP[continu][l] = [+], MP[voisé][l] = [+]) alors ci,dl $\subset$ $C_1C_2C_3$. **(5)**

Puisque les CSM n'opèrent que sur chaque allomorphe pris isolément, elles ne rendent compte que des contraintes à l'intérieur d'un même morphème. D'où, le processus de génération des verbes nécessite l'utilisation d'autres outils, qui sont les matrices lexicales (MOUSSA, 1973).

2.2 Matrices Lexicales

2.2.1 Matrices Lexicales Trilitères (MLT)

Ce sont des matrices bidimensionnelles qui représentent la position des consonnes dans une racine trilitère, ces matrices sont extraites de la référence "تاج العروس ", avec quelques

transformations afin de l'utiliser dans ce travail (HADDAD, 2004). Aux 28 consonnes de la langue arabe correspondent 28 MLT. Les 28 matrices sont issues d'une statistique élaborée par Amr Helmi MOUSSA sur le dictionnaire تاج العروس, en transformant les racines trilitères du dictionnaire en des matrices décrivant les racines attestées par ce dictionnaire.

Ce sont des matrices binaires Mi, avec $1 \leq i \leq n$ (n = 28 : nombre des consonnes).

Mi [j][k] exprime les raines $C_iC_jC_k$ (avec i,j et k Є [1..28])(exemple كتب KTB), tel que :

Mi[][] indique la lettre qui est en première position dans la racine $C_iC_jC_k$ (ك) K

Mi[j][] indique la lettre qui est en deuxième position dans la racine $C_iC_jC_k$ (ت) T

Mi[][k] indique la lettre qui est en troisième position dans la racine $C_iC_jC_k$ (ب) B

Nous distinguons les cas suivants :

- Si Mi [j][k]= 1 alors la racine $C_iC_jC_k$ est une racine attestée par le dictionnaire تاج العروس (exemple كتب)
- Sinon (Mi [j][k]= 0) alors la racine $C_iC_jC_k$ n'est pas attestée par le dictionnaire " تاج العروس ".(exemple ظضذ).

Nous pouvons schématiser comme suit cette représentation:

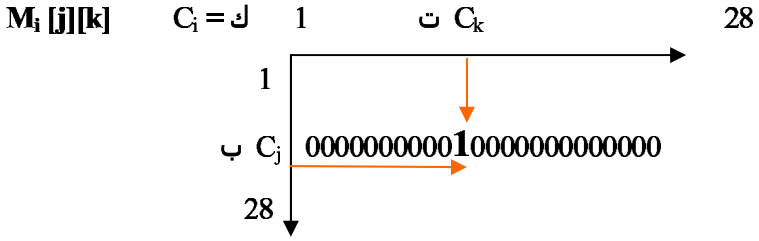


Figure 1 : Représentation de la matrice lexicale

2.2.2 Matrices Lexicales Quadrilitères (MLQ)

Les matrices lexicales quadrilitères sont des matrices bidimensionnelles qui représentent la position des consonnes dans une racine quadrilitère. En s'inspirant de "تاج العروس "، et du "الشامل في تصريف الأفعال العربيّة "، nous avons pu établir 28 matrices comme suit :

Soit Mi une matrice, avec $1 \leq i \leq 28$. Soit Q une représentation d'une racine quadrilitère quelconque attestée par la langue arabe, soit C1C2C3 une représentation d'une racine trilitère attestée et qui a donnée la racine quadrilitère Q, avec C_1, C_2 et C_3 Є C. Mi [j][k] exprime les racines $C_iC_jC_k$ (avec i, j et k Є [1..28])

Ces matrices bidimensionnelles sont formulées de la manière suivante :

- Si Mi [j][k]= 1 alors la racine $C_iC_jC_k$ est une racine attestée et Q est la racine quadrilitère générée de $C_iC_jC_k$ par dérivation avec le schème "فاعل", comme "كاتب".
- Si Mi [j][k]= 2 alors la racine $C_iC_jC_k$ est une racine attestée et Q est la racine quadrilitère générée de CiCjCk par dérivation avec le schème "فعّل", comme "بعّد".
- Si Mi [j][k]= 3 alors la racine $C_iC_jC_k$ est une racine attestée et Q est la racine quadrilitère générée de $C_iC_jC_k$ par dérivation avec le schème "أفعل" , comme " أبعد".
- Si Mi [j][k]= 4 alors la racine $C_iC_jC_k$ est une racine attestée et Q est la racine quadrilitère générée de CiCjCk par dérivation avec le schème "فعلل" , comme "زلزل".
- Si Mi [j][k]= x, avec x Є [أ ب ج د...ه و ي] , alors Q = $C_iC_jC_k$ x , comme "حوقل".
- Sinon (Mi [j][k]= 0) alors la racine $C_iC_jC_k$ n'est pas attestée par le dictionnaire " " تاج العروس.

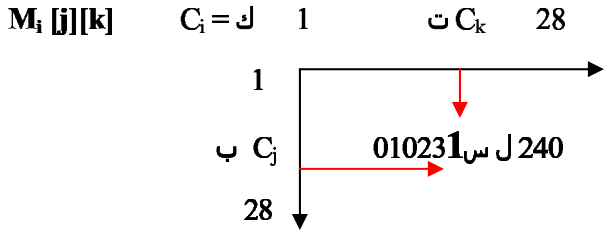


Figure 2 : Représentation de la matrice lexicale quadrilitère

3. Description du système réalisé :

3.1. Les dictionnaires :

Le système développé est composé de deux fonctions qui sont :

1. la génération des dictionnaires,
2. la consultation des dictionnaires.

3.1.1. La génération automatique de cinq dictionnaires de racines trilitères et quadrilitères arabes :

- Le premier dictionnaire est théorique (21952 racines = (28)^3). Il contient toutes les racines trilitères théoriquement possibles del'arabe standard.
- Le deuxième dictionnaire(20415 racines) : c'est le dictionnaire des racines trilitères admissibles. C'est-à-dire les racines qui n'enfreignent aucune des (CSM).
- Le troisième dictionnaire (7836): c'est le dictionnaire des racines trilitères attestées ; c'est-à-dire utilisées dans la langue arabe et qui sont tirées des tableaux de répartitions construits à partir du grand dictionnaire arabe(الصحاح لإبن الجوهري).
- Le Quatrième dictionnaire (13023 racines): c'est le dictionnaire des racines admissibles par la langue arabe mais non attestées. Ces racines peuvent être utilisées pour enrichir la langue arabe par d'autres mots nouveaux.
- Le cinquième dictionnaire (4000 racines) : c'est le dictionnaire des racines quadrilitères attestées ; qui sont tirées des matrices lexicales quadrilitéres.

Le schéma suivant illustre le diagramme de données relatif à la génération automatique des différents dictionnaires :

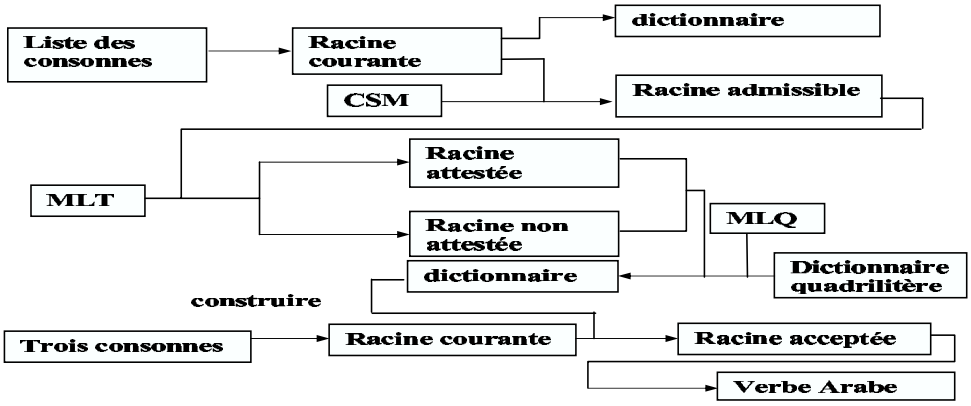


Figure 3 : Diagramme de génération des différents dictionnaires arabes

Certaines racines trilitères attestées n'obéissent pas à une ou plusieurs CSM : nous avons créé un sixième dictionnaire (203 racines) qui regroupe ces racines, avec pour chacune, l'affichage de la CSM qui n'est pas vérifiée. Exemple : la racine (ببب) est attestée mais ne vérifie pas la condition CSM4.

3.1.2. La consultation de ces dictionnaires dans le but, de la recherche d'une racine, ou l'affichage d'une liste de racines d'un dictionnaire bien déterminé, ou sa mise à jour.

3.2 Structure interne des données du dictionnaire électronique :

Pour réaliser le dictionnaire capital, nous avons adopté la structure de listes chaînées. Cette structure permet un accès simple et une recherche facile des informations. La figure suivante décrit cette représentation interne ainsi que les différents niveaux des données :

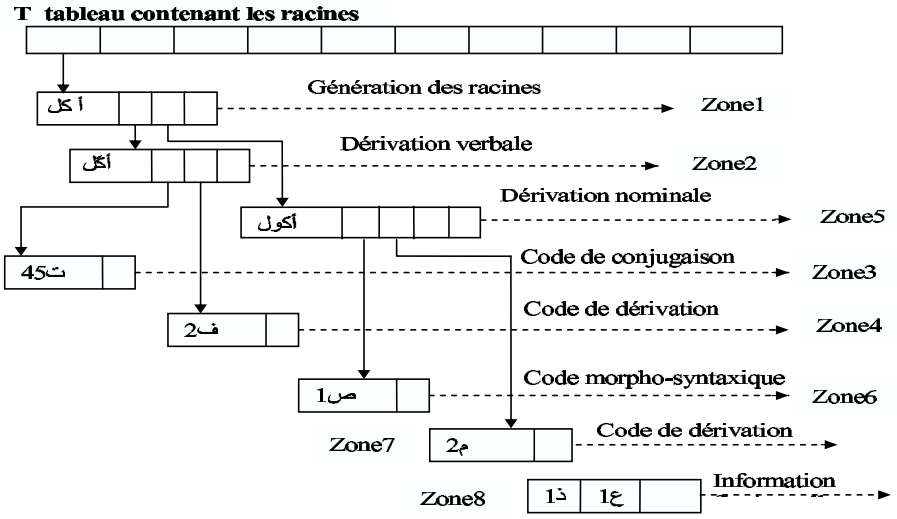


Figure 4 : Structure interne des données

Conclusion

La génération automatique du dictionnaire des racines trilitères et quadrilitères en utilisant les CSM et les ML fait l'originalité de ce travail. Ce dictionnaire sera à la base de toute analyse morpho-syntaxique de l'arabe, il regroupe les racines du grand dictionnaire (معجم الصحاح), auquel on peut ajouter d'autres dictionnaires.

Le dictionnaire capital, resultat de ce système, contient 36876216 verbes et mots de l'arabe : ce dictionnaire est généré automatiquement à la demande de l'utilisateur donc ne pose pas de problèmes d'encombrement en mémoire.

Références

M. SILBERZTEIN, (1993).Dictionnaires électroniques et analyse automatique de textes (Le système INTEX). (Masson, Paris)

H. HABAILI, (1976). Contraintes de structure morphématique en Arabe, DEA en linguistique, Canada, université de Montréal.

A. HADDAD, (2004). Un système de génération automatique de dictionnaires linguistiques et thématiques de la langue arabe. Mastère en informatique, Ecole Nationale des Sciences de l'informatique, TUNISIE.

A. H. MOUSSA, (1973). Statistical study of Arabic roots in moijam arous. Kouyet .

R ZAAFRANI,(2004). Un dictionnaire électronique pour apprenant de l'arabe (langue seconde) basé sur corpus. JEP-TALN 2004, Fès, Maroc.

T.SAIDANE, A.HADDAD, M.ZRIGUI, Pr. M. BEN AHMED, (2004). Réalisation d'un système hybride de synthèse de la parole arabe utilisant un dictionnaire de polyphones JEP-TALN 2004, Fès, Maroc.

Segmentation de textes arabes basée sur l'analyse contextuelle des signes de ponctuations et de certaines particules

Lamia Hadrich Belguith (1), Leila Baccour (1) et Ghassan Mourad (2)

(1) Laboratoire de recherche LARIS – Faculté des Sciences Economiques et de Gestion de Sfax
B.P. 1088, 3018, Sfax, Tunisie
l.belguith@fsegs.rnu.tn
leila_freind@techemail.com
(2) Equipe LaLICC – Paris Sorbonne
96, Bd Raspail, 75006 Paris
Ghassan.Mourad@paris4.sorbonne.fr

Mots clés : Segmenteur de textes arabes, segmentation en phrases, exploration contextuelle, expressions rationnelles.

Keywords: Arabic text tokenizer, sentence tokenization, contextual exploration, regular expressions.

Résumé Nous proposons dans cet article une approche de segmentation de textes arabes non voyellés basée sur une analyse contextuelle des signes de ponctuations et de certaines particules, tels que les conjonctions de coordination. Nous présentons ensuite notre système STAr, un segmenteur de textes arabes basé sur l'approche proposée. STAr accepte en entrée un texte arabe en format txt et génère en sortie un texte segmenté en paragraphes et en phrases.

Abstract We propose in this paper an approach to segment non-vowelled Arabic texts. Our approach is based on a contextual analysis of the punctuation marks and a list of particles, such as the coordination conjunctions. Then, we present our system STAr, a tokenizer based on the proposed approach. The STAr input is an Arabic text (in .txt format) and its output is a segmented text into paragraphs and sentences.

Introduction

Pour la plupart des applications de traitement automatique des langues naturelles (e.g., l'analyse de texte, l'extraction d'information, le résumé automatique) la segmentation devient une phase importante pour repérer les segments contenant les informations recherchées. Ainsi par exemple, commencer une analyse d'un texte sans le segmenter en phrases conduit à des résultats peu fiables; de même, avoir un mauvais segmenteur conduit à accumuler les erreurs du traitement automatique du texte (Mourad, 2001).

La segmentation consiste à désambiguïser les frontières des phrases et des paragraphes et se base généralement sur un ensemble de règles de segmentation. C'est une phase non triviale pour toute application en TALN. En effet, segmenter un texte nécessite le repérage des frontières formelles marquées par des signes typographiques. Par ailleurs, dans les textes arabes actuels, les signes de ponctuation ne sont pas très utilisés et dans le cas où ils y figurent, ils ne sont pas gérés par des règles d'utilisation. De plus, d'après l'observation de corpus, nous avons constaté que certaines particules (e.g., "و" (et), "ف" (donc)) jouent un rôle principal dans la séparation de phrases.

Dans ce qui suit, nous présentons un bref aperçu sur les travaux de segmentation. Ensuite, nous détaillons les difficultés rencontrées lors de la segmentation des textes arabes. Nous proposons, ensuite, notre approche de segmentation de textes arabes. Après, nous présentons le système STAr, un segmenteur de textes arabes basé sur l'approche proposée. Enfin, nous présentons l'évaluation de STAr.

1 Bref aperçu sur les travaux de segmentation

Les travaux sur la segmentation ne sont pas nombreux. Pour certaines langues latines, ils existent des segmenteurs fonctionnels. Alors que pour l'arabe, il y a peu de travaux sur la segmentation de textes en phrases et il n'existe pas des segmenteurs fonctionnels et spécifiques à l'arabe.

Dans ce qui suit nous présentons quelque segmenteurs pour le français et l'anglais.

- Le segmenteur INTEX (Silberztein, 93) utilise un transducteur pour découper un texte français en phrases en s'appuyant sur les signes de ponctuation.
- Le segmenteur SATZ (Palmer, Hearst, 1994) de textes anglais utilise les catégories lexicales au voisinage des signes de ponctuation et applique une méthode d'apprentissage en utilisant les réseaux de neurones.
- Le système SegATex (Mourad, 2001) est un segmenteur de textes français qui conçoit des règles de segmentation en étudiant les voisinages des signes de ponctuation et des marques typographiques en appliquant la méthode d'exploration contextuelle (DESCLES, 1997).

2 La segmentation de textes arabes : particularités et difficultés

La segmentation automatique de textes arabes présente plusieurs difficultés spécifiques à la langue arabe. Nous présentons dans ce qui suit certaines ambiguïtés qui rendent la segmentation difficile à réaliser sans une étude approfondie sur un corpus à large couverture.

- L'ambiguïté vocalique des mots : un texte arabe non voyellé est fortement ambigu. La proportion des mots ambigus passe à plus de 90% si les comptages portent sur les voyellations globales de ces mots (Debili, Achour, Souissi, 2002). Ainsi, un mot non voyellé peut avoir plusieurs caractéristiques morphologiques possibles (Chaâben, Belguith, 2003). Par exemple le mot "فهم" peut être un nom, un verbe, ou un pronom personnel précédé d'une conjonction de coordination.
- L'ambiguïté dérivationelle : le mot arabe n'est pas le résultat d'une simple concaténation de morphèmes comme c'est le cas pour l'anglais mais c'est à partir d'une racine, d'une combinaison de voyelles, de préfixes, d'infixes, de suffixes et d'un schème morphologique qu'on obtient un mot (Beesley, 1996). Ainsi, l'identification de la catégorie grammaticale de certains mots est ambiguë ce qui entraîne des difficultés au niveau de la segmentation automatique.
- L'ambiguïté structurelle : la phrase arabe est relativement longue et complexe en comparaison avec d'autres langues, tels que le français ou l'anglais. Ainsi il n'est pas rare de trouver des phrases arabes composées de plusieurs dizaines de mots.
- L'utilisation des signes de ponctuation : l'arabe n'est pas appuyée principalement sur les signes de ponctuations et les marqueurs typographiques; ces derniers ont généralement un rôle pausale. Ainsi, nous pouvons trouver tout un paragraphe arabe ne contenant aucun signe de ponctuation à part un point à la fin de ce paragraphe.
- L'agglutination : les conjonctions de coordinations jouent un rôle important dans la segmentation de textes arabes. Cependant, elles sont toujours agglutinées aux mots qui les suivent. Ainsi par exemple la lettre "و" (w) dans le mot "وهم" peut représenter une lettre du mot en question (i.e., « wahmun » (imagination)) ou une conjonction de coordination suivie d'un pronom personnel (i.e., "و"+ "هم" « wa+hum » (et + ils)).

3 Approche proposée pour la segmentation de textes arabes

Afin de surmonter les problèmes de segmentation que nous venons de présenter, nous proposons une approche de segmentation de textes arabes basée sur l'exploration contextuelle des signes de ponctuation, des mots connecteurs jouant le rôle de séparateur de phrases (e.g., "لكن" (lakin), "لقد" (laqad) et "أمّا" ('amma)) ainsi que celles de certaines particules tel que les conjonctions de coordination ("و" (wa) et "ف" (fā)).

L'exploration contextuelle repose sur une étude des indices linguistiques déclencheurs appelés indicateurs et des indices complémentaires associés à ces indicateurs et sur un ensemble de règles (Descles, 1991). Ainsi, nous proposons d'utiliser l'exploration contextuelle pour étudier les contextes droit et gauche de chaque mot ou particule jouant le rôle de séparateur de phrases. Pour ce faire, nous avons étudié un corpus de textes issus de quatre livres de l'enseignement tunisien (voir figure 1).

Corpus	Nombre de textes	Nombre de paragraphes	Nombre de mots
Livre de 5ème année primaire	70	618	19 254
Livre de 6ème année primaire	72	173	18 317
Livre de 7ème année de base	65	202	20 221
Livre de 8ème année de base	73	256	25 886
Total	*279*	*1 249*	*82 678*

Figure 1 : Corpus utilisé pour la conception des règles de segmentation

L'étude de ce corpus (segmenté manuellement par des linguistes) nous a permis de concevoir 183 règles de segmentation. Ces règles ont le format suivant :

Soit un marqueur déclencheur X SI le contexte gauche de X est G ET/OU SI le contexte droit de X est D ALORS prendre la décision Y (fin ou non fin d'un segment)

Figure 2 : Format de règles conçues pour la segmentation de textes (Mourad, 2001)

Ces règles peuvent être classées en trois classes relatives aux trois types de marqueurs déclencheurs à savoir les signes de ponctuation, les particules et les mots connecteurs (Baccour, Mourad, Belguith Hadrich, 2003). Nous présentons dans ce qui suit un exemple d'une règle relative à la virgule.

Contexte gauche			Marqueur	Contexte droit		
	Verbe	Espace	,			وفي صباح

SI	la virgule est suivie par un espace
ET Si	l'espace est suivi d'un verbe
ET SI	le contexte droit de la virgule commence par " وفي صباح"
ALORS	la virgule ne marque pas la fin de la phrase

C'est le cas par exemple de l'énoncé suivant :

وفي صباح مشرق من أصباح الصّيف، مرّ بابن عمّه إسماعيل.

Et à une des matinées ensoleillées de l'été, il a passé à son cousin Ismail.

4 Présentation du système STAr

STAr (voir figure 3) est un segmenteur de textes arabes basé sur l'approche de segmentation

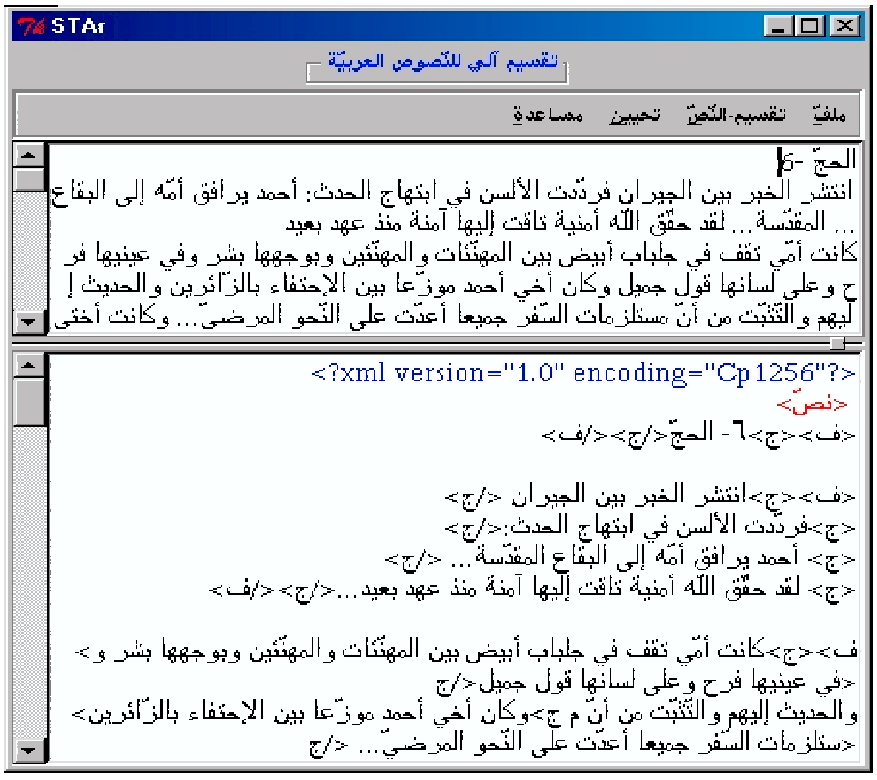


Figure 3 : Un exemple d'exécution de STAr

proposée. Il est réalisé avec le langage de programmation Perl. Il accepte en entrée un texte arabe en format txt et génère en sortie un texte segmenté en paragraphes et en phrases.

La figure 3 montre un texte segmenté par STAr. Dans le premier éditeur, figure le texte source (texte à segmenter de type .txt) et dans le deuxième éditeur figure le texte segmenté par STAr. Ce texte est généré dans un fichier XML. Les balises <نصّ> et </نصّ> indiquent le début et la fin d'un texte, les balises <ف> et </ف> représentent le début et la fin d'un paragraphe et les balises <ج> et </ج> représentent le début et la fin d'une phrase.

5 Evaluation de STAr

L'évaluation du système STAr a été réalisée sur deux corpus différents (voir figure 4).

Corpus	*Nombre de textes*	*Nombre de paragraphes*	*Nombre de mots*
Deux livres (4ème année primaire et 9ème année de base)	*144*	*991*	*40 3431*
Articles de journaux	60	510	38 062

Figure 4 : Corpus d'évaluation de STAr

Les mesures de rappel et de précision obtenues pour le premier corpus sont meilleures que ceux trouvés pour le deuxième corpus (voir figure 5). Ceci s'explique par le fait que les articles de journaux contiennent des erreurs typographiques (i.e. insertion d'un espace après la conjonction de coordination "و" (wa), omission de la lettre "الشّدّة" ('chadda), des constructions erronées, etc.) qui augmentent le taux d'erreur au niveau de la segmentation en mots, de l'identification de la catégorie grammaticale des mots et par conséquent le taux d'erreur au niveau de la segmentation en phrases augmente.

Corpus	*Rappel*	*Précision*
Livres	88.26%	80.65%
Articles de journaux	75.81%	65.66%

Figure 5 : Les mesures de rappel et de précision obtenues pour les deux corpus d'évaluation

7 Conclusion et perspectives

Dans ce papier nous avons proposé une approche de segmentation de textes arabes non voyellés qui se base sur l'analyse contextuelle des signes de ponctuation, de certaines particules et certains mots connecteurs.

Nous avons aussi présenté notre système STAr, un Segmenteur de Textes Arabes, basé sur l'approche proposée. STAr est actuellement intégré dans le système MASPAR (Multi Agent System for Parsing Arabic) d'analyse de textes arabes non voyellés (Aloulou, Belguith, Ben Hamadou, 2000), (Aloulou, Belguith, Hadj Kacem, Ben Hamadou, 2004). Ce système est composé de 5 agents (segmentation, morphologie, syntaxe, ellipse, anaphore) (Aloulou, Belguith, Hadj Kacem, Hammami, 2003). Ainsi STAr est intégré dans MASPAR en tant qu'agent pour la segmentation de textes en phrases et pourrait collaborer avec l'agent morphologie qui a pour objectif de déterminer pour chaque mot sa catégorie grammaticale ainsi que ses caractéristiques morphologiques (genre, nombre, temps, personne, etc.) (Belguith Hadrich, Ben Hamadou, 2004).

Comme perspectives, nous envisageons d'étudier la collaboration de STAr avec l'agent syntaxe. En effet, certaines ambiguïtés de segmentation ne peuvent être levées qu'à l'aide d'informations syntaxiques. De plus certaines particules utilisées dans la segmentation

peuvent à leur tour être ambiguës. Ainsi, nous envisageons de faire une étude approfondie de ces cas d'ambiguïtés.

Références

Aloulou C., Belguith Hadrich L., Hadj Kacem A., Ben Hamadou A., (2004), Conception et développement du système MASPAR d'analyse de l'Arabe selon une approche agent, *14ème Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle*, du 28 au 30 janvier 2004 à Toulouse - France.

Aloulou C., Belguith Hadrich L., Ben Hamadou A., (2000), Vers un système d'analyse syntaxique robuste pour l'Arabe: Application au recouvrement des erreurs de la reconnaissance, *7ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'2000)*, 16 – 18 octobre 2000, Lausanne, SUISSE.

Blachère R., Gaudefroy-Demombynes M. (1975), *Grammaire de l'arabe classique*, Éditions Maisonneuve & Larose 15, rue Victor-cousin 75005 Paris.

Beesley K. (1996), Arabic Finite-State Morphological Analysis and Generation; *COLING96*, Vol. 1, pages 89-94.

Belguith Hadrich L., Ben Hamadou A. (2004), Traitement des erreurs d'accord : une analyse syntagmatique pour la vérification et une analyse multicritère pour la correction, *Revue d'Intelligence Artificielle (RSTI- RIA)*, Hermès-Lavoisier, Vol. 18, N 5 et 6.

Belguith Hadrich L. (1999), *Traitement des erreurs d'accord de l'Arabe basé sur une analyse syntagmatique étendue pour la vérification et une analyse multicritères pour la correction*, Thèse de doctorat en informatique, Faculté des Sciences de Tunis.

Baccour L., Mourad G., Belguith Hadrich L. (2003), Segmentation de textes arabes en phrases basée sur les signes de ponctuation et les mots connecteurs, *troisième journées scientifiques des jeunes chercheurs en génie électrique et informatique*, du 25-27 mars, Mahdia, Tunisie.

Chaâben N., Belguith Hadrich L (2003), L'étiquetage morpho-syntaxique: Comment lever l'ambiguïté dans les textes arabes non voyellés *?, troisième journées scientifiques des jeunes chercheurs en génie électrique et informatique*, du 25-27 mars, Mahdia, Tunisie.

Descles J.-P., (1997), *Systèmes d'exploration contextuelle. Co-texte et calcul du sens.* , éd. Claude Guimier, Presses Universitaires de Caen, pp. 215-232.

Debili F., Achour H., Souissi E. (2002), La langue arabe et l'ordinateur, de l'étiquetage grammatical à la voyellation automatique, *Correspondances n° 71 juillet-août 2002.*

Hammami S., Aloulou C., Belguith Hadrich L., Hadj Kacem A. (2003), Implémentation du système MASPAR selon une approche multi-agent*, IWPT'03 (International Workshop on Parsing Technologies),* 23-25 avril 2003, Nancy, France.

Mourad G. (2001), *Analyse informatique de signes typographiques pour la segmentation de textes et l'extraction automatique des citations*, Thèse de doctorat en informatique linguistique, université de Paris - Sorbonne.

Palmer D., Hearst M. (1994), Adaptive sentence boundary disambiguation, *Report No. UCB/CSD 94/797*, Computer Science Division (EECS), University of California, Berkeley, California 94720.

Silberztein M. (1993), Dictionnaires électroniques et analyse automatique de textes, Le système INTEX, Paris, Masson.

A Descriptive Characterization of Multicomponent Tree Adjoining Grammars

Laura Kallmeyer
TALaNa/Lattice, UFRL, University Paris 7
2 place Jussieu, Case 7003, 75005 Paris
laura.kallmeyer@linguist.jussieu.fr

Mots-clefs : Grammaires d'Arbres Adjoints, MCTAG, formalismes grammaticaux

Keywords: Tree Adjoining Grammars, MCTAG, grammar formalisms

Résumé Il a été montré que les Grammaires d'Arbres Adjoints Ensemblistes (Multicomponent Tree Adjoining Grammars, MCTAG) sont très utiles pour des applications TAL. Pourtant, la définition des MCTAG est problématique parce qu'elle fait référence au procès de dérivation même : une contrainte de simultanéité est imposée concernant la façon dont on ajoute les membres d'un même ensemble d'arbres. En regardant uniquement le résultat d'une dérivation, c'est-à-dire l'arbre dérivé et l'arbre de dérivation, cette simultanéité n'est plus visible. Par conséquent pour vérifier la contrainte de simultanéité, il faut toujours considérer l'ordre concret des pas de la dérivation. Afin d'éviter cela, nous proposons une caractérisation alternative de MCTAG qui permet une abstraction de l'ordre de dérivation : Les arbres générés par la grammaire sont caractérisés par les propriétés de leurs arbres de dérivation.

Abstract Multicomponent Tree Adjoining Grammars (MCTAG) is a formalism that has been shown to be useful for many natural language applications. The definition of MCTAG however is problematic since it refers to the process of the derivation itself: a simultaneity constraint must be respected concerning the way the members of the elementary tree sets are added. Looking only at the result of a derivation (i.e., the derived tree and the derivation tree), this simultaneity is no longer visible and therefore cannot be checked. I.e., this way of characterizing MCTAG does not allow to abstract away from the concrete order of derivation. Therefore, in this paper, we propose an alternative definition of MCTAG that characterizes the trees in the tree language of an MCTAG via the properties of the derivation trees the MCTAG licences.

1 Introduction

1.1 Tree Adjoining Grammars

Tree Adjoining Grammar (TAG, Joshi et al., 1975) is a tree-rewriting formalism. A TAG consists of a finite set of trees (*elementary* trees) with nonterminals and terminals as node labels (terminals only label leaf nodes). Starting from the elementary trees, larger trees are derived by

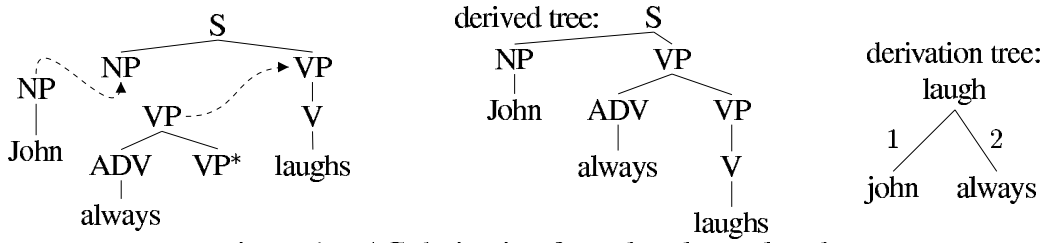


Figure 1: TAG derivation for *John always laughs*

substitution (replacing a leaf with a new tree) and *adjunction* (replacing an internal node with a new tree). In case of an adjunction, the new tree is a so-called *auxiliary* tree that has exactly one leaf marked as the foot node (marked with an asterisk). All other elementary trees are called *initial* trees. When adjoining an auxiliary tree β to a node μ, in the resulting tree, the subtree with root node μ from the old tree is put below the foot node of β. Each derivation starts with an initial tree. In the final derived tree, all leaves must have terminal labels. See for example Fig. 1 : Starting from the *laughs* tree, the tree for *John* is substituted for the NP leaf and the tree for *always* is adjoined at the VP node.

TAG derivations are represented by derivation trees that record the history of how the elementary trees are put together. A derived tree is the result of carrying out the substitutions and adjunctions. Each edge in the derivation tree stands for an adjunction or a substitution. The edges are labelled with Gorn addresses of the nodes where the substitutions/adjunctions take place.[1] E.g., in Fig. 1 the derivation tree indicates that the elementary tree for *John* is substituted for the node at address 1 and *always* is adjoined at node address 2.

1.2 Multicomponent TAG

Multicomponent TAG (MCTAG, Joshi, 1987; Weir, 1988) is a TAG extension useful for linguistic applications. An MCTAG contains sets of elementary trees. Starting with an initial tree, in each derivation step, all trees from one of the tree sets are added simultaneously. Depending on the nodes to which these trees attach, different kinds of MCTAGs are distinguished: if all nodes are required to be part of the same elementary tree, the MCTAG is *tree-local*; if all nodes are required to be part of the same tree set, the grammar is *set-local*; otherwise the grammar is *non-local*.[2] Consider for example the non-local MCTAG derivation in Fig. 2: the tree for *to be certain* adjoins to the lower S node of *like*, the WH and NP nodes of *like* are substituted for *what* and *John* respectively, and *does* and *seem* are adjoined simultaneously to the upper S node of *like* and the root node of *to be certain* respectively. These last two operations cannot be performed before having added *to be certain* to *like*, otherwise the simultaneity requirement is not satisfied.

Intuitively, the requirement of adding all elements of an elementary set simultaneously is easy to understand and this definition of MCTAG seems very clear. However, the simultaneity requirement imposes certain derivation orders even though a different order might lead to the same adjunctions and substitutions and to the same derived tree. E.g., in Fig. 2 one might as well

[1] The root has the address ϵ, and the jth child of the node with address p has address pj.

[2] Cases where MCTAGs have been argued to be useful are extractions out of complex NPs as in "which painting did you buy a copy of" where the two parts of the complex NP should be part of one elementary structure but cannot be part of the same elementary tree. For such examples Kroch and Joshi (1987) propose to use tree-local MCTAGs.

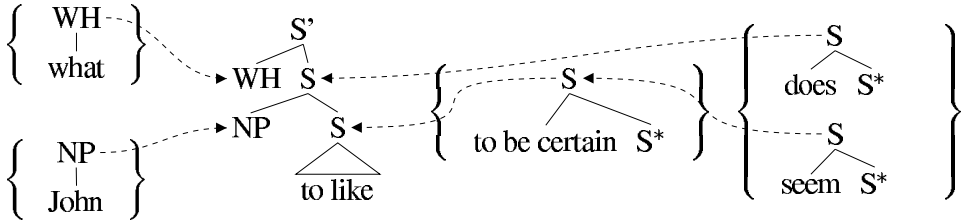


Figure 2: Derivation for *what does John seem to be certain to like* in an MCTAG G_M

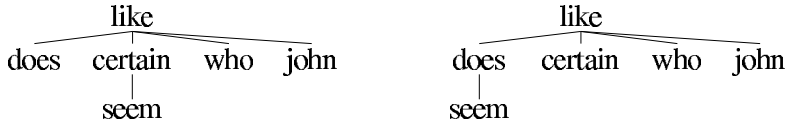


Figure 3: TAG derivation trees for the TAG underlying Fig. 2

start by adding *does* to *like* (at the higher S node), then adjoin *to be certain* to *like* (at the lower S node) and then adjoin *seem* to *to be certain*. This yields the same derived tree with the same adjunctions and substitutions. But the simultaneity requirement is not respected. Consequently, in order to check whether a given tree is part of the tree language, one has to check the possible derivations of this tree including the different derivation orders. In contrast to this, in a TAG it is sufficient to check whether there is a derivation tree yielding the tree in question. I.e., one can abstract away from the order of the derivations steps. E.g., in Fig. 1, no matter in which order *John* and *always* are added, the derivation tree and consequently the derived tree are the same.

For MCTAG as well one would like to abstract away from differences with respect to derivation order that do not make any difference concerning the substitutions and adjunctions that are performed. One way to achieve this is to consider an MCTAG as a TAG G with additional multicomponent tree sets (sets of initial and auxiliary trees from G) where certain derivation trees in G are disallowed since they do not satisfy certain constraints. E.g., the derivation trees in Fig. 3 are both possible in a TAG with the elementary trees from the MCTAG G_M in Fig. 2. The first derivation tree is the one for the derivation from Fig.2. Since we know that only *does* and *seem* are in one set and since *does* and *seem* are dominated by different daughters of *like* (namely *does* and *certain* respectively), this is a possible TAG derivation tree in G_M. The second derivation tree is possible in the underlying TAG but not in G_M: since *seem* adjoins into *does*, it is not possible to add *does* and *seem* simultaneously to different nodes in an already derived tree. With this characterization of MCTAG one gets rid of the problematic simultaneity requirement. Instead, one characterizes in a descriptive way the properties of the derivation trees licensed by the grammar. The advantage of this non-operational perspective is that one needs not to check all possible derivation orders with respect to the simultaneity constraint.

In section 2, standard definitions of TAG and MCTAG are given. Then, in section 3, an alternative descriptive characterization of MCTAG is proposed.

2 Standard definitions of TAG and MCTAG

We assume that the definitions of initial and auxiliary trees and the definitions of substitution and adjunction are already known.[3] a TAG (see, e.g., Vijay-Shanker, 1987) is a tuple $G =$

[3]For formal definitions of initial and auxiliary trees with certain alphabets of nonterminal and terminal symbols and also for formal definitions of the operations substitution and adjunction see for example Kallmeyer (1999).

$\langle I, A, N, T\rangle$ with N abd T being finite sets of nonterminals and terminals, and I and A being finites sets of initial and auxiliary trees with nonterminals N and terminals T.

In a TAG $G = \langle I, A, N, T\rangle$, a *derivation step* is defined as follows: Let γ and γ' be finite trees. $\gamma \Rightarrow \gamma'$ in G iff there is a node position p and a tree $\gamma_0 \in I \cup A$[4] such that $\gamma' = \gamma[p, \gamma_0]$.[5] $\stackrel{*}{\Rightarrow}$ is the reflexive transitive closure of $\Rightarrow$. The *tree language* of G is then $L_T(G) := \{\gamma \mid$ there is an $\alpha \in I$ such that $\alpha \stackrel{*}{\Rightarrow} \gamma$ and all leaves in γ have terminal labels$\}$.

Each node address p in a derived tree points at a node belonging to some elementary tree γ_e. In γ_e this node has some address p_e. In the following we assume that the address p in a derivation step $\gamma \Rightarrow \gamma'$ of the node where the adjunction/substitution takes place is the corresponding tuple $\langle p_e, \gamma_e\rangle$. This is possible since each node in a derived tree in TAG belongs uniquely to one of the elementary trees used in the course of the derivation. E.g., the address of the ADV node in the derived tree in Fig. 1 is $\langle 1, always\rangle$. Using these addresses we can define derivation trees: A *derivation tree* is a tuple $\langle \mathcal{N}, \mathcal{E}\rangle$ of nodes and edges. $\mathcal{N}$ is a finite set of instances of elementary trees and $\mathcal{E} \subset \mathcal{N} \times \mathcal{N} \times \mathbb{N}^*$ where $\mathbb{N}^*$ is the set of Gorn addresses. (The edges are directed from the mother node to the daughter.)[6] For a TAG $G = \langle I, A, N, T\rangle$ and a derivation $\gamma_0 \Rightarrow \gamma_1 \Rightarrow \gamma_2 \cdots \Rightarrow \gamma_n$ in G, the derivation tree $\langle \mathcal{N}, \mathcal{E}\rangle$ is then as follows: $\gamma_0 \in \mathcal{N}$, and for all derivation steps $\gamma_i \Rightarrow \gamma_{i+1}$, $0 \leq i < n$ in the derivation such that there is a node position $\langle p_e, \gamma_e\rangle$ and a tree $\gamma \in I \cup A$ with $\gamma_{i+1} = \gamma_i[\langle p_e, \gamma_e\rangle, \gamma]$: $\gamma \in \mathcal{N}$ and $\langle \gamma_e, \gamma, p_e\rangle \in \mathcal{E}$. These are all nodes and edges. In a derivation tree $D = \langle \mathcal{N}, \mathcal{E}\rangle$, the *parent* relation is the relation between mothers and daughters, $\mathcal{P}_D := \{\langle n_1, n_2\rangle \mid$ there is a $p \in \mathbb{N}^*$ such that $\langle n_1, n_2, p\rangle \in \mathcal{E}\}$. The *dominance* relation is its reflexive transitive closure, $\mathcal{D}_D := \{\langle n_1, n_2\rangle \mid n_1, n_2 \in \mathcal{N}$ and either $n_1 = n_2$ or there is a n_3 such that $\langle n_1, n_3\rangle \in \mathcal{P}_D$ and $\langle n_3, n_2\rangle \in \mathcal{D}_D\}$.

Finally, we define multicomponent TAG (MCTAG, Joshi, 1987; Weir, 1988): A *multicomponent TAG* (MCTAG) is a tuple $G = \langle I, A, N, T, \mathcal{A}\rangle$ such that: $G_{TAG} := \langle I, A, N, T\rangle$ is a TAG, and $\mathcal{A} \subseteq P(I \cup A)$ is a set of subsets of $I \cup A$, the set of elementary tree sets.[7] $\gamma \Rightarrow \gamma'$ is a *multicomponent derivation step* in G iff there is an instance $\{\gamma_1, \ldots, \gamma_n\}$ of an elementary tree set in $\mathcal{A}$ and there are pairwise different node addresses $p_1, \ldots, p_n$ such that $\gamma' = \gamma[p_1, \gamma_1] \ldots [p_n, \gamma_n]$ where $\gamma[p_1, \gamma_1] \ldots [p_n, \gamma_n]$ is the result of adding the γ_i $(1 \leq i \leq n)$ at node positions p_i in γ. As in TAG, a derivation starts from an initial tree and in the final derived tree, all leaves must be labelled by terminals.

In each MCTAG derivation step, the trees from a new elementary tree set are added to the already derived tree. Since they are added to pairwise different nodes, one can as well add them one after the other, i.e., each multicomponent derivation in an MCTAG $G = \langle I, A, N, T, \mathcal{A}\rangle$ corresponds to a derivation in the TAG $G_{TAG} := \langle I, A, N, T\rangle$. Let us define the *TAG derivation tree* of such a multicomponent derivation as the corresponding derivation tree in G_{TAG}.[8]

[4] To be precise, this must be an occurrence of an elementary tree. Henceforth, whenever we use an elementary tree in a derivation we actually mean an occurrence of this elementary tree.

[5] As usual, we use the following notations for substitution and adjunction. For trees γ and γ' and for node positions p, $\gamma[p, \gamma']$ is defi ned as follows: If γ' is (derived from) an initial tree with root label $X \in N$ and the node at position p in γ is a substitution node with label X, then $\gamma[p, \gamma']$ is the tree one obtains by substitution of γ' into γ at node position p. If γ' is (derived from) an auxiliary tree with root label $X \in N$and if the node at position p in γ is an internal node with label X, then $\gamma[p, \gamma']$ is the tree one obtains by adjunction of γ' to γ at node position p. Otherwise $\gamma[p, \gamma']$ is undefi ned.

[6] Linear precedence is not needed in a derivation tree since it does not influence the result of the derivation.

[7] $P(X)$ is the set of subsets of some set X.

[8] This TAG derivation tree is not the MCTAG derivation tree defi ned in Weir (1988). The nodes of Weir's MCTAG derivation trees are labelled by sequences of elementary trees (i.e., by elementary tree sets) and each edge stands for simultaneous adjunctions/substitutions of all elements of such a set.

3 A descriptive characterization of MCTAG

The TAG derivation trees for MCTAG derivations have certain properties resulting from the requirement that the elements of elementary tree sets must be added simultaneously: Firstly, if an elementary tree set is used, then all trees from this set must occur in the derivation tree. Secondly, one tree from an elementary tree set cannot be substituted or adjoined into another tree from the same set. Thirdly, different tree sets cannot be interleaved. More concretely there cannot be n tree sets such a tree from the first is added to a tree from the second, a tree from the second to a tree from the third etc. (which amounts to adding first the nth tree set, then the $(n-1)$th etc.), while at the same time a tree from the nth set is added to a tree from the first set. For non-local MCTAG, these are all constraints the TAG derivation tree needs to satisfy.

Lemma 1 *Let $G = \langle I, A, N, T, \mathcal{A}\rangle$ be an MCTAG, $G_{TAG} := \langle I, A, N, T\rangle$. Let $D = \langle \mathcal{N}, \mathcal{E}\rangle$ be a derivation tree in G_{TAG} with the corresponding derived tree t being in $L(G_{TAG})$.*

D is a possible TAG derivation tree in G with $t \in L(G)$ iff D is such that

- **(MC1)** *The root of D is an instance of an initial tree $\alpha \in I$ and all other nodes are instances of trees from tree sets in $\mathcal{A}$ such that for all instances Γ of elementary tree sets from $\mathcal{A}$ and for all $\gamma_1, \gamma_2 \in \Gamma$: if $\gamma_1 \in \mathcal{N}$, then $\gamma_2 \in \mathcal{N}$.*
- **(MC2)** *For all instances Γ of elementary tree sets from $\mathcal{A}$ and for all $\gamma_1, \gamma_2 \in \Gamma$, $\gamma_1 \neq \gamma_2$: $\langle \gamma_1, \gamma_2\rangle \notin \mathcal{D}_D$.*
- **(MC3)** *For all pairwise different instances $\Gamma_1, \Gamma_2, \ldots, \Gamma_n$, $n \geq 2$ of elementary tree sets from $\mathcal{A}$: there are no $\gamma_1^{(i)}, \gamma_2^{(i)} \in \Gamma_i$, $1 \leq i \leq n$ such that $\langle \gamma_1^{(1)}, \gamma_2^{(n)}\rangle \in \mathcal{D}_D$ and $\langle \gamma_1^{(i)}, \gamma_2^{(i-1)}\rangle \in \mathcal{D}_D$ for $2 \leq i \leq n$.*

The proof is given in Kallmeyer (2005). The lemma gives us a way to characterize non-local MCTAG via the properties of the TAG derivation trees the grammar licenses and thereby to get rid of the original simultaneity requirement: The corresponding properties are now captured in the three constraints (MC1)–(MC3). Since these constraints need to hold only for the TAG derivation trees that correspond to derived trees in the tree language, sub-derivation trees need not satisfy them. In other words, γ_1 and γ_2 from the same tree set can be added at different moments of the derivation as long as the final TAG derivation tree satisfies (MC1)–(MC3).

We can now define tree-local and set-local TAG derivation trees by imposing further conditions: Let $G = \langle I, A, N, T, \mathcal{A}\rangle$ be an MCTAG. Let $D = \langle \mathcal{N}, \mathcal{E}\rangle$ be a TAG derivation tree for some $t \in L(\langle I, A, N, T\rangle)$. D is a *multicomponent* derivation tree iff it satisfies (MC1)–(MC3). D is *tree-local* iff for all instances $\{\gamma_1, \ldots, \gamma_n\}$ of elementary tree sets with $\gamma_1, \ldots, \gamma_n \in \mathcal{N}$: there is one γ such that $\langle \gamma, \gamma_1\rangle, \ldots \langle \gamma, \gamma_n\rangle \in \mathcal{P}_D$. D is *set-local* iff for all instances $\{\gamma_1, \ldots, \gamma_n\}$ of elementary tree sets with $\gamma_1, \ldots, \gamma_n \in \mathcal{N}$: there is an instance Γ of an elementary tree set such that for all $1 \leq i \leq n$ there is a $t_i \in \Gamma$ with $\langle t_i, \gamma_i\rangle \in \mathcal{P}_D$.

The following lemma is immediate.

Lemma 2 *Let G be an MCTAG.*

- *G is a* tree-local MCTAG *iff the set of trees generated by G, $L_T(G)$, is defined as the set of those trees that can be derived with a tree-local multicomponent TAG derivation tree in G.*
- *G is a* set-local MCTAG *iff the set of trees generated by G, $L_T(G)$, is defined as the set of those trees that can be derived with a set-local multicomponent TAG derivation tree in G.*

4 Conclusion

MCTAG is an extension of TAG that has been shown to be useful for many natural language applications. Therefore a profound understanding of the mathematical properties of the formalism is indispensable. In a TAG, the central structure of a derivation, the derivation tree abstracts away from the order of derivation steps as long as the result of the derivation is the same: in the derivation tree, the adjunction/substitution operations corresponding to different daughters of the same node can be performed in any order without influencing the derived tree one obtains. Consequently, the derivation trees are unordered with respect to linear precedence.

This way of abstracting away from the concrete order of derivation steps is not possible with the classical MCTAG definition. The definition is problematic since it refers to the process of the derivation itself: a simultaneity constraint must be respected concerning the way the members of the elementary tree sets are added. Looking only at the result a derivation (i.e., the derived tree and the derivation tree), this simultaneity is no longer visible and therefore cannot be checked. I.e., this way of characterizing MCTAG does not allow to abstract away from the concrete order of derivation. Therefore, in this paper, we propose an alternative definition of MCTAG that characterizes the trees in the tree language of an MCTAG via the properties of the TAG derivation trees the MCTAG licences. In this way, in MCTAG like in TAG, the TAG derivation tree can be considered being the central structure of the formalism and the desired abstraction can be obtained.

Apart from the fact that this descriptive characterization of MCTAG helps to understand the mathematical properties of the grammar formalism, it probably also has an impact on parsing. Parsing can be done independently from concrete derivations since the simultaneity constraint need not be checked. Only the outcoming derivation trees need to be checked for well-formedness in the sense of (MC1)–(MC3). However, we do not pursue this further here and we leave the subject for future research.

Références

Joshi, A. K.: 1987. An introduction to Tree Adjoining Grammars. *in* A. Manaster-Ramer (ed.), *Mathematics of Language*. John Benjamins. Amsterdam. pp. 87–114.

Joshi, A. K., Levy, L. S. and Takahashi, M. 1975. Tree Adjunct Grammars. *Journal of Computer and System Science* **10**, 136–163.

Kallmeyer, L.: 1999. *Tree Description Grammars and Underspecifi ed Representations*. PhD thesis. Universität Tübingen. Technical Report IRCS-99-08 at the Institute for Research in Cognitive Science, Philadelphia.

Kallmeyer, L. 2005. Tree-local multicomponent tree adjoining grammars with shared nodes. *Computational Linguistics*. To appear.

Kroch, A. S. and Joshi, A. K.: 1987. Analyzing extraposition in a tree adjoining grammar. *in* G. J. Huck and A. E. Ojeda (eds), *Syntax and Semantics: Discontinuous Constituency*. Academic Press, Inc.. pp. 107–149.

Vijay-Shanker, K.: 1987. *A Study of Tree Adjoining Grammars*. PhD thesis. University of Pennsylvania.

Weir, D. J.: 1988. *Characterizing mildly context-sensitive grammar formalisms*. PhD thesis. University of Pennsylvania.

Approches en corpus pour la traduction: le cas MÉTÉO

Philippe Langlais, Thomas Leplus
Simona Gandrabur et Guy Lapalme
RALI
Université de Montréal
http://rali.iro.umontreal.ca/

Mots-clefs : Mémoire de traduction, traduction probabiliste, alignements multiples, réordonnancement à postériori

Keywords: Memory-based translation, statistical translation, multiple alignment, rescoring

Résumé La traduction automatique (TA) attire depuis plusieurs années l'intérêt d'un nombre grandissant de chercheurs. De nombreuses approches sont proposées et plusieurs campagnes d'évaluation rythment les avancées faites. La tâche de traduction à laquelle les participants de ces campagnes se prêtent consiste presque invariablement à traduire des articles journalistiques d'une langue étrangère vers l'anglais; tâche qui peut sembler artificielle. Dans cette étude, nous nous intéressons à savoir ce que différentes approches basées sur les corpus peuvent faire sur une tâche réelle. Nous avons reconstruit à cet effet l'un des plus grands succès de la TA: le système MÉTÉO. Nous montrons qu'une combinaison de mémoire de traduction et d'approches statistiques permet d'obtenir des résultats comparables à celles du système MÉTÉO, tout en offrant un cycle de développement plus court et de plus grandes possibilités d'ajustements.

Abstract Machine Translation (MT) is the focus of extensive scientific investigations driven by regular evaluation campaigns, but which are mostly oriented towards a somewhat artificial task: translating news articles into English. In this paper, we investigate how well current MT approaches deal with a real-world task. We have *rationally reconstructed* one of the only MT systems in daily production use: the METEO system. We show how a combination of a sentence-based memory approach, a phrase-based statistical engine and a neural-network rescorer can give results comparable to those of the current system while offering a faster development cycle and better customization possibilities.

1 Introduction

Depuis la reprise des campagnes d'évaluation NIST[1] la traduction automatique (TA) revêt un caractère de plus en plus compétitif. La tâche partagée à laquelle se prêtent les "compétiteurs" de ces campagnes d'évaluation consiste à traduire vers l'anglais des textes journalistiques. S'il est clair que cette tâche répond en partie à des préoccupations concrètes du pays organisateur, il n'est cependant pas immédiat d'imaginer des applications réelles de la technologie évaluée.

Des tâches de traduction plus spécifiques existent cependant. Lors du workshop IWSLT (Akiba *et al.*, 2004) dont l'objectif premier était de proposer un protocole d'évaluation adapté à la traduction de corpus oralisés, la tâche partagée consistait à traduire des phrases du corpus BTEC (Basic Travel Expression Corpus). Ce corpus regroupe des phrases susceptibles d'être utiles à un touriste à l'étranger. Une autre tâche de traduction plus ciblée et qui a fait l'objet de nombreuses études est la tâche Verbmobil (Wahlster, 2000) qui consiste à traduire des dialogues de tâches précises (comme la prise de rendez-vous) pour la paire de langue anglais/allemand.

Dans cette étude, nous nous intéressons à une tâche encore plus précise et dont l'applicabilité ne fait cette fois-ci aucun doute puisqu'elle est reconnue comme l'un des plus grand succès de la traduction automatique: la traduction de l'anglais vers le français de bulletins météorologiques émis par *Environnement Canada* (EC)[2]. Nous baptisons cette tâche MÉTÉO.

Récemment, Leplus et al. (2004) montraient qu'à l'aide d'une mémoire de traduction phrastique peuplée de bulletins météorologiques déjà traduits, il était possible d'obtenir des traductions de bonne qualité. Ils expliquaient leur succès par un fort taux de répétitivité des phrases que le système MÉTÉO traduit. Dans ce travail, nous étudions la pertinence de plusieurs approches basées sur les corpus à traduire les bulletins météorologiques.

2 Protocole

Nous avons utilisé dans ce travail le bitexte décrit dans (Leplus *et al.*, 2004). Nous avons repris le même découpage en trois partie de ce bitexte: TRAIN pour l'entraînement des systèmes, BLANC pour leur ajustement et TEST pour tester les différentes approches. Ce découpage avait été choisi de manière à ce que les textes soient d'une période disjointe et que la tranche de test soit d'une période postérieure à celle de l'entraînement; ceci afin de simuler autant que faire se peut les conditions réelles d'utilisation du système.

Pour évaluer nos différentes approches, nous utilisons des métriques automatiques qui bien que discutables n'en sont pas moins largement utilisées: deux taux d'erreurs — WER au niveau des mots et SER au niveau des phrases — que l'on cherchera à minimiser et deux mesures de couverture n-gramme — NIST et 100×BLEU — que l'on voudra maximiser, toutes les deux calculées par le script `mteval` (version 11a) disponible depuis le site de NIST.

[1] Consulter `http://www.nist.gov/speech/tests/mt/` pour plus d'information.

[2] Le bulletin en cours peut être consulté à l'adresse `http://meteo.ec.gc.ca/forecast/textforecast_f.html`

3 Mémoire de traduction phrastique

Nous avons mesuré que 83% des phrases du corpus BLANC sont présentes *verbatim* dans le corpus TRAIN. Cette couverture atteint 87% si nous introduisons quelques classes de mots comme les jours, les mois ou encore les numéros de téléphones. Nous avons donc commencé par reproduire l'approche mémoire de traduction phrastique proposée par Leplus et al. (2004).

Nous avons construit une mémoire en gardant de chaque phrase source de TRAIN, un maximum de 5 traductions. En pratique, 89% des phrases anglaises de TRAIN n'ont qu'une seule traduction, probablement en raison du fait que la plupart des phrases ont été produites automatiquement (nous reviendrons sur ce point dans la section 7).

Pour une nouvelle phrase à traduire, nous recherchons les phrases sources les plus proches (en terme de distance d'édition) dans la mémoire et trions les traductions associées selon un score dont le détail est décrit dans (Langlais *et al.*, 2005). Dans cette expérience, la première phrase cible retournée est la traduction retenue.

mémoire				Leplus et al.			
WER%	SER%	NIST	BLEU	WER%	SER%	NIST	BLEU
8.42	23.43	10.9571	87.68	9.18	23.56	10.8983	86.95

Table 1: Évaluation de l'approche mémoire phrastique sur le corpus TEST et comparaison avec l'approche Leplus et al. (2004).

Les scores sont très bons si on les compare avec ceux observés dans d'autres tâches de traduction. Nous référons le lecteur à l'étude de Zens et Ney (2004) pour des performances état de l'art sur trois tâches de traduction incluant Verbmobil. Nos performances sont également légèrement supérieures à celles mentionnées par (Leplus *et al.*, 2004). Il n'en reste cependant pas moins que le taux d'erreur au niveau des phrases (c'est-à-dire le pourcentage de traductions produites non identiques à la traduction de référence) n'est pas particulièrement bas.

4 Approche probabiliste

Nous avons testé dans un deuxième temps une approche état de l'art en traduction statistique (Koehn *et al.*, 2003). Elle s'appuie sur un modèle de la distribution conditionnelle d'une séquence de mots dans une langue étant donnée une séquence dans l'autre langue. Les détails de l'obtention des modèles probabilistes sous-jacents sont donnés dans (Langlais *et al.*, 2005). Nous avons fait usage du décodeur PHARAOH (Koehn, 2004) disponible gratuitement pour des fins de recherche.

Les performances du système probabiliste sont présentées en table 2. Une comparaison directe avec les résultats mesurés avec l'approche mémoire milite en faveur de la mémoire, surtout si l'on observe le taux d'erreur au niveau des phrases. Cependant, nous remarquons que la performance du traducteur probabiliste lorsque mesurée sur les phrases à traduire qui n'ont pas été vues *verbatim* dans le corpus TRAIN sont de loin supérieures à celles obtenues par l'approche mémoire. Nous reviendrons sur la complémentarité de ces deux approches en section 7.

WER%	SER%	NIST	BLEU
7.46	32.01	10.8725	84.03

Table 2: Évaluation de l'approche statistique sur TEST.

5 Approche consensuelle

Bangalore et al. (2002) ont montré qu'il était possible de combiner des traductions produites par différents moteurs de traduction afin de générer des traductions d'une qualité supérieure à celles produites par un seul des moteurs. L'idée sous-jacente à cette approche (*bootstrapping*) est l'alignement de plusieurs traductions candidates afin d'isoler des îlots de confiance capables de diriger la génération d'une traduction dite consensuelle. Nous retrouvons cette idée dans certains systèmes d'acquisition et de génération de paraphrases.

Nous avons reproduit cette approche et avons pour cela adapté à nos besoins le programme CLUSTALW (Thompson *et al.*, 1994) écrit pour aligner entre-elles plusieurs séquences de protéines. À partir d'un alignement multiple de traduction (dont le lecteur trouvera les détails dans (Langlais *et al.*, 2005)), nous pouvons construire un treillis qui permet de produire en sus des traductions alignées de nouvelles phrases que l'on espère plus robustes. Nous utilisons le package CARMEL (Knight & Al-Onaizan, 1999) pour trouver dans un treillis la traduction consensuelle; c'est-à-dire le chemin de plus faible coût dans le treillis.

Les résultats de cette approche sont présentés en table 3 pour les seules phrases de BLANC non rencontrées *verbatim* dans le corpus ayant servi à créer la mémoire. Nous observons que la traduction par consensus améliore la qualité (telle que mesurée) des traductions produites. Le taux d'erreur au niveau des phrases est en particulier réduit de 9 points (en absolu), ce qui constitue une amélioration notable.

mémoire				mémoire + consensus			
WER%	SER%	NIST	BLEU	WER%	SER%	NIST	BLEU
18.69	94.82	9.7853	66.56	18.97	85.53	9.9314	68.86

Table 3: Performance de l'approche consensuelle sur la sortie de la mémoire de traduction pour les 13 010 phrases de BLANC non rencontrées dans le corpus TRAIN.

6 Ré-ordonnancement par apprentissage neuronal

Dans notre cadre, le *rescoring* consiste à ré-ordonner une liste d'alternatives produites par un système (dit natif) avec l'espoir que des informations supplémentaires, ou différentes façons de les utiliser, permettent de produire un ordonnancement plus pertinent. Le *rescoring* a fait l'objet d'études récentes en traduction probabiliste (Blatz *et al.*, 2004).

Dans notre contexte, cela consiste à reclasser la liste des meilleures traductions générées par PHARAOH (Koehn, 2004) pour une phrase donnée. Chaque alternative de traduction t_j est représentée par un vecteur de traits v_j et est étiquetée comme correcte si elle est identique à la traduction de référence et incorrecte sinon. Nous avons utilisé le package TORCH (Collobert

et al., 2002) pour entraîner un réseau perceptron multi-couche à estimer $p(\oplus|v_j)$, la probabilité conditionnelle de la correctitude d'une alternative t_j.

Nous avons testé différentes configurations de la couche cachée du réseau et avons considéré de nombreux traits pour représenter nos alternatives, chacun encodant des caractéristiques particulières. Les plus utiles étaient a) le ratio des longueurs de la phrase source et de la traduction candidate, b) la probabilité *a posteriori* de l'alternative et c) les scores $p(t_j|s)$ calculés par les modèles IBM 1 et 2 (Brown *et al.*, 1993). De plus amples informations sur cette approche sont disponibles dans (Langlais *et al.*, 2005).

Nous présentons en table 4 les performances mesurées par l'étape de rescoring.

smt				smt + rescoring			
WER%	SER%	NIST	BLEU	WER%	SER%	NIST	BLEU
7.46	32.01	10.8725	84.03	5.73	25.03	10.9828	87.40

Table 4: Comparaison des performances du moteur probabiliste (smt) seul et des traductions produites par reclassement (smt + *rescoring*) sur TEST.

7 Discussion

La diversité des approches que nous avons implémentées nous donne la souplesse de pouvoir les combiner. Pour illustrer ce point, nous avons évalué une combinaison très simple où la mémoire seule est consultée lorsque la phrase à traduire est déjà dans la mémoire, et où le moteur de traduction probabiliste *rescoré* est consulté sinon. Les performances ainsi mesurées (voir la table 5) sont meilleures que celles de chaque approche prise isolément.

WER%	SER%	NIST	BLEU
4.85	20.80	11.3021	89.59

Table 5: Performance sur TEST de la combinaison de la mémoire et du moteur de traduction probabiliste reclassé.

Il est cependant approprié de s'interroger quant à la performance véritable d'un tel système. Il est en particulier intéressant de contraster ces résultats avec ceux mesurés par le *bureau de la traduction du Canada* (BTC) qui est en charge de produire les traductions des bulletins météorologiques produits par *Environnement Canada* (EC). Le BTC utilise en effet le système MÉTÉO pour traduire automatiquement les bulletins anglais, mais a la responsabilité de réviser tout ou partie des traductions ainsi produites.

(Macklovitch, 1985) décrit une évaluation du système MÉTÉO-II conduite par le BTC. L'auteur a sélectionné 1257 phrases françaises publiées sur une période de 24 heures par EC et a compté le nombre de fois où le système produisait exactement la même traduction que celle qui a été publiée. Les erreurs dues à des fautes flagrantes non imputables au système étaient cependant écartées (typos, erreur de transmission, etc.). Il rapporte que seulement 11% des phrases testées étaient différentes de celles publiées.

Ce protocole d'évaluation correspond grossièrement au nôtre lorsque nous mesurons un taux d'erreur au niveau des phrases. Les approches que nous avons implémentées ne montrent pas un tel niveau de performance. Cependant, une comparaison directe des deux protocoles n'est pas adéquate. Premièrement, nous évaluons nos approches sur un corpus bien plus grand (36 228 phrases). Deuxièmement, nous avons mesuré un bruit d'environ 7% dans notre référence. Troisièmement, une évaluation informelle d'un échantillon de 1000 traductions (choisies aléatoirement) différentes de celles de notre référence, nous a révélé que 77% d'entre-elles étaient des traductions correctes.

Références

AKIBA Y., FEDERICO M., KANDO N., NAKAIWA H., PAUL M. & TSUJII J. (2004). Overview of the IWSLT04 evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, p. 1–12, Kyoto.

BANGALORE S., MURDOCK V. & RICCARDI G. (2002). Bootstrapping bilingual data using consensus translation for a multilingual instant messaging system. In *Proceedings of COLING*, p. 50–56, Taipei.

BLATZ, J., FITZGERALD, E., FOSTER, G., GANDRABUR, S., GOUTTE, C., KULESZA, A., SANCHIS, A., UEFFING & N. (2004). Confidence estimation for machine translation. In *Proceedings of COLING*, p. 315–321, Geneva.

BROWN P., PIETRA S. D., PIETRA V. D. & MERCER R. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, **19**(2), 263–311.

COLLOBERT R., BENGIO S. & MARIÉTHOZ. J. (2002). *Torch: a modular machine learning software library*. Rapport interne IDIAP-RR 02-46, IDIAP.

KNIGHT K. & AL-ONAIZAN Y. (1999). *A Primer on Finite-State Software for Natural Language Processing*. http://www.isi.edu/licensed-sw/carmel/carmel-tutorial2.pdf.

KOEHN P. (2004). Pharaoh: a beam search decoder for phrase-based smt. In *Proceedings of AMTA*, p. 115–124, Washington.

KOEHN P., OCH F. & MARCU D. (2003). Statistical phrase-based translation. In *Proceedings of HLT*, p. 127–133, Edmonton.

LANGLAIS P., LEPLUS T., GANDRABUR S. & LAPALME G. (2005). From the real world to real words: The meteo case. In *in 10th Annual Conference of the European Association for Machine Translation*, Budapest, Hungary.

LEPLUS T., LANGLAIS P. & LAPALME G. (2004). Weather report translation using a translation memory. In *Proceedings of AMTA*, p. 154–163, Washington.

MACKLOVITCH E. (1985). *A Linguistic Performance Evaluation of METEO 2*. Rapport interne, Canadian Translation Bureau.

THOMPSON J., HIGGINS D. & GIBSON T. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**(22), 4673–4680.

WAHLSTER, Ed. (2000). *Verbmobil: Foundations of speech-to-speech translations*. Berlin, Germany: Springer Verlag.

ZENS R. & NEY H. (2004). Improvements in phrase-based statistical machine translation. In *Proceedings of HLT/NAACL*, p. 257–264, Boston.

Simplification interactive pour la production de textes adaptés aux personnes souffrant de troubles de la compréhension

Aurélien Max
LIR-LIMSI-CNRS
Université Paris 11
Orsay, France
aurelien.max@limsi.fr

Mots-clefs : troubles du langage, simplification syntaxique, règles de réécriture, validation interactive, traitements de texte

Keywords: language disorders, syntactic simplification, rewriting rules, interactive validation, word processors

Résumé Cet article traite du problème de la compréhensibilité des textes et en particulier du besoin de simplifier la complexité syntaxique des phrases pour des lecteurs souffrant de troubles de la compréhension. Nous présentons une approche à base de règles de simplification développées manuellement et son intégration dans un traitement de texte. Cette intégration permet la validation interactive de simplifications candidates produites par le système, et lie la tâche de création de texte simplifié à celle de rédaction.

Abstract This paper addresses the issue of text readability and in particular the need for simplifying the syntactic complexity of sentences for language-impaired readers. The proposed approach uses handcrafted simplification rules and has been integrated into a word processor. This allows interactive validation of candidate simplified sentences produced by the system, and integrates the task of creating simplified texts into that of authoring.

1 Simplification de texte

La *compréhensibilité* est une propriété cruciale d'un texte, et l'usage d'une langue sans contraintes particulières peut mener à des textes difficiles à comprendre. Or la vocation première d'un texte informatif est de véhiculer une information auprès de ses lecteurs (soit d'atteindre certains *buts communicatifs*), ce qui impose des contraintes sur la rédaction de ce texte. Des troubles du langage peuvent néanmoins interdire la compréhension de textes qui seraient autrement jugés raisonnablement compréhensibles. L'*aphasie de Broca*[1] a pour effets de rendre compliquée l'expression ainsi que l'interprétation d'énoncés contenant une certaine *complexité*

[1]L'aphasie est un trouble du langage pouvant survenir à la suite d'une attaque cérébrale touchant les lobes frontaux et temporaux de l'hémisphère gauche du cerveau. Plusieurs catégories d'aphasies existent, et elles peuvent affecter la compréhension aussi bien que la production du langage.

linguistique, alors que les connaissances sémantiques du sujet sont intactes. Il semble donc important de pouvoir produire des textes adaptés à ces lecteurs, ou au moins de dériver des textes simplifiés à partir de textes existants.

La simplification de texte a récemment été l'objet de plusieurs travaux, avec comme motivation la simplification de texte en vue d'une analyse syntaxique (Chandrasekar et al, 1996) ou la simplification de texte destinée aux lecteurs souffrant de troubles du langage (Carroll et al, 1998; Devlin, 1999; Liben-Nowell, 2000; Canning, 2002). L'approche de Chandrasekar et al. se base sur des règles de simplification syntaxique qui sont apprises depuis un corpus simplifié manuellement. Les autres approches utilisent des règles écrites manuellement qui sont appliquées sur la sortie d'un analyseur syntaxique. Plus récemment, (Siddharthan, 2003) s'est intéressé à la conservation de la cohérence des textes résultant de simplifications syntaxiques.

L'approche que nous présentons ici se distingue des précédentes par le fait qu'elle vise à intégrer un module de simplification de texte à base de règles dans un outil de rédaction traditionnel (un traitement de texte). Il nous semble en effet important de pouvoir considérer la production d'un texte simplifié comme une activité liée à celle de la rédaction du texte dont il est issu, ce qui permet d'impliquer l'auteur dans les choix pour lesquels un système automatique ne pourrait prendre de décision satisfaisante dans le cas général. Dans cet article, nous nous concentrerons sur la simplification syntaxique des phrases[2]. Dans l'approche que nous avons choisie, les simplifications sont le résultat de l'application non-déterministe de règles de réécriture développées manuellement par un linguiste. Cette approche est en partie justifiée par la difficulté à obtenir des corpus alignés de textes et de leur simplification qui pourraient servir à un apprentissage de ces règles[3] ou à des techniques basées sur l'exemple.

2 Système de simplification syntaxique par réécriture

Nous avons choisi une approche à base de règles pour la simplification syntaxique des phrases d'un texte en anglais s'inspirant de travaux sur la transduction d'arbres. L'utilisation de règles permet d'exprimer des conditions sur leur applicabilité qui offrent une meilleure maîtrise du contexte de simplification qui peut être plus ou moins spécifié (des mots jusqu'aux catégories syntaxiques profondes les dominant par exemple). Le développement des règles peut être incrémental, ce qui autorise une évaluation progressive du système sur un corpus de test constant, afin de contrôler que l'ajout de règles ne dégrade pas les performances. Par ailleurs, la non-couverture d'une structure syntaxique particulière ne réalise pas de simplification non souhaitée, et laisse donc localement la phrase inchangée, ce qui garantit au pire la conservation de sa complexité syntaxique. Une telle approche nous a semblé d'autant plus acceptable si elle est intégrée à un traitement de texte qui autorise facilement les révisions ou modifications.

Le niveau de description des phrases utilisé pour décrire les patrons de réécriture est choisi par le linguiste en charge de l'écriture des règles. Ce niveau doit être dérivable automatiquement à partir du texte, ce qui pose la question du compromis entre la robustesse de l'analyse et la finesse de description. Au minimum, la séquence de parties du discours des mots du texte peut

[2] Un système complet de simplification de texte doit également au minimum pouvoir permettre de résoudre les anaphores ainsi que de simplifier la complexité lexicale.

[3] Comme le note (Siddharthan, 2003), la simplification manuelle d'un corpus de texte permettrait vraisemblablement de proposer un jeu de règles de simplification qui devrait avoir une performance au moins comparable à celle de règles apprises sur ce corpus simplifié.

être utilisée comme une structure plate, mais elle n'offre que peu de possibilités de simplification. Des structures syntaxiques plus ou moins profondes seront donc utilisées en fonction des analyseurs disponibles[4].

Une contrainte importante de notre approche de réécriture par règles est que les informations nécessaires à la génération des phrases simplifiées doivent pouvoir être dérivées à partir des phrases en entrée (bien que l'auteur ait ensuite l'opportunité de modifier les phrases simplifiées). Si passer par une représentation sémantique intermédiaire des phrases pour faire de la simplification par regénération serait souhaitable, cela pose d'importants problèmes d'analyse et de représentation et implique la disponibilité d'un moteur de génération de texte paramétrable. Les règles syntaxiques permettent de réutiliser le contenu présent en entrée dans la sortie du simplificateur, et de nouveaux éléments peuvent soit être directement spécifiés dans la sortie, soit exprimés par le biais de conditions sur les règles.

Il n'est cependant pas entièrement possible de s'affranchir de certaines fonctionnalités de génération morphologique, puisque certaines modifications de la syntaxe impliqueront la production de formes de surface adaptées au nouveau contexte[5]. Divers modules d'analyse et de génération linguistiques peuvent ainsi être utilisés dans les conditions des règles.

La simplification d'un texte est opérée phrase par phrase. Les patrons spécifiant les structures syntaxiques à réécrire sont recherchés dans la structure obtenue pour la phrase en entrée, et ce à quelque profondeur que ce soit. Sous réserve que les éventuelles conditions d'application soient remplies, le patron spécifiant la sortie remplace le patron d'entrée, ce qui peut générer autant de simplifications qu'il y a d'occurrences du patron d'entrée dans la phrase. Les patrons d'entrée des règles sont récursivement recherchés dans les sorties du système pour produire l'ensemble des simplifications possibles tenant compte des différents ordres d'application des règles.

Le format des patrons de sortie doit donc être compatible avec le format des patrons d'entrée, mais les étiquettes de sortie peuvent être volontairement modifiées afin, soit d'interdire l'application ultérieure d'autres règles sur un constituant particulier, soit au contraire de déclencher l'application d'une règle réalisant un traitement particulier[6].

Enfin, la nature des réécritures effectuées requiert la possibilité d'insérer dans la sortie du système de nouvelles phrases, qui pourront précéder ou suivre la phrase dans laquelle le patron d'entrée a été trouvé.

Format des règles de réécriture Nous avons donné une importance particulière au fait que les règles puissent être écrites par des linguistes familiers notamment avec des notations d'arbres de dérivation syntaxiques. Les règles sont fortement basées sur la notion d'unification de varia-

[4]Un analyseur probabiliste robuste pour l'anglais tel que RASP (Bricoe et Carroll, 2002) offre une solution robuste pouvant retourner une forêt de solutions annotées par leur probabilité. Cela pose notamment le problème de la sélection de l'analyse retenue pour opérer la simplification. Bien qu'il soit possible de considérer par défaut l'analyse la plus probable retournée par le système, une certaine désambiguïsation parmi les plus probables pourrait être envisagée dans notre contexte interactif.

[5]C'est par exemple le cas du verbe de la proposition principale lors du passage d'une phrase de la voix passive à la voix active, puisque celui-ci doit s'accorder en personne et en nombre avec l'agent du verbe qui était précédemment situé dans un syntagme prépositionnel, ex: *The cat is chased by the dog.* → *The dog* ***chases*** *the cat.*

[6]Par exemple, il est ainsi possible de marquer un groupe nominal indéfini comme devant être transformé en groupe nominal défini par une règle appropriée lorsqu'il est repris dans une nouvelle phrase, ex.:*A terrifying dog chases the cat.* → *A dog is chasing the cat.* ***The dog*** *is terrifying.*

bles, qui peuvent correspondre à des littéraux, à des arbres, ou à des forêts d'arbres[7].

La règle ci-dessous donne un exemple pour le passage à la voix active de phrases à la voix passive contenant un agent exprimé[8]. Les catégories et structures syntaxiques de l'exemple sont celles utilisées dans le Penn TreeBank qui nous a servi comme corpus de test initial.

```
define Activise passive sentences with overt agent

if      [be, InflBe] is analyzeVerb(TagBe, [Be]);
        ?OptAdvs contains only advp rb; ?OptPart contains only prt;
rewrite [s      [       ?Opt1
                        [np-Index NPTheme]
                        ?Opt2
                        [vp     [       [TagBe [Be]]
                                        [vp  [  ?OptAdvs
                                                [vbn Verb]
                                                ?OptPart
                                                [np [[none [Trace-Index]]]]
                                        ?Opt3
                                        [pp     [       [prep [by]]
                                                        [np/lgs NPAgent]]]
                                                        ?Opt4]]]]
                        ?Opt5]]
as      [s      [       ?Opt1
                        [np NPAgent]
                        ?Opt2
                        [vp     [       [SurfaceTag [SurfaceVerb]]
                                        ?OptPart
                                        [np NPTheme]
                                        ?Opt3
                                        ?Opt4]]
                        ?Opt5]]
where   [Number, Person] is number(NPAgent);
        [BaseForm, Infl] is analyzeVerb(vbn, Verb);
        [SurfaceVerb, SurfaceTag] is generateVerb(BaseForm, InflBe, Number, Person);
```

La première condition de la clause **if** qui fait l'appel de la fonction linguistique `analyzeVerb` indique que le verbe de la clause principale doit avoir pour lemme *be*, et que la flexion du verbe (temps et personne) doit se retrouver dans la variable `InflBe`. Cette dernière information sera utilisée dans la clause **where** pour produire la version de surface du verbe principal dans la phrase à la voix active (fonction `generateVerb`). Les deux autres conditions de la clause **where** tentent de reconnaître la personne et le nombre du groupe nominal agent, ainsi que le lemme du verbe au participe passé (fonctions `number` et `analyzeVerb`).

Les constituants optionnels représentant des forêts d'arbres (ex: `?OptAdvs`) peuvent être réutilisés ou non dans les patrons de sortie, mais il n'est pas possible de faire référence à leur structure interne dans la règle dans laquelle ils sont reconnus, ce qui pourra être fait par l'application de règles ultérieures si ces constituants sont réinjectés dans un patron de sortie. Cette fonctionnalité est particulièrement utile pour des constituants qui ne sont pas pertinents pour la règle de simplification en cours d'application mais qui sont dominés par le constituant racine d'un pa-

[7] L'implémentation du moteur de simplification a été réalisée en langage Prolog.

[8] Une estimation (Canning, 2002) indique qu'environ 80% des phrases en anglais au passif n'ont pas d'agent exprimé, et que dans ce cas il est particulièrement difficile, voire impossible, de le retrouver (ex: *She was taken to the hospital*). Notre système n'a pas la prétention de simplifier de telles phrases, bien qu'il serait possible dans certains cas d'insérer un agent indéfini tel que *something* ou *someone* (ce qui requiert néanmoins la détermination du caractère "animé" de cet agent).

tron d'entrée (par exemple, `?Opt4` et `?Opt5` regrouperont l'ensemble des constituants suivant le syntagme prépositionnel contenant l'agent et appartenant au groupe verbal englobant).

Sélection de la meilleure simplification L'application des règles de simplification syntaxique produit une liste de sorties qui sont ordonnées par ordre de production par le moteur de simplification, qui dépend de l'ordre d'analyse des règles. Puisque nous avons choisi d'intégrer la simplification syntaxique dans un traitement de texte, il est possible de demander à l'utilisateur de choisir la simplification qui lui semble la plus appropriée. Cependant, il est souhaitable de pouvoir au préalable ordonner les simplifications candidates par un score décroissant qui indique leur *qualité*. Idéalement, cette qualité impliquerait une prise en compte de la compréhensibilité des phrases, ainsi que de la conservation du sens par rapport à la phrase de départ. Or ce dernier élément relève du jugement expert de l'auteur du texte (ou de l'utilisateur de notre système)[9].

De nombreuses mesures de compréhensibilité (*readability*) des textes ont été proposées (voir par ex. (Siddharthan, 2003)), comme par exemple la formule de Flesch qui combine le nombre de syllabes par mot et le nombre de mots par phrase. Une hypothèse fondamentale de notre approche est qu'une règle de simplification "casse" une structure syntaxique difficile pour la remplacer par une structure syntaxique plus facile à comprendre[10]. Ainsi, la simplification syntaxique effectuée doit corréler avec le nombre de règles appliquées. L'ordonnancement des simplifications utilisé prend donc en compte le nombre d'applications de règles sans doublons dans la sortie[11], puis la taille moyenne des phrases. À la demande de l'utilisateur, une phrase est analysée puis simplifiée, puis les 5 simplifications obtenant les meilleurs scores sont proposées.

Évaluation initiale du système Nous avons développé un jeu de règles destiné à simplifier les textes pour les personnes souffrant d'aphasie. Pour cela, nous avons notamment suivi les résultats expérimentaux obtenus par Caplan (Caplan, 1987) et avons extrait des structures syntaxiques présentant une complexité particulière pour ces patients. Le tableau 1 illustre certains types de phénomènes traités avec un exemple de simplification proposée.

Nous n'avons pas été en mesure de pouvoir nous-mêmes évaluer l'impact sur la compréhension de phrases par des patients aphasiques. L'évaluation de la conservation du sens est un problème très difficile auquel est confronté toute approche de paraphrase. L'applicabilité de notre approche vient du fait que l'auteur du texte d'origine a la possibilité de valider interactivement la simplification proposée par le système qui lui semble la plus adéquate. Il nous reste à évaluer empiriquement l'acceptabilité et l'utilisabilité de notre système au sein d'un traitement de texte.

[9] Sans décision humaine, la simplification obtenant le meilleur score peut être proposée par défaut.

[10] Ceci ne concerne néanmoins pas toutes les règles, puisque certaines ont pour but de réaliser des changements nécessaires pour assurer la cohérence du texte, comme la règle mentionnée plus tôt transformant un groupe nominal indéfini en groupe nominal défini.

[11] Afin de contourner ce problème, nous souhaitons modifier le moteur de simplification pour qu'il applique d'abord les règles qui simplifient les structures les plus "profondes" dans l'arbre syntaxique de la phrase en entrée. Cette approche semble également intuitivement plus proche de la méthodologie de simplification suivie par un humain.

Type de phrase	Exemple de simplification
Passive	The elephant$_i$ was hit t_i by the monkey → The monkey hit the elephant.
Cleft Object	It was the elephant$_i$ [that the monkey hit t_i] → The monkey hit the elephant.
Dative	The elephant gave the rabbit the monkey → The elephant gave the monkey. The rabbit received the monkey.
Conjoined	The elephant [[hit the monkey] and [hugged the rabbit]] → The elephant hit the monkey. The elephant hugged the rabbit.
Subject-Object relative	The elephant$_i$ [that the monkey hit t_i] hugged the rabbit → The monkey hugged the elephant. The elephant hugged the rabbit.
Object-Subject relative	The elephant hit [the monkey that hugged the rabbit] → The monkey hugged the rabbit. The elephant hit the monkey.

Figure 1: Exemples d'application de règles de simplification

3 Perspectives

Un point important pour l'acceptabilité du système concerne le fait que le système doit éviter de demander à l'utilisateur plusieurs validations pour des phrases similaires, et propager des décisions lorsque cela est possible. La validation interactive permet de constituer progressivement un corpus aligné de phrases et de leur simplification, ainsi que d'attribuer des poids aux règles de simplification étant donné l'historique d'une série d'applications de règle. Ces poids pourraient alors être utilisés par le moteur pour guider les simplifications ultérieures d'après la *mémoire de simplification* ainsi constituée.

Remerciements L'auteur remercie David Liben-Nowell pour les nombreuses discussions sur le thème de la simplification syntaxique et pour le travail commun sur le système initial.

Références

Briscoe, Edward et John Carroll (2002), Robust accurate statistical annotation of general text, *Actes de Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Gran Canaria.

Canning, Yvonne (2002), *Syntactic simplification of text*, Thèse de PhD, Université de Sunderland.

Caplan, D. (1987), *Neurolinguistics and Linguistic Aphasiology*, Cambridge University Press, Cambridge.

Carroll, John, Guido Minnen, Yvonne Canning, Siobhan Devlin et John Tait (1998), Practical Simplification of English Newspaper Text to Assist Aphasic Readers, *Actes de AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, Madison, États-Unis.

Chandrasekar, R., C. Doran et B. Srinivas (1996), Motivations and methods for text simplification, *Actes de COLING'96*, Copenhague, Danemark.

Devlin, Siobhan (1999), *Simplifying natural language for aphasic readers*, Thèse de PhD, Université de Sunderland.

Liben-Nowell, David (2000), *Syntactic Simplification*, Thèse de MPhil, Université de Cambridge.

Siddharthan, Advaith (2003), *Syntactic Simplification and Text Cohesion*, Thèse de PhD, Université de Cambridge.

Indexation automatique de ressources de santé à l'aide de *paires* de descripteurs MeSH

Aurélie Névéol[1,2] Alexandrina Rogozan[1] Stéfan J. Darmoni[1,2]

[1]Laboratoire PSI – FRE 2645 CNRS Université et INSA de Rouen
{aneveol, arogozan}@insa-rouen.fr
[2]Equipe CISMeF, CHU de Rouen - 1, rue de Germont - 76031 Rouen
stefan.darmoni@univ-rouen.fr

Mots-clés: Indexation Automatique, Terminologie Médicale, Vocabulaire Contrôlé.

Keywords: Automatic Indexing, Medical Terminology, Controlled Vocabulary.

Résumé: Depuis quelques années, médecins et documentalistes doivent faire face à une demande croissante dans le domaine du codage médico-économique et de l'indexation des diverses sources d'information disponibles dans le domaine de la santé. Il est donc nécessaire de développer des outils d'indexation automatique qui réduisent les délais d'indexation et facilitent l'accès aux ressources médicales. Nous proposons deux méthodes d'indexation automatique de ressources de santé à l'aide de paires de descripteurs MeSH. La combinaison de ces deux méthodes permet d'optimiser les résulats en exploitant la complémentarité des approches. Les performances obtenues sont équivalentes à celles des outils de la littérature pour une indexation à l'aide de descripteurs seuls.

Abstract: The increasing number of health documents available in electronic form, and the demand on both practitioners and librarians to encode these documents with controlled vocabulary information calls for automatic tools and methods to help them perform this task efficiently. In this article, we are presenting and comparing two methods for the automatic indexing of heath resources with pairs of MeSH descriptors. A combination of both methods achieves better results by exploiting the complementarity of the approaches. This performance matches the tools described in the litterature for single term indexing.

Introduction

Depuis quelques années, le nombre de documents électroniques augmente de manière exponentielle. Se pose alors la question de l'exploitation efficace de ces ressources. Dans les domaines de la santé et de la bio-médecine, de nombreux travaux ont été entrepris afin de guider les utilisateurs dans leur recherche d'information. Ainsi, la base documentaire MEDLINE® recense 11 millions d'articles scientifiques en langue anglaise indexés à l'aide du thésaurus MeSH® (Medical Subject Headings) développé et maintenu par la NLM

(National Library of Medicine). En Europe, la fondation HON (Health On the Net-http://www.hon.ch/) ou le Catalogue et Index des Sites Médicaux Francophones (CISMeF-http://www.cismef.org) se proposent de guider les internautes vers une information médicale de qualité. Dans ces deux projets européens, la description des ressources[1] s'appuie sur la version française du MeSH. Notre objectif est de formaliser cette démarche et de proposer une méthode d'indexation automatique en adéquation avec les caractéristiques de l'indexation manuelle. Nous présentons deux méthodes répondant à ces critères. Après les avoir évaluées séparément puis combinées, nous discutons de l'intérêt d'utiliser de tels outils dans le cadre de l'aide à l'indexation.

1 Méthodes d'indexation

1.1 Indexation Manuelle

Les principes de l'indexation manuelle à l'aide de descripteurs MeSH (mots clés et qualificatifs) sont clairement exposés dans (Dailland et al., 2003). Les caractéristiques de cette indexation sont: (a) l'utilisation de descripteurs obligatoires (des mots clés particuliers, par exemple <*sujet âgé*>), (b) l'association de qualificatifs pour préciser les mots clés le cas échéant (par exemple, la paire <*diabète/chimiothérapie*> sera utilisée de préférence au mot clé isolé <*diabète*> pour évoquer les traitements médicamenteux du diabète) et (c) l'adaptation du nombre de mots clés (ou de paires) utilisés pour indexer une ressource en fonction de son contenu (par exemple, dans le catalogue CISMeF, les ressources sont indexées avec un nombre de mots clés (ou paires) pouvant aller de zéro à plusieurs dizaines). De manière plus générale, on retiendra également qu'il existe un consensus entre les experts (Anderson et Pérez-Carballo, 2001) selon lequel l'indexation manuelle s'effectue en deux étapes: 1- l'analyse du document, afin d'en retirer le contenu et 2- la traduction de ce contenu dans le langage retenu pour la description (par exemple, le MeSH). En pratique, une troisième étape de relecture et de révision peut également être ajoutée.

1.2 Indexation Automatique

1.2.1 Méthode de Traitement Automatique de la Langue Naturelle (TALN)

Cette méthode, détaillée dans (Névéol, 2004), suit les étapes de l'indexation manuelle (analyse, traduction et révision). Une analyse de surface utilisant un dictionnaire MeSH et une bibliothèque de graphes permet de reconnaître les différentes formes prises par les concepts médicaux (flexions, synonymes, hyperonymes, …) et de les comptabiliser afin d'attribuer un score à chaque concept. Les informations nécessaires à la traduction des concepts sous forme MeSH sont contenues dans le dictionnaire (pour les mots clés) et dans les graphes (pour les paires). Les relations hiérarchiques entre termes MeSH permettent de réduire la liste des candidats en reportant les occurrences des mots clés pères vers leurs fils afin d'indexer au

[1] Afin de rendre compte de la multiplicité des documents électroniques indexés tant du point de vue du format, du type de document, que des usages auxquels ils sont destinés, nous utiliserons le terme de "ressource".

plus précis. L'application de règles d'indexation dans un deuxième temps permet de réviser la liste des candidats avant d'obtenir l'indexation finale avec la fonction de rupture décrite en 1.2.3. Les règles d'indexation sont de deux types :

1. Des **règles d'indexation issues de la NLM** préconisant l'utilisation d'un mot clé MeSH de préférence à une paire mot clé/qualificatif pour représenter un concept. La règle « $MC_1/Q_1 \Rightarrow MC_2$ » indique qu'il convient de remplacer la paire MC_1/Q_1 par le seul mot clé MC_2. Par exemple:<*cœur/transplantation*> => <*transplantation cardiaque*>

2. Des **règles d'indexation CISMeF** préconisant l'introduction d'un mot clé (ou paire). Ainsi, la règle « $MC_1/Q_1^2 + MC_2/Q_2^2$ » indique qu'il convient d'ajouter la paire MC_2/Q_2 à l'indexation d'une ressource déjà indexée avec la paire MC_1/Q_1. Par exemple : <*appendicectomie*> + <*appendicite/chirurgie*>

1.2.2 Méthode des k plus proches voisins (k-PPV)

La méthode des *k*-PPV est une référence dans le domaine de la classification. Son principe est simple. Soit C une collection de ressources étiquetées notées r_i. Soit $r \notin C$ une ressource à étiqueter. On calcule la similarité $s(r, r_i)$ de r à chaque ressource r_i de C, et on sélectionne ses k plus proches voisins, c'est à dire $r_1, \ldots r_k$ tels que $s(r, r_i)$ soit maximum pour $i=1, \ldots k$. La similarité $s(r, r_i)$ entre deux ressources correspond au nombre de mots pleins en commun entre le titre de r et celui de r_i. Chaque ressource est représentée par un sac de mots issu du titre après filtrage des mots grammaticaux à l'aide d'un anti-lexique (Salton et Mc Gill, 1983). Dans le cas où on cherche à étiqueter les ressources avec une seule classe, la nouvelle ressource peut être étiquetée avec la classe dominante parmi ses k plus proches voisins. Dans le cadre de l'indexation, les classes avec lesquelles nous allons étiqueter les ressources sont des mots clés ou des paires mot clé/qualificatif MeSH. De plus, l'indexation d'une ressource doit être composée d'un nombre *a-priori* inconnu de mots clés (ou paires). Nous associons donc à r une liste de candidats MeSH auxquels un score S (compris entre 1 et k) est attribué en fonction du vote des k voisins. Le choix d'une valeur de k est évoqué à la section 2. La fonction de rupture décrite en 1.2.3 permet de sélectionner les candidats retenus pour l'indexation finale.

1.2.3 Fonction de rupture

Soit N le nombre de mots clés (ou paires) candidats à l'indexation extraits à l'aide de l'une des méthodes ci-dessus. Soit S_i le score attribué au i-ème candidat. On suppose que les candidats sont classés par ordre de scores décroissants, de sorte que $S_1 > \ldots > S_i > \ldots > S_N$. Pour $i = 1, \ldots, N-1$, on calcule $F = \frac{S_i - S_{i+1}}{S_i + S_{i+1}}$. Le seuil retenu sera i tel que F soit maximum.

---

[2] On considère ici que les qualificatifs peuvent être « vides », c'est à dire que les règles CISMeF peuvent statuer sur des mots clés seuls ou associés à un qualificatif.

1.2.4 Méthodes Mixtes

Afin d'évaluer la complémentarité des méthodes présentées ci-dessus (1.2.1 et 1.2.2), nous avons combiné les indexations obtenues selon deux procédés, l'un prenant en compte le rang des mots clés (ou paires) résultant de l'indexation TALN et k-PPV, l'autre prenant en compte le score attribué aux mots clés (ou paires) par ces méthodes. Ainsi, à chaque mot clé (ou paire) est attribué un nouveau score égal à la somme de ses rangs (resp. scores relatifs) dans les deux méthodes. Si un mot clé n'a été extrait que par une seule méthode, son rang (resp. score relatif) pour la deuxième méthode est considéré comme nul. Les mots clés (ou paires) sont ensuite classés par score décroissant. Dans les deux cas, nous avons placé en tête du classement les mots clé (ou paires) extraits conjointement par les deux méthodes.

2 Expérimentation

Le corpus d'évaluation utilisé est composé de 82 ressources extraites aléatoirement du catalogue CISMeF. Chaque ressource du corpus a été indexée automatiquement à l'aide des méthodes TALN et k-PPV successivement. Les résultats obtenus par les deux méthodes ont ensuite été combinés comme indiqué en 1.2.4. Dans chaque cas, l'indexation automatique obtenue a été comparée à l'indexation manuelle de référence. Les performances sont évaluées avec les mesures habituelles de précision et de rappel, ainsi que la F-measure qui combine ces deux dernières. Un poids équivalent est alors accordé à la précision et au rappel (Manning et Schütze, 1999). La Figure 1 présente la F-measure en fonction du rang pour chacune des méthodes.

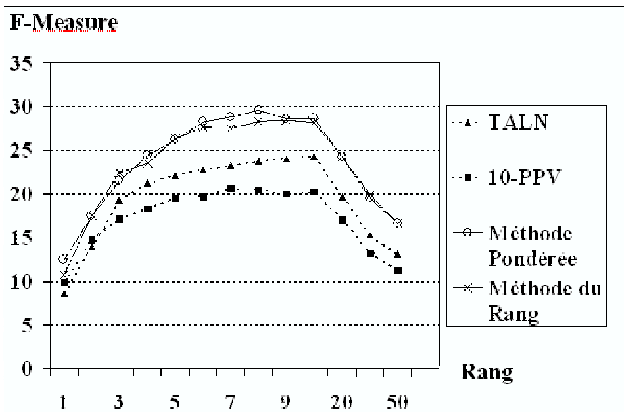


Figure 1 : Courbes F-measure en fonction du rang

Le Tableau 1 présente la précision et le rappel obtenus pour chaque méthode séparément (colonnes 1 à 3), et pour la combinaison des méthodes fondée sur le rang et sur les scores (colonnes 4 et 5). Pour la méthode des k-PPV, plusieurs valeurs de k ont été testées (k=1, 3, 5, 10, 15) et les meilleurs résultats ont été observés avec k=10, bien que pour certaines ressources, il ne soit pas possible de trouver dix voisins (il arrive même qu'aucun voisin ne soit trouvé). Dans ce cas, les valeurs de précision et de rappel de la méthode sur une telle ressource sont considérées comme nulles. Dans la troisième colonne (N=73) nous donnons les performances obtenues pour les soixante treize ressources pour lesquelles au moins dix voisins ont été trouvés.

Rang	TALN	10-PPV (N=82)	10-PPV (N=73)	Mixte Rang	Mixte Score
	P - R	**P - R**	**P - R**	**P - R**	**P - R**
1	36 – 5	51 - 6	57 - 6	51 - 6	49 - 6
4	32 – 16	30 – 13	34 – 15	38 - 17	36 - 18
10	22 – 27	20 – 21	22 – 23	26 - 31	27 - 33
50	8 – 40	7 – 35	8 - 40	10 - 53	10 - 53
T	**27 - 21** (T_{moyen}=12)	**32 – 16** (T_{moyen} =5)	**36 – 18** (T_{moyen} =5)	***34 – 24*** (*T_{moyen} =9*)	***32 – 25*** (*T_{moyen} =9*)

Tableau. 1 : Précision et Rappel de chaque méthode sur un corpus de 82 ressources

3 Discussion

3.1 Performances des différentes méthodes

D'après le Tableau 1 et la Figure 1, on peut remarquer que la combinaison des deux méthodes d'indexation présentées offre une précision supérieure ou égale à chacune des méthodes. Rappel et F-measure sont également meilleurs. Cela rejoint les conclusions de (Aronson et al. 2004) qui observent une amélioration des performances du système MTI lorsqu'un module statistique vient compléter le module de traitement linguistique. Les résultats de la combinaison des deux méthodes sont équivalents à ceux des extracteurs MeSH francophones de la littérature (Névéol et al. 2005). Cependant, le système que nous présentons a l'avantage d'indexer à l'aide de *paires* mot clé/qualificatifs alors que les autres sytèmes existants – par exemple, (Pouliquen, 2002) ou (Gaudinat et al., 2002), extraient des termes isolés. La fonction de rupture est efficace dans la mesure où, la précision au seuil T est généralement plus élevée que la précision obtenue au rang fixe équivalent (par exemple, la précision est de 32 à T=5 contre 29 au rang fixe 5 pour la méthode k-PPV). Pour les méthodes combinées, les chiffres sur le seuil (en italique) sont donnés à titre indicatif, et ont été obtenus en effectuant une moyenne entre précision et rappel aux seuils pour les méthodes TALN et k-PPV.

3.2 Complémentarité des méthodes

Le dictionnaire MeSH utilisé par la méthode TALN permet de ne pas limiter l'indexation aux termes MeSH déjà présents dans la base de ressources indexées (~11.000 termes MeSH sur près de 23.000). L'approche k-PPV utilise une indexation manuelle de référence, et permet ainsi de proposer une indexation cohérente avec la base existante. De plus, elle peut extraire des mots clés qui sont pour l'instant difficilement repérables avec la méthode TALN, comme <*étude comparative*>. En effet, ces termes apparaissent rarement dans le texte des ressources et doivent être déduits d'une analyse globale de la ressource. Par exemple, un article faisant état d'une même étude épidémiologique réalisée en France et au Canada pourra être indexé avec le mot clé <*étude comparative*> bien que celui-ci ne soit jamais explicitement utilisé. Ainsi, dès les premiers rangs, le rappel obtenu avec la fusion des méthodes est supérieur au rappel obtenu avec l'une des méthodes utilisée séparément. Cependant, la méthode des k-PPV

est fondée sur le titre des ressources, ce qui peut poser problème si le titre n'est pas suffisamment explicite, ou si la base comporte trop peu de ressources aux titres similaires. Ainsi, on peut observer un silence de la méthode (11% des cas sur le corpus utilisé) ou bien une indexation approximative. L'utilisation conjointe de la méthode TALN permet de minimiser l'impact de ces erreurs. Le fait de placer les termes communs aux deux méthodes en tête de l'indexation combinée permet de mettre en avant des termes qui avaient obtenu un score bas avec les deux méthodes.

4 Conclusion et perspectives

Nous avons présenté deux méthodes d'indexation automatique de documents médicaux inovantes dans la mesure où elles proposent une indexation à l'aide de *paires* de descripteurs MeSH. Ces approches ont été évaluées séparément, puis combinées sur un corpus de 82 ressources. Il apparaît que la fusion des deux approches offre des performances supérieures à chacune des méthodes seules. Pour la poursuite de ce travail, nous préparons une évaluation auprès d'indexeurs professionnels afin de déterminer si la révision de l'indexation automatique proposée par les systèmes combinés permet un gain de temps et/ou une réduction du silence de l'indexation manuelle.

Références

ANDERSON, JD., PÉREZ-CARBALLO, J. (2001) The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part I. *IPM* 37(2), 231-254.

ARONSON AR., MORK JG., GAY CW., HUMPHREY SM., ROGERS WJ (2004). The NLM Indexing Initiative's Medical Text Indexer. Actes de *MEDINFO* 2004; 268-71.

DAILLAND, F. LEUTHEREAU, A. AND VALLEE, H. (2003). Aide mémoire d'indexation MeSH et F-MeSH pour le catalogage. Rapport Technique de l'INSERM. Bibliothèque de la Faculté de Médecine de Paris XI.

GAUDINAT, A., BOYER, C., BAUJARD, V., RUCH P. (2002) Evaluation de l'extraction de termes MeSH pour les systèmes de recherche d'information dans le domaine médical. Actes des JFIM.

MANNING, C., SHÜTZE, H. (1999) Foundations of Statistical NLP, MIT Press.

NEVEOL, A. (2004) Indexation automatique de ressources de santé à l'aide d'un vocabulaire contrôlé. Actes de *RECITAL*. 2004;105-114.

NÉVÉOL, A., MARY V, GAUDINAT, A., ROGOZAN A., DARMONI SJ. (2005) Benchmark evaluation of the French MeSH indexing systems; soumis à AIME 2005.

POULIQUEN B. (2002) Indexation de textes médicaux par indexation de concepts, et ses utilisations, Thèse de Doctorat, Université Rennes 1.

SALTON, G., MC GILL, M.J. (1983) Introduction to Modern Information Retrieval. New York : McGraw-Hill.

Réseau bayésien pour un modèle d'utilisateur et un module de compréhension pour l'optimisation des systèmes de dialogues

Olivier Pietquin

Supélec, Campus de Metz – Equipe STS
2 rue Edouard Belin – F-57070 Metz
olivier.pietquin@supelec.fr

Mots-clés : Systèmes de dialogue, simulation de dialogues, modèle d'utilisateur, optimisation.
Keywords: Spoken dialog systems, dialog simulation, user modeling, optimization

Résumé Dans cet article, un environnement modulaire pour la simulation automatique de dialogues homme-machine est proposé. Cet environnement comprend notamment un modèle d'utilisateur consistant dirigé par le but et un module de simulation de compréhension de parole. Un réseau bayésien est à la base de ces deux modèles et selon les paramètres utilisés, il peut générer un comportement d'utilisateur cohérent ou servir de classificateur de concepts. L'environnement a été utilisé dans le contexte de l'optimisation de stratégies de dialogue sur une tâche simple de remplissage de formulaire et les résultats montrent qu'il est alors possible d'identifier certains dialogues problématiques du point de vue de la compréhension.

Abstract In this paper we present a modular environment for simulating human-machine dialogues by computer means. This environment includes a consistent goal-directed user model and a natural language understanding system model. Both models rely on a special Bayesian network used with different parameters in such a way that it can generate a consistent user behaviour according to a goal and the history of the interaction, and been used as a concept classifier. This environment was tested in the framework of optimal strategy learning for the simple form-filling task. The results show that the environment allows pointing out problematic dialogues that may occur because of misunderstanding between the user and the system.

1 Introduction

Dans cet article, nous traitons essentiellement de simulation de dialogues homme-machine. Initialement, les systèmes de simulation étaient destinés essentiellement à la validation de modèles du discours (Power, 1979). Avec l'apparition des interfaces vocales sont aussi arrivés les problèmes de conception. La conception de ces interfaces est un processus cyclique dans lequel interviennent successivement des phases de développement, de tests, d'évaluations et d'améliorations. La phase la plus sujette aux contraintes de temps et d'argent et bien souvent celle de l'évaluation et de test. Pour cette raison, la simulation en vue de l'évaluation automatique des interfaces s'est répandue depuis la fin des années 1990 (Eckert et al., 1998). De cette combinaison de la simulation et de l'automatisation de l'évaluation a assez vite découlé une nouvelle application : l'apprentissage automatique de stratégies optimales (Levin, Pieraccini, 1997) (Singh et al., 1999). Dans cet article, un environnement de simulation de dialogues est proposé dans le cadre de cette dernière application.

De tels environnements existent donc déjà. Certains utilisent des modèles statistiques de transitions entre états obtenus d'après observation de dialogues réels, (Singh et al., 1999). D'autres utilisent un modèle d'utilisateur sans mémoire (Levin, Pieraccini, 1997) et n'incluent pas de modélisation de l'erreur. Ici, nous décrivons un environnement de simulation comprenant un modèle d'utilisateur consistant étant donné l'historique de l'interaction (avec mémoire) et un but. Cet environnement comprend aussi un modèle de système de reconnaissance vocale ainsi qu'un module simulant la compréhension du langage naturel. En incluant ces modules dans l'environnement, nous espérons que les stratégies apprises tiendront comptes de leurs lacunes.

2 Un modèle formel pour le dialogue vocal homme-machine

De manière formelle et comme le décrit la Figure 1, un dialogue vocal homme-machine peut être considéré comme un processus séquentiel dans lequel un utilisateur humain et un système de gestion de dialogue (DM : *Dialogue Manager*) communiquent grâce à la parole au travers d'un canal de transmission. Ce canal est composé de différents modules qui manipulent chacun l'information pour lui faire prendre une forme utilisable par le ou les modules suivants. Le but d'un système de dialogue étant souvent de fournir de l'information à l'utilisateur, le système de gestion de dialogues peut donc accéder à une base de connaissance.

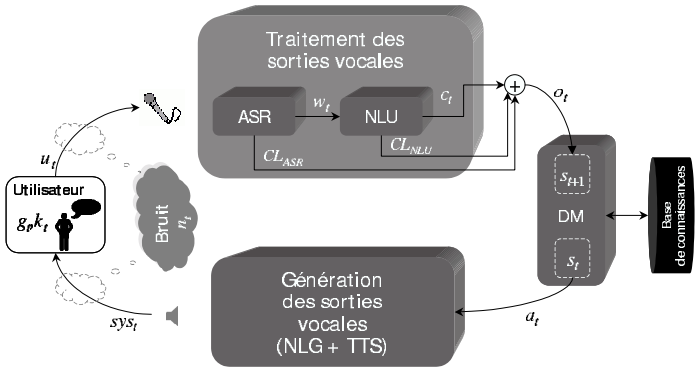


Figure 1 : Modèle de dialogue vocal homme-machine

Le processus étant séquentiel, il peut être discrétisé en *tours t*. A chaque tour, le gestionnaire de dialogue génère un ensemble d'*actes de communication* a_t sur base de son *état* interne s_t pouvant se matérialiser en une invite, une question, une aide, une demande de confirmation, la fermeture du dialogue etc. Afin d'être compris par l'utilisateur, cet ensemble est transformé en un signal de parole sys_t par les systèmes de génération de sorties vocales. En fonction de ce qu'il a pu comprendre de ce signal, de sa *connaissance* au moment t (k_t) et du *but* qu'il poursuit en communiquant avec le système (g_t), l'utilisateur produit à son tour un signal de parole u_t. Dans le cas particulier des systèmes de dialogue, le terme 'connaissance' peut faire référence à la connaissance de l'utilisateur concernant l'*historique* de l'interaction, la *tâche*, le *système* lui-même ou le *monde* en général. Les deux signaux vocaux u_t et sys_t sont entachés par le *bruit* ambiant n_t au moment de leur production. Le système de reconnaissance vocale (ASR) traite alors le signal u_t et le transforme en un ensemble de *mots* w_t. Au passage, le module ASR produit une mesure CL_{ASR} indiquant le degré de *confiance* qu'il accorde à son résultat. L'ensemble w_t est ensuite passé au système de compréhension de parole (NLU) qui doit en retirer une représentation sémantique que nous supposerons mise sous la forme d'un ensemble de *concepts* c_t. Le module NLU produit lui-aussi une mesure de confiance CL_{NLU} associée à l'ensemble c_t. L'ensemble $\{c_t, CL_{ASR}, CL_{NLU}\}$ compose une *observation* o_t qui est utilisée pour réaliser une mise à jour de son état interne. D'un point de vue probabiliste, le comportement de l'utilisateur peut être résumé par la probabilité conjointe suivante :

$$\begin{aligned} P(u,g,k|sys,a,s,n) &= \underbrace{P(k\,|\,sys,a,s,n)}_{\text{MAJ de connaissance}} \cdot \underbrace{P(g\,|\,k,sys,a,s,n)}_{\text{Modification du but}} \cdot \underbrace{P(u\,|\,g,k,sys,a,s,n)}_{\text{Sortie utilisateur}} \\ &= \underbrace{P(k\,|\,sys,s,n)}_{\text{MAJ de connaissance}} \cdot \underbrace{P(g\,|\,k)}_{\text{Modification du but}} \cdot \underbrace{P(u\,|\,g,k,sys,n)}_{\text{Sortie utilisateur}} \end{aligned} \quad (1)$$

Les simplifications dans (1) tiennent compte de plusieurs faits, notamment on peut raisonnablement admettre que la connaissance de l'utilisateur n'est pas modifiée par l'acte a puisque l'utilisateur n'a pas accès directement à cette valeur. De même, sa réponse ne dépend ni de l'acte a qu'il ne connait

pas, ni de l'état s qu'il a du intégrer dans sa connaissance de l'historique de l'interaction. Enfin, une modification du but de l'utilisateur doit passer par une modification de sa connaissance uniquement. Les trois termes de (1) mettent en évidence les relations étroites qui existent entre le processus de production de parole et le couple {but, connaissance}. Néanmoins, la modification de la connaissance est un processus incrémental (mise à jour) et se base donc aussi sur la connaissance préalable de l'utilisateur :

$$\begin{aligned} P(k \mid sys, s, n) &= \sum_{k^-} P(k \mid k^-, sys, s, n) \cdot P(k^- \mid sys, s, n) \\ &= \sum_{k^-} P(k \mid k^-, sys, n) \cdot P(k^- \mid s) \end{aligned} \qquad (2)$$

Ici, k^- représente la variable k_{t-1}. La simplification du second facteur de la somme provient du fait évident que la connaissance de l'utilisateur au temps t-1 ne peut pas dépendre des signaux de parole ou de bruit au temps t.

3 Le modèle d'utilisateur

3.1 Un réseau bayésien dynamique

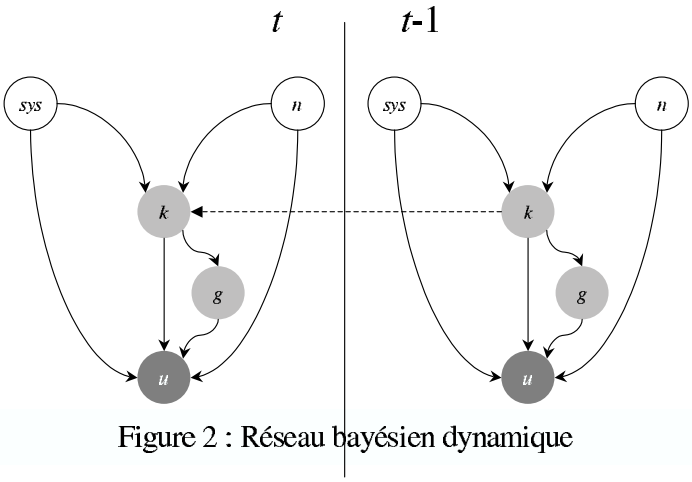


Figure 2 : Réseau bayésien dynamique

Les équations (1) et (2) permettent de dire qu'un réseau bayésien dynamique (DBN : Dynamic Bayesian Network) pourrait encoder la factorisation particulière des probabilités associées à l'utilisateur (Pearl, 1988). Les nœuds du réseau sont donnés par les variables présentes dans les équations (*sys*, *n*, *k*, *g*, *u*) et les arcs sont donnés par les probabilités conditionnelles. La consistance de tour en tour est assurée par la dépendance dans le temps de la variable *k*. Le réseau dynamique obtenu est montré sur la Figure 2. Les variables *sys* et *n* sont des variables extérieures à l'utilisateur (cercles vides), les variables *k* et *g* sont des variables internes (cercles gris-clair) et la variable *u* est une variable de sortie (cercles gris-foncé).

3.2 Utilisation du Modèle

Le DBN de la Figure 2 paraît relativement simple, néanmoins la définition des variables qu'il fait intervenir est plus ou moins floue. Ici, nous avons choisi une représentation des variables en paires « *attribut-valeur* » (paires AV) dérivées de la description en « *Matrice attribut-valeur* » de la tâche. Dans ce cadre, chaque acte de communication est considéré comme un ensemble de paires AV. Dans ce qui suit, Le signal de parole *sys* émis par le système est alors modélisé par un ensemble de pairs AV dont l'ensemble des attributs, noté $S=\{s^\sigma\}$, contient des éléments qui peuvent prendre des valeurs booléennes indiquant si oui ou non l'attribut associé est présent dans *sys*. Un attribut spécial non booléen A_S sera inclus à S et sa valeur définira le type d'acte de communication associé à *sys*. Les types acceptés peuvent être 'invite', 'question', 'demande de relaxation', 'proposition', 'demande de confirmation', 'fermeture du dialogue', … Une question directe sera alors caractérisée par un attribut A_S égal à 'question' et un seul attribut s^σ dont la valeur sera *vraie*. La réponse u de l'utilisateur sera modélisée par une autre paire AV dans laquelle les attributs appartiennent à $U = \{u^\upsilon\}$ et l'ensemble des valeurs possibles pour chaque attribut u^υ sera note $V = \{ v_i^\upsilon \}$. Un attribut spécial C_U est ajouté à U et sa valeur booléenne indique si l'utilisateur a décidé de clore le dialogue dans sa réponse. Le but et la

connaissance de l'utilisateur seront représentées respectivement par les paires $G = \{[g^\gamma, gv_i^\gamma]\}$ et $K = \{[k^\kappa, kv_i^\kappa]\}$ ou g^γ et k^κ sont des attributs et gv_i^γ et kv_i^κ sont les valeurs possibles. En fonction de ces nouvelles notations, le réseau de la Figure 2 devient celui de la Figure 3 ou la dépendance dans le temps a été volontairement omise pour plus de clarté ainsi que le bruit dont la modélisation est trop complexe. Chaque valeur ou *état* possible pour chaque variable de ce réseau est une combinaison des attributs et des valeurs, ce qui signifie que les états sont discrets et en nombre fini. On peut donc définir une version factorisée de ce réseau dans laquelle figureraient les variables $A_S, s^\sigma, v_i^\sigma, u^\upsilon, v_i^\upsilon, g^\gamma, gv_i^\gamma, k^\kappa, kv_i^\kappa$ et U_C.

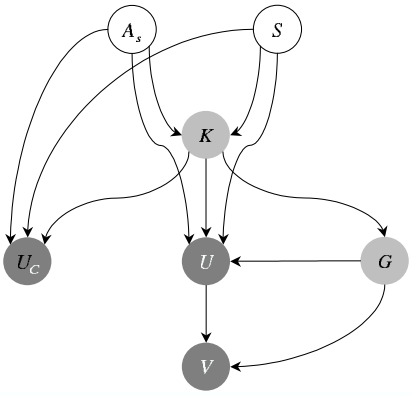


Figure 3 : Réseau bayésien basé sur les paires AV

Considérons une tâche simple consistant à remplir un formulaire composé de deux entrées : $S = \{s^1, s^2\}$. Le système peut utiliser 4 types d'actes de communication : 'invite', 'question directe', 'demande de confirmation' et 'fermeture'. Pour simplifier, considérons que la connaissance de l'utilisateur se compose de simples compteurs, chacun associé à un élément de S, initialisés à 0 et qui sont incrémentés à chaque fois que le système pose une question ou demande une confirmation sur l'entrée associée. Ceci est suffisant pour permettre au modèle d'utilisateur de rester consistant par rapport à l'historique de l'interaction et de réagir à un comportement insatisfaisant du système (en réagissant lorsqu'une entrée a été demandée plusieurs fois). Le but de l'utilisateur est alors de transmettre au système les valeurs correctes pour les attributs représentés par les entrées du formulaire (Figure 4).

Goal		Know.
Att.	Val.	Count
g^1	gv_1	k^1
g^2	gv_2	k^2

Figure 4 : But et connaissance de l'utilisateur

L'utilisateur peut donc inclure dans ses réponses u les deux attributs u^1 et u^2 (il y a autant d'attributs dans U que dans S). Afin de simuler la réponse de l'utilisateur à l'invite, il suffit alors d'entrer l'évidence suivante dans le moteur d'inférence :

A_S	k^1	k^2	g^1	g^2	gv^1	gv^2
invite	0	0	1	1	gv_1	gv_2

Figure 5 : Evidence pour une réponse à l'invite

Les valeurs 1 associées aux variables g^i signifient que les attributs g^i sont bien présents dans le but. Grâce à cette évidence, le moteur d'inférence produira les probabilités $P(u^1=1)$, $P(u^2=1)$, $P(U_C=1)$ et leurs compléments. Tout d'abord, le modèle choisit de manière aléatoire un nombre réel entre 0 et 1, si ce nombre est inférieur à $P(U_C=1)$, le dialogue est clos. Dans le cas contraire, le même processus est répété pour choisir les attributs présents dans la réponse de l'utilisateur. En supposant que u^1 est sélectionnée pour être présente dans la réponse de l'utilisateur, l'évidence suivante est alors entrée dans le moteur d'inférence :

u^1	u^2	gv^1	gv^2
1	0	gv_1	gv_2

Figure 6 : Inférence pour une valeur de réponse

4 Simulation de la compréhension de parole

La simulation de NLU peut se faire en utilisant le réseau bayésien décrit plus haut comme classificateur. Pour ce faire, nous considèrerons que les erreurs de reconnaissances vocales n'affectent que les valeurs des paires AV alors que les erreurs d'associations attribut-valeur sont dues au module

de compréhension. En considérant que le processus de reconnaissance vocale a transformé les valeurs $V = \{ v_i^v \}$ générées dans sa réponse u par le modèle d'utilisateur en un ensemble de valeur $W = \{w_j\}$ et en reprenant l'exemple simple du remplissage de formulaire expliqué dans la section précédente, les évidences suivantes peuvent être introduites dans le moteur d'inférence pour simuler la compréhension de la réponse à l'invite :

A_S	s^1	s^2	with	v^1_1	or	V^1_2
invite	0	0		w_j		w_j

Figure 7 : Evidence pour la compréhension de la réponse à l'invite

A moins que w_j ne soit pas une valeur acceptable pour un des attributs testés, ces deux différentes évidences vont fournir des valeurs pour les probabilités $P(u^1 \mid A_S = \text{greet}, v^1_1 = w_j)$ and $P(u^2 \mid A_S = \text{greet}, v^1_2 = w_j)$. Le système de simulation de compréhension va alors affecter la valeur w_j à l'attribut u^i ayant produit la probabilité la plus haute. Des situations plus complexes peuvent évidemment être rencontrées mais il est toujours possible de les transformer en évidence utilisable par le moteur d'inférence. Cette méthode peut aussi produire une sorte de niveau de confiance de compréhension. Dans le cas de la classification d'une seule valeur, le niveau de confiance de compréhension est simplement la probabilité fournie par le moteur d'inférence. Lorsque plusieurs valeurs ont du être associée à des attributs par le module de compréhension, une mesure de confiance peut être affectée à chaque paire ou une mesure globale peut être donnée en multipliant toutes les valeurs.

5 Apprentissage de stratégies optimales par simulation

Le modèle décrit ci-dessus a été développé dans le but de l'apprentissage automatique de stratégies de dialogue homme-machine optimales. Nous avons donc mis notre environnement en présence d'un agent d'apprentissage par renforcement comme proposé dans (Levin, Pieraccini, 1997). Pour se faire, il faut définir un critère d'optimisation. On peut en trouver plusieurs dans la littérature néanmoins, l'hypothèse selon laquelle la contribution de chaque acte à la satisfaction de l'utilisateur est une bonne mesure de l'évaluation d'une stratégie est retenue ici. Selon (Singh et al, 1999) une fonction de coût basée sur une mesure de la complétion de la tâche, les performances de reconnaissance et de compréhension et la durée en tours du dialogue serait satisfaisante. Dans notre expérience, les utilisateurs sont invités à fournir des informations à propos d'un voyage en train. Les attributs sont donc une ville de départ, une ville de destination, une heure de départ, une heure d'arrivée désirée et la classe. Il y a 50 valeurs possibles pour les villes (les mêmes pour le départ et l'arrivée) et les heures possibles sont les heures plaines (de 0 à 24). Les types d'actes de communications possibles sont 'invite', 'question directe', 'question ouverte', 'confirmation explicite' et 'fermeture du dialogue'. Nous réalisons plusieurs expériences différentes dans lesquelles l'agent d'apprentissage évolue dans un espace d'état construit sur base de l'historique de l'interaction et d'une valeur binaire indiquant si le niveau de confiance de la dernière interaction est *haut* ou *bas*. Les expériences varient entre autre par la définition du niveau de confiance qui peut être uniquement CL_{ASR} (espace d'états S_1 dans la suite) et CL_{ASR}*CL_{NLU} (espace d'états S_2 dans la suite). De même la fonction de coût intègre l'une ou l'autre mesure de confiance. Au début de chaque dialogue, un but d'utilisateur est construit assignant des valeurs aux 5 attributs. La mesure de complétion de la tâche est alors définie comme le rapport entre le nombre d'attributs dont la valeur a été correctement assignée au nombre d'attributs en tout (5 ici). On définit aussi deux environnements de simulation. Le premier (Sim_1) intègre le modèle d'utilisateur et un module de simulation de reconnaissance vocale introduisant des erreurs et une mesure de confiance de reconnaissance. Le second environnement (Sim_2) intègre, en plus, le module de compréhension. Nous avons réalisé trois expériences différentes en combinant différemment les espaces d'états et les environnements de simulation. Les résultats de l'apprentissage sont montrés dans les tableaux de la Figure 8. Dans le tableau de gauche sont indiqués les résultats des mesures objectives pouvant être obtenues lors d'un dialogue moyen suivant la stratégie apprise (mesures obtenues en calculant la moyenne des mesures faites sur 10 000 dialogues simulés). Dans le tableau de droite sont indiquées les fréquences moyennes d'occurrences de chaque type d'acte de communication.

	N	*TC*
Sim_1, S_1	5.39	0.81
Sim_2, S_1	7.03	0.74
Sim_2, S_2	5.82	0.79

	invite	constQ	openQ	expC	Close
Sim_1, S_1	1.0	0.85	1.23	1.31	1.0
Sim_2, S_1	1.0	1.25	1.18	2.60	1.0
Sim_2, S_2	1.0	1.05	1.18	1.58	1.0

Figure 8 : Résultats de l'expérience

Grâce aux tableaux de la Figure 8, nous pouvons conclure que lors de la première expérience (sans erreur de compréhension), il y a plus de question ouvertes que de questions directes. Les erreurs de reconnaissances étant prises en compte par l'introduction de CL_{ASR} dans S_1 et Sim_1, il y a souvent des demandes de confirmations. Dans la deuxième expérience, des erreurs de compréhensions sont introduites mais elles ne peuvent pas être détectées par les mesures de confiance. On observe une augmentation du nombre de confirmations puisque le système ne peut jamais être certain que les valeurs sont bien assignées. La longueur moyenne du dialogue s'en trouve augmentée et la complétion de la tâche diminue. En ajoutant CL_{NLU} dans S_2, les performances s'améliorent et on retrouve presque les résultats de la première expérience. Ceci est du au fait que certaines questions ouvertes sont évitées parce qu'elles résultent en une très mauvaises mesure de confiance. En effet la stratégie est modifiée et les questions ouvertes concernant les deux villes en même temps sont très peu probables car elles induisent des confusions et des niveaux de confiance plus faibles.

6 Conclusions et perspectives

Dans cet article, un environnement de simulation de dialogues dans lequel ont été introduit un modèle d'utilisateur consistant et un module de simulation de compréhension de parole a été décrit. Cet environnement a été développé dans le but d'un 'apprentissage de stratégies de dialogues optimales et il a pu être démontré par expérience que cet environnement permettait de mettre en évidence des problèmes éventuels de compréhension et d'adapter la stratégie automatiquement en conséquence. Quelques particularités de l'environnement n'ont pas été exploitées dans ce travail et il serait probablement intéressant de s'y atteler dans le futur. Par exemple, la relation avec le fonctionnement parallèle de l'utilisateur et le gestionnaire de dialogue et le phénomène de *grounding* intervenant dans les dialogues homme-homme a été brièvement mentionné dans la section 2 mais n'a pas vraiment été exploitée. Le besoin d'introduire des sous-dialogues permettant la mise en phase des connaissances supposées de l'utilisateur et de l'état réel du gestionnaire pourrait être détecté par la l'inconsistance entre l'état du système et des valeurs inférées de la connaissance de l'utilisateur.

Références

ECKERT W., LEVIN E., PIERACCINI R. (1998) Automatic Evaluation of Spoken Dialogue Systems, *Technical Report TR98.9.1, AT&T Labs Research.*

LEVIN E., PIERACCINI R. (1997), A Stochastic Model of Computer-Human Interaction for Learning Dialogue Strategies, *Proc. Eurospeech'97*, Rhodes, Greece, pp. 1883-1886.

PEARL J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers, Inc. San Francisco, California.

PIETQUIN O., DUTOIT T. (2002) Modélisation d'un Système de Reconnaissance dans le Cadre de l'Evaluation et l'Optimisation Automatique des Systèmes de Dialogue, *Actes des Journées d'Etude de la Parole, JEP 2002*, Nancy (France).

POWER R. (1979) The Organization of Purposeful Dialogues, *Linguistics 17*, pp. 107-152.

SINGH S., KEARNS M., LITMAN D., WALKER M., (1999) Reinforcement Learning for Spoken Dialogue Systems, *Proc. NIPS'99*, Denver, USA.

Correction automatique en temps réel
Contraintes, méthodes et voies de recherche

Roger Rainero

Société Diagonal
1, Traverse des Brucs – Valbonne — Sophia Antipolis
roger.r@prolexis.com

Mots-clés :

Correction automatique, temps réel, analyse syntaxique, grammaire de contraintes.

Résumé

Cet article expose un cas concret d'utilisation d'une grammaire de contraintes. Le produit qui les applique a été commercialisé en 2003 pour corriger automatiquement et en temps réel les fautes d'accord présentes dans les sous-titres des retransmissions en direct des débats du Sénat du Canada. Avant la mise en place du système, le taux moyen de fautes était de l'ordre de 7 pour 100 mots. Depuis la mise en service, le taux d'erreurs a chuté à 1,7 %.

Nous expliquons dans ce qui suit les principaux atouts des grammaires de contraintes dans le cas particulier des traitements temps réel, et plus généralement pour toutes les applications qui nécessitent une analyse au fur et à mesure du discours (c.-à-d. sans attendre la fin des phrases).

Keywords:

Automatic correction, real-time, syntactic analysis, grammar of constraints.

Abstract

This article sets out a concrete use case of a grammar of constraints. The product which applies them was commercialised in 2003 to automatically correct in real time the errors of agreement present in the sub-titles of live televised debates from the Senate of Canada.

Before the introduction of this system, the average rate of mistakes was in the order of 7 per 100 words. With the introduction of this system, the rate of errors has fallen to 1.7%.

In the following section, we explain the main advantages of a grammar of constraints in the specific case of real-time processing, and more generally for all applications which require an analysis during the speech (that is, without waiting until the end of sentences).

1. Exposé du problème

Le Sénat du Canada diffuse certains de ses débats en direct sur une chaîne de télévision spécialisée. Chaque sénateur pouvant s'exprimer dans sa langue maternelle (français ou anglais), les interventions se succèdent indifféremment dans ces deux langues. Les téléspectateurs ont la possibilité d'afficher des sous-titres, soit en français, soit en anglais, mais lorsqu'une langue est choisie, la totalité des débats est transcrite dans cette langue (fonction légale pour les malentendants).

Les sous-titrages français sont obtenus par retranscription sténotypée soit directe (locuteur français) soit indirecte (locuteur anglais traduit simultanément en français, la sténotypiste enregistrant alors la traduction). Les sténotypistes francophones et anglophones utilisent la même méthode de saisie mise au point en Amérique du Nord, très performante pour les langues globalement phonétiques (où la majorité des lettres se prononcent). Cette méthode donne ainsi d'excellents résultats en anglais. Mais pour le français qui comporte de nombreuses syllabes finales muettes, les ajustements ont été longs et fastidieux, et la mise en ondes a été maintes fois repoussée, à la recherche d'un taux acceptable de transcription exacte. Les résultats ont régulièrement progressé jusqu'en 2002, où ce taux a plafonné aux alentours de 93 %. (sur 100 mots, seuls 93 étaient corrects).

Bien que ce taux paraisse très élevé, il génère un nombre d'incidents de lecture très au-delà de ce qui est acceptable. Il suffit, pour s'en convaincre, de constater qu'il correspond à 8 fautes par minute de lecture. Le Sénat du Canada a alors fait un appel d'offres international dans le but de trouver une solution automatique susceptible d'améliorer cette situation. La solution proposée devait permettre de corriger automatiquement le plus grand nombre de fautes résiduelles possibles, sans ajouter de fautes là où il n'y en a pas. Par ailleurs, l'automate devait s'intercaler dans le processus d'acquisition du texte sténotypé (logiciel Eclipse déjà installé) sans le ralentir de façon notable.

La société Diagonal a soumissionné en proposant une adaptation spécifique de ses moteurs d'analyse déjà utilisés dans les logiciels de correction ProLexis et Myriade. Cette solution retenue par le Sénat a été livrée en octobre 2003.

2. Exigences dynamiques de la correction automatique des sous-titrages en temps réel

Le système demandé par le Sénat imposait de faire les corrections au fur et à mesure de la saisie, c'est-à-dire sans que l'on puisse attendre la fin des phrases. Cette obligation vient essentiellement du direct : les sous-titres suivent à peu près les paroles des orateurs. En théorie donc, les corrections doivent être faites quasi immédiatement après les fautes.

En pratique, nous disposions des souplesses suivantes :

- les diffusions sont en léger différé d'une à deux secondes,
- les sous-titres sont découpés en lignes qui ne partent à l'antenne que lorsqu'elles sont pleines.

Exemple avec la phrase : *« Les filles jouent aux billes, les garçons jouent au ballon. »*

Voici ce que saisit la sténotypiste par tranche de 0,5 seconde (avec les fautes) :

0,5 s	*Les*
1,0 s	*Les fille*
1,5 s	*Les fille joue*
2,0 s	*Les fille joue au*
2,5 s	*Les fille joue au bille,*

3,0 s *Les fille joue au bille, les*
3,5 s *Les fille joue au bille, les garçon*
4,0 s *Les fille joue au bille, les garçon joue*
4,5 s *Les fille joue au bille, les garçon joue au*
5,0 s *Les fille joue au bille, les garçon joue au ballon.*

Et voici ce que doit voir le téléspectateur (entre parenthèses, les corrections à faire) :

0 s (rien)
3 s LES FILLE(S) JOUE(NT) AU(X) BILLE(S),
6 s LES GARÇON(S) JOUE(NT) AU BALLON.

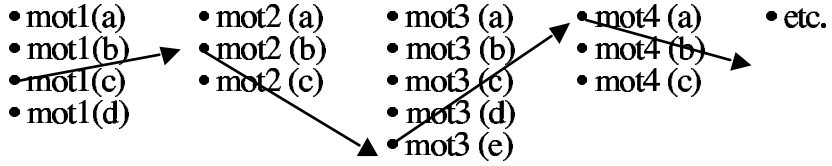
L'automate doit corriger les fautes de la 1re ligne au plus tard au temps 5,0 (temps réel 3,0 + 2 secondes de différé), c'est-à-dire lorsque le 4e mot de la ligne suivante vient d'arriver. Pour la première faute (fille), il dispose d'un retard de 7 mots, mais pour la dernière faute (bille), il ne bénéficie plus que d'un retard de 4 mots. L'automate ignorant totalement à quel moment les lignes sont déclarées « pleines », il est obligé de s'astreindre à faire toutes ses corrections avec un maximum de quatre mots de retard !

C'était bien là la plus grande difficulté à laquelle nous allions être confrontés.

3. Les atouts des systèmes d'analyse basés sur les contraintes

Ce qui a rendu la chose possible avec ProLexis tient dans le fait essentiel que notre moteur exploite un principe de propagation de contraintes.

Comme dans les principaux systèmes basés sur la satisfaction de contraintes (cf. (Blache00)), notre approche ne vise pas à construire un arbre syntaxique de la phrase, mais plutôt à optimiser, dans un réseau de contraintes, le chemin menant du premier au dernier mot de la phrase :



Sur ce schéma, les contraintes ne sont pas figurées. Elles ne se manifestent qu'au travers du choix du chemin affiché qui est censé satisfaire le maximum d'entre elles.

Chaque nouveau mot apporte son lot de variantes possibles, mais aussi son lot de contraintes potentielles pour toutes les variantes établies depuis le début de la phrase :

→ Appliquer une contrainte revient à calculer son coefficient de satisfaction (son « poids ») dans toutes les variantes établies.

→ Propager une contrainte revient à reconsidérer l'application des contraintes déjà appliquées, comme conséquence de l'arrivée du nouveau jeu de contraintes.

Notre système ne construit donc pas un arbre, mais il pondère un réseau. À la fin de la phrase, un algorithme simple peut restituer l'arborescence de la structure syntaxique, si le besoin s'en fait sentir, mais cela n'est pas nécessaire pour diagnostiquer les erreurs et les corriger.

La pondération est souple, dans la mesure où une contrainte mal satisfaite n'est qu'une indication d'un écart par rapport à la norme formalisée par cette contrainte. En ce sens, elle produit des analyses robustes qui s'accommodent de déviances parfois fortes par rapport aux usages ou même de l'absence de certains mots.

Tout cela est bien connu et caractérise les grammaires basées sur les contraintes, mais n'est pas déterminant dans le cas qui nous intéresse.

Le plus gros avantage de notre grammaire, dans le cadre de la correction automatique en temps réel, provient de sa capacité à délivrer une analyse pondérée des variantes après chaque mot. Bien sûr, l'analyse est réputée optimale lorsque tous les mots de la phrase sont connus, mais cette analyse est néanmoins disponible en phase intermédiaire après chaque mot.

Enfin, le mécanisme de propagation après chaque mot présente un autre avantage de taille dans le cas de notre application « temps réel » : il permet, par un choix judicieux des coefficients de pondération, de marginaliser assez vite certaines variantes isolées et de limiter à un niveau raisonnable le nombre de variantes concurrentes que le logiciel évalue en parallèle à chaque itération.

Ce mécanisme peut même s'autoréguler en détruisant systématiquement les variantes les moins probables à chaque passe, de telle sorte que leur nombre total reste en dessous d'un seuil critique pour le temps d'exécution.

Évidemment, toutes ces actions aveugles sont préjudiciables à la qualité de l'analyse *in fine*. Toute la question était de quantifier leur influence réelle sur la fiabilité des corrections attendues.

4. Tests de fiabilité des résultats intermédiaires à « mot + 4 »

Jusqu'à présent, nous n'avions jamais testé la fiabilité des résultats intermédiaires avant la fin de la phrase. Il nous fallait donc vérifier que dans le contexte du Sénat, nous disposions effectivement de suffisamment d'indices pour décider des corrections au maximum à mot + 4.

En théorie, en effet, tout nouveau mot dans une phrase peut changer totalement son analyse. C'est un exercice bien connu auquel se livrent volontiers les professionnels de l'analyse syntaxique. Et c'est aussi avec de tels exemples que l'on peut démontrer que ce que nous avons fait est impossible. En voici quelques-uns :

Début : *Le chien regarde le chat et la souris…*
Suite 1 : *Le chien regarde le chat et la souris mais…*
Suite 2 : *Le chien regarde le chat et la souris prise…*
Suite 2a : *Le chien regarde le chat et la souris prise… (au piège ?)*
Suite 2b : *Le chien regarde le chat et la souris prise… (du tabac ?)*

Mais quelle est la portée statistique réelle d'un tel phénomène et quelle est son influence sur la fiabilité d'un système de correction automatique après quatre mots ?

Pour l'estimer, nous avons fabriqué un prototype du produit fini simulant le comportement de l'outil d'acquisition Eclipse. Ce logiciel lisait un extrait des débats du Sénat obtenu par sténotypie et l'envoyait signe à signe à l'automate qui gardait trace dans un fichier de sortie de toutes les corrections faites et du moment où elles pouvaient être faites.

Le tableau suivant montre un extrait de ce fichier de sortie, concernant un début de phrase telle qu'elle est délivrée par Eclipse, fautes comprises (colonne de gauche). À droite, les fautes corrigées sont intercalées après le mot qui les rend possibles :

Après la saisie de…	Les corrections suivantes sont faites…
Il faut également des solution pratique qui…	solution ➔ solutions pratique ➔ pratiques
soit sensé…	soit ➔ soient
pour…	sensé ➔ sensées
ceux qui travaille sur…	travaille ➔ travaillent

On constate que les deux premières erreurs « solution » et « pratique » sont corrigées dès la saisie de « qui », donc respectivement à mot + 2 et mot + 1.

Regardons de plus près les analyses qui sont faites à ce stade : le mot « pratique » est ambigu : ce peut être un nom, un adjectif ou un verbe. Toutes ces formes génèrent potentiellement autant de variantes. Les variantes verbales paraissent improbables, mais, comme toujours en pareil cas, une petite réflexion permet d'en découvrir certaines formes légitimes :

« Il faut également des solutions, pratique ce sport et tu verras ! »
« Il faut également des solutions(,) explique César… »

Bien sûr, la virgule semble cruciale, mais la pondération de son absence n'est pas suffisante ici. En revanche, l'arrivée du mot suivant « qui » est déterminante, elle rend la flexion verbale pour le mot « pratique » quasiment impossible. En théorie, la flexion verbale ne peut être totalement exclue, car l'hypothèse d'oubli d'un mot peut toujours la justifier. Mais les réglages actuels de nos seuils de probabilités pour des textes de provenance sténotypée font qu'elle est rejetée ici.

La suite est compréhensible : déterminé nominal, le groupe précédent est fautif sur l'accord GN. Deux formes correctes sont possibles : « une solution pratique » et « des solutions pratiques ». C'est là qu'interviennent des automates spécialisés dans la correction automatique spécifique du Sénat du Canada : ces automates choisissent de façon probabiliste la correction au pluriel.

Appliqué au texte de référence de 2 000 mots fourni par le Sénat, contenant 149 fautes et correspondant à un débat réel de 20 minutes, le prototype a donné les résultats suivants :

Nombre de fautes corrigées :	…avec un retard de :
89	1 mot
17	2 mots
3	3 mots
1	4 mots

La faible incidence des situations ambiguës sur les corrections automatiques envisagées pour les débats du Sénat du Canada paraissait donc confirmée, au moins sur le texte étudié. Et la propagation de contraintes montrait là une capacité tout à fait étonnante à résoudre le problème posé. Restaient à démontrer son efficacité et sa stabilité à grande échelle.

5. Méthode d'évaluation à grande échelle.

L'inconvénient des systèmes probabilistes est que leur comportement ne peut être totalement déduit de tests à petite échelle. Typiquement, dans le cas du Sénat, la nature même des débats influe grandement sur le vocabulaire, les intervenants et donc sur les types de phrases prononcées. Nul doute qu'un simple échantillon de 2 000 mots ne pouvait représenter correctement la totalité des situations auxquelles devrait faire face l'automate après sa mise en service.

Après avoir été choisis par le Sénat du Canada pour exécuter le marché, nous avons donc lancé en parallèle les deux réalisations suivantes : d'une part, le logiciel lui-même (bien entendu), et d'autre part, l'étalonnage d'un corpus de 50 000 mots destiné à valider les tests d'usine du logiciel, avant sa livraison chez le client.

Ce corpus a été extrait des transcriptions sténotypées de débats récents représentant un peu plus de 10 heures d'antenne réparties sur une période de deux mois. On ne s'est limité à cette taille que pour des contraintes de temps. Deux personnes ont travaillé pendant un mois pour sélectionner les textes, éliminer les passages en double, détecter les fautes et les baliser dans le texte. Quelque 3 500 fautes y ont été repérées.

L'application de l'automate sur ce corpus a corrigé plus de 2 500 fautes sur 3 500, établissant un taux de reconnaissance à grande échelle stable à 98,31 %. Sur ce même corpus, l'automate n'a introduit que 23 fautes, soit un taux moyen de surcorrection de 1/2100.

6. Perspectives et voies de recherche

Il faut se méfier de l'idée fausse qui consiste à penser que le temps d'exécution n'est pas un problème majeur pour les algorithmes d'analyse automatique. On entend souvent dire : « de toute manière, les machines iront de plus en plus vite et un jour viendra où les algorithmes lents s'exécuteront vite ! ».

Tout cela est vrai, sauf pour les applications « temps réel ». La correction automatique des sous-titrages est un exemple, mais il y en a bien d'autres. La reconnaissance vocale multilocuteur et la traduction simultanée sont deux domaines qui pointent déjà à l'horizon et qui n'attendent que l'émergence de nouvelles technologies linguistiques pour se déployer à grande échelle.

Le découpage du mécanisme d'analyse en strates successives ou le contrôle intégré du nombre de variantes concurrentes que permettent aujourd'hui les grammaires guidées par la satisfaction de contraintes semble être un atout de poids dans les applications temps réels.

Références

Blache P. (2000) Le rôle des contraintes dans les théories linguistiques et leur intérêt pour l'analyse automatique, Actes de *TALN 2000*.

Blache P. (2000) Constraints, Linguistic Theories and Natural Language Processing, *Lecture Notes in Artificial Intelligence,* Springer-Verlag.

Harper M. P., Helzerman R. A. (1995). Extensions to constraint dependency parsing for spoken language processing. *Computer Speech and Language*, Vol. 9 (3), pp 187-234.

Maruyama, H. (1990). Constraint dependency grammar and its weak generative capacity. *Computer Software*.

Les Méta-RCG, un formalisme linguistique non-linéaire : description et mise en œuvre

Benoît Sagot
INRIA - Projet Atoll
Domaine de Voluceau, B.P. 105, 78153 Le Chesnay Cedex, France
benoit.sagot@inria.fr

Mots-clefs : Analyse syntaxique, interface syntaxe-sémantique, grammaires non-linéaires, Grammaires à Concaténation d'Intervalles (RCG)

Keywords: Parsing, syntax-semantics interface, non-linear grammars, Range Concatenation Grammars (RCG)

Résumé Nous présentons dans cet article un nouveau formalisme linguistique qui repose sur les Grammaires à Concaténation d'Intervalles (RCG), appelé *Méta-RCG*. Nous exposons tout d'abord pourquoi la non-linéarité permet une représentation adéquate des phénomènes linguistiques, et en particulier de l'interaction entre les différents niveaux de description. Puis nous présentons les Méta-RCG et les concepts linguistiques supplémentaires qu'elles mettent en œuvre, tout en restant convertibles en RCG classiques. Nous montrons que les analyses classiques (constituants, dépendances, topologie, sémantique prédicat-arguments) peuvent être obtenues par *projection partielle* d'une analyse Méta-RCG complète. Enfin, nous décrivons la grammaire du français que nous développons dans ce nouveau formalisme et l'analyseur efficace qui en découle. Nous illustrons alors la notion de projection partielle sur un exemple.

Abstract In this paper, we present a novel linguistic formalism based on Range Concatenation Grammars (RCG), called *Meta-RCG*. We first expose why non-linearity allows a satisfying representation of linguistic phenomena, and in particular of the interaction between the different levels of description. Then we introduce Meta-RCGs and the extra linguistic concepts they manipulate, while remaining compilable into classical RCGs. Moreover, we show that classical analyses (consituency, dependency, topology, predicate-arguments semantics) can be obtained by *partial projection* of a full Meta-RCG parse. Finally, we describe the grammar for French we develop in this new formalism and the associated efficient parser. We illustrate the notion of partial projection with an example.

1 Introduction

Dans (Sagot & Boullier, 2004), les auteurs montrent que les Grammaires à Concaténation d'Intervalles (Range Concatenation Grammars, ci-après RCG) sont bien adaptées à la description du langage naturel. Toutefois, et malgré la disponibilité d'un analyseur (Boullier, 2004), ils ne présentent pas de système d'analyse complet incluant une grammaire linguistique.

L'objectif de cet article est de décrire un tel système d'analyse. Mais cela ne se limite pas à une grammaire : en réalité, les RCG ne sont pas directement utilisables pour représenter des phénomènes linguistiques. Si le principe qui leur est sous-jacent, celui de la concaténation d'intervalles de la chaîne d'entrée, est adapté à la description du langage naturel, d'autres concepts linguistiques de base sont à prendre en considération, tels les phénomènes d'homonymie, de syntagmes à têtes, de traits, d'extraction ou d'interface syntaxe-sémantique.

Les langages définis par les RCG couvrent tout *PTIME*, c'est-à-dire l'ensemble des langages analysables en temps polynomial. Ceci est dû à une propriété fondamentale des RCG, leur non-linéarité (ou clôture par intersection). Dans la première section, nous montrons que cette non-linéarité permet une description appropriée des réalités linguistiques, alors que la plupart des formalismes développés jusqu'à présent ont un squelette linéaire (car clos par homomorphisme, voir plus bas). En particulier, cette non-linéarité permet la définition d'un formalisme qui étend les RCG en prenant en compte les concepts linguistiques de base cités plus haut, tout en restant convertible en RCG standard. La deuxième section donne un aperçu de ce formalisme, appelé Méta-RCG. Enfin, nous présentons dans la troisième section une grammaire du français écrite en Méta-RCG. Ainsi, nous montrons que la grammaire que nous avons écrite fait interagir intimement tous les niveaux d'analyse linguistique, de la morphologie à la sémantique lexicale en passant par la syntaxe. De ce fait, les analyses classiques (en constituants, en dépendances, en boîtes topologiques, en prédicats sémantiques) peuvent être extraites de nos analyses Méta-RCG par *projection partielle*.

2 Non-linéarité et grammaires pour les langues naturelles

La complexité des langues naturelles est au-delà de celle des Grammaires Non-Contextuelles (CFG). Pour des raisons pratiques et théoriques, tous les formalismes envisagés pour décrire les langues naturelles étendent les CFG, même si la façon dont ils les étendent varie. La plupart reposent sur une architecture à deux niveaux. Ils utilisent tout d'abord un *squelette syntaxique* qui étend lui-même les CFG tout en étant clos par homomorphisme[1]. Au dessus de ce squelette, ils mettent en jeu des structures, dites *décorations*, qui sont calculées sur les analyses syntaxiques (souvent des arbres) fournies par le squelette, et ce le plus souvent par des mécanismes reposant sur l'unification. C'est par exemple le cas des Grammaires Fonctionnelles-Lexicales (LFG) ou des Grammaires d'Arbres Adjoints avec décorations (*Feature-based TAG*).

Ce n'est pourtant pas le seul choix possible. En réalité, un formalisme qui étend les CFG et qui est de complexité raisonnable[2] ne peut être clos à la fois par intersection et par homomorphisme. Et choisir, comme cela se fait généralement, la clôture par homomorphisme a une grande influence sur la nature des formalismes. En effet, de nombreux faits linguistiques de différentes

[1]Un homomorphisme étant un cas particulier de substitution, la clôture par homomorphisme découle de la clôture par substitution, plus classiquement utilisée et décrite.

[2]Plus précisément, qui définit un sous-ensemble strict des langages récusivement énumérables.

natures peuvent concerner les mêmes parties d'une phrase. La linéarité induite par la clôture par homomorphisme impose alors de faire un choix : une seule classe de faits linguistiques est décrite par le squelette syntaxique et les autres ne peuvent effectivement qu'être relégués dans des décorations à la complexité mal contrôlée. C'est la raison d'être des architectures à deux niveaux (TAG et traits pour les FTAG, CFG et équations fonctionnelles pour les LFG, etc.), qui ont certains inconvénients[3]. Il est donc intéressant d'explorer l'autre piste, celle des formalismes clos par intersection, ou *non-linéaires*, puisqu'ils permettent de se débarasser des décorations et ont de nombreux avantages computationnels et linguistiques.

Les formalismes non-linéaires à un seul niveau ont la puissance d'expression nécessaire pour décrire le langage naturel, même si l'on fait comme ici l'hypothèse que l'on peut se limiter à des formalismes polynomiaux[4]. En prenant les RCG comme exemple d'un tel formalisme, (Boullier, 2003) présente ainsi des grammaires pour les langages $\{a^{2^n}\}$ et $\{a^n b a^{n-1} b \ldots bab\}$ (nombres chinois, génitifs en géorgien ancien,...) et pour une version abstraite du *scrambling*. De nombreux langages modélisant des phénomènes purement syntaxiques non représentables dans un formalisme syntaxique linéaire peuvent être décrits par des grammaires non-linéaires. Mais cette augmentation de la complexité ne se fait pas au détriment de l'efficacité d'analyse[5].

Par ailleurs, la non-linéarité est appropriée pour décrire les langues car celles-ci reposent elles-mêmes sur des mécanismes non-linéaires : si dans les formalismes à deux niveaux il faut reporter dans les décorations le traitement de certains phénomènes, c'est bien parce que la non-linéarité des langues ne peut être représentée par un formalisme linéaire. On peut citer, comme exemples de phénomènes linguistiques non-linéaires, les mécanismes non-linéaires purement syntaxiques[6], comme les verbes à contrôle ou à montée, et l'interaction entre différents types de contraintes linguistiques, et en particulier la syntaxe et la sémantique lexicale. De fait, un formalisme non-linéaire permet par exemple de contraindre trivialement la construction d'une dépendance à des contraintes syntaxiques (accord, ...) et à des contraintes sémantiques (restrictions de sélection, ...) de manière simultanée, voire à toutes sortes de contraintes[7].

Outre une meilleure représentation des faits linguistiques, la non-linéarité permet une meilleure efficacité d'analyse. En effet, la multiplication des contraintes de diverses natures induit un abandon précoce des analyses invalides : dès qu'une contrainte échoue, l'analyse échoue. Ainsi, plus il y a de contraintes, et plus elle a de chances d'échouer vite. C'est l'antithèse des formalismes à deux niveaux comme LFG, où l'on peut être amené à effectuer un nombre colossal d'analyses de premier niveau (CFG) et à calculer sur les arbres produits de coûteuses décorations (que ce soit fait en une seule phase ou en deux phases).

Un formalisme non-linéaire étendant les CFG couvre au moins *PTIME*. Or nous faisons l'hypothèse que l'on peut se cantonner aux formalismes polynomiaux. Suivant ainsi (Sagot & Boullier, 2004), nous sommes donc intéressés par les formalismes couvrant exactement *PTIME*.

[3]Ces inconvénients sont à la fois computationnels et linguistiques : l'unification induit une complexité algorithmique trop élevée (analyseurs exponentiels), des mécanismes d'interaction (successifs ou simultanés) entre le squelette et les décorations, et un arbitraire linguistique dans la position exacte de la séparation entre squelette et décorations. De plus, il existe des phénomènes purement syntaxiques qui dépassent strictement la puissance d'expression des squelettes linéaires (CFG, TAG, MC-TAG), etc.

[4]Dont les grammaires définissent des langages analysables en temps polynomial en la longueur de l'entrée.

[5]Ainsi, on peut convertir une CFG ou une TAG en une RCG fortement équivalente qui s'analyse respectivement en $O(n^3)$ ou en $O(n^6)$. On peut analyser en $O(n^3)$ toute intersection de CFG. Le langage $\{a^{2^n}\}$, quoiqu'ayant peu de rapport avec la linguistique, montre la puissance de la non-linéarité : il ne respecte pas la propriété de croissance constante (CGP) mais est reconnaissable en temps logarithmique et donc sublinéaire avec une RCG.

[6]Bien qu'aucun formalisme ne les reconnaisse comme tels, puisqu'ils ne sont pas en mesure de les traiter ainsi

[7]Des expériences sont en cours concernant la structuration du discours.

Deux de ces formalismes ont été plus particulièrement étudiés : les *simple Literal Movement Grammars* (sLMG, (Groenink, 1996)) et les Grammaires à Concaténation d'Intervalles (RCG, (Boullier, 2004)). Leur définition formelle et leurs propriétés ont été déjà publiées à maintes reprises, et ne seront pas répétées ici. Rappelons simplement que dans les deux cas une grammaire est un ensemble de clauses à la Horn qui mettent en jeu des prédicats sur des portions de la chaîne d'entrée. Pour plusieurs raisons, les RCG sont plus adaptées aux descriptions linguistiques[8]. Mais ce choix n'est pas véritablement déterminant, et nous l'avons fait en partie aussi en raison de la disponibilité d'un analyseur extrêmement efficace pour les RCG (Boullier, 2004).

3 Les Méta-RCG : une brève description

Si la notion d'intervalle est bien adaptée à la description linguistique, il manque aux RCG un certain nombre de concepts primaires pour pouvoir prétendre au titre de formalisme linguistique. Nous avons donc défini un formalisme qui étend la syntaxe des RCG pour rendre accessibles ces concepts. Ce formalisme, décrit dans (Sagot, 2005), s'appelle Méta-RCG. Il n'introduit pas de décorations formant un second niveau, comme dans les formalismes au squelette syntaxique linéaire : la non-linéarité permet à ces extensions d'être incluses à l'intérieur de la grammaire, par *compilation* de toute Méta-RCG en une RCG classique. Nous avons donc écrit un compilateur qui effectue cette transformation, ainsi qu'un convertisseur permettant de traduire l'analyse d'une phrase obtenue à l'aide de cette RCG en une analyse Méta-RCG.

3.1 Extensions linguistiques de la syntaxe des RCG

Une Méta-RCG, comme une RCG, est constituée de clauses manipulant des prédicats. Et toute RCG est une Méta-RCG. Cependant, les prédicats et les arguments Méta-RCG sont des extensions des prédicats et des arguments RCG. Nous allons donc passer en revue successivement les trois familles d'extensions : les têtes de syntagmes, les traits et numéros d'homonymes, et les contextes.

La notion de tête d'un syntagme est répandue dans la littérature linguistique. Dans le cas des Méta-RCG, l'idée est formulée de la façon suivante : il est impossible, en raison de la non-linéarité des RCG et donc des Méta-RCG, de rendre visible une portion d'analyse à différents endroits de l'analyse, car cela dépasserait *PTIME*. Or on a parfois besoin de rendre visible plus d'informations qu'un simple intervalle (ce qui est la version basique de la non-linéarité). De ce fait, on essaie de construire des représentations partielles de portions d'analyses. On fait alors l'hypothèse suivante : on n'a jamais besoin de connaître à propos d'un syntagme et à l'extérieur de celui-ci plus que ses têtes (il y en a plusieurs en cas de coordination) et un certain nombre de traits (voir ci-dessous). Outre un intervalle simple, un argument Méta-RCG peut donc représenter en plus la liste de ses têtes[9]. On appelle *argument syntagmatique* un tel argument à têtes.

[8] Il est toujours possible de convertir une sLMG en une RCG et inversement. La différence entre RCG et sLMG réside dans ce que l'on entend par « portion » de la chaîne d'entrée. Pour les sLMG, il s'agit de sous-chaînes, alors que pour les RCG il s'agit d'intervalles (c'est-à-dire d'occurrences de sous-chaînes). En conséquence, et à titre d'exemple, il n'y a qu'une seule chaîne vide pour les sLMG, alors que pour les RCG il y a $n+1$ intervalles vides distincts entre les mots d'une phrase de longueur n, ce qui semble plus satisfaisant.

[9] On dispose naturellement d'opérateurs d'empilement d'une tête dans une telle liste.

Les prédicats Méta-RCG étendent les prédicats RCG. Tout d'abord, un prédicat Méta-RCG est décoré par des *traits* définis sur des domaines finis qui permettent de représenter de manière élégante des propriétés de syntagmes, c.-à-d. des propriétés de portions d'analyses et non seulement d'intervalles. Ensuite, on associe à chaque nom de prédicat une liste de positions d'arguments dits *à numéros d'homonymes* : ces arguments, remplis par un intervalle vide ou par un « mot », sont associés à des *numéros* qui différencient les homonymes. Enfin, à chaque nom de prédicat est associée une liste de *contextes* (des intervalles syntagmatiques et/ou des traits) qui peuvent être traités comme des intervalles (ou des traits) standard. Sauf indication contraire, un contexte présent dans deux prédicats d'une même clause a une valeur unique[10].

3.2 Projection partielle d'une analyse globale en analyses classiques

La non-linéarité des Méta-RCG permet de traiter au même niveau les phénomènes de dépendance, de constituance, de sémantique lexicale, et autres. De plus, le fondement même des RCG, à savoir la concaténation d'intervalles, rend la vision topologique de la langue inhérente à toute grammaire Méta-RCG. Aussi l'analyse d'une phrase obtenue à partir d'une grammaire Méta-RCG regroupe-t-elle des faits linguistiques participant de ces diverses classes de phénomènes.

Il est ainsi aisé d'extraire d'une analyse Méta-RCG des *vues partielles* qui constituent des analyses classiques (en constituants, en dépendances, en boîtes topologiques, en relations prédicats-arguments sémantiques, etc.). Cette extraction se fait par *projection partielle* : aucune information n'est calculée à partir de l'analyse Méta-RCG, elles en sont extraites par projection.

4 Une Méta-RCG du français

Le développement du formalisme Méta-RCG s'est fait en parallèle avec celui d'une grammaire du français. Faute de place, nous ne pouvons en donner ici de fragment. Actuellement, elle comporte 370 clauses grammaticales (non lexicales) se compilant en une RCG à 625 clauses, d'arité 73 et de degré maximal 40. Elle couvre déjà un grand nombre de phénomènes linguistiques.

À titre d'illustration, considérons les deux phrases suivantes :

(1) Les entreprises dans lesquelles le Japon veut que la Commission accepte que l'Europe investisse fabriquent des ordinateurs.

(2) Cette idée pose un problème à Nancy.

La phrase (1) inclut une relative qui modifie un verbe enchâssé à travers un contrôle sujet et une complétive. L'efficacité du système permet d'obtenir une analyse unique en 0,38s[11], dont la projection partielle pour obtenir un graphe de dépendances produit la figure 1. La phrase (2) illustre l'interaction entre syntaxe et sémantique au niveau de la grammaire. En effet, *Nancy* étant ambigu (prénom ou lieu), et *pose* (verbe transitif direct ou verbe support à complément d'objet indirect), seules les contraintes sémantiques permettent de n'obtenir que deux analyses, ce qui est bien le cas ici (en 0.01s)[12]. Nous montrons dans la figure 2 l'arbre de constituance (commun aux deux analyses) obtenu par projection partielle de notre analyse.

[10] La notion de *barrière* est alors facilement implémentée : en n'associant pas un certain contexte à un certain prédicat, on l'empêche d'être visible dans la portion d'analyse dont ce prédicat est la racine. D'où un traitement élégant de divers phénomènes, tels que dépendances à longue distance, contrôle, montée, etc.

[11] L'architecture utilisée est un Apple Powerbook avec processeur G4 à 1,5 GHz, et `gcc` 3.3.

[12] L'ambiguïté sur *pose* est levée, mais pas celle sur *Nancy*. Naturellement, si l'on remplace *un problème* par *un vase*, on n'obtient qu'une seule analyse, où *Nancy* est un lieu.

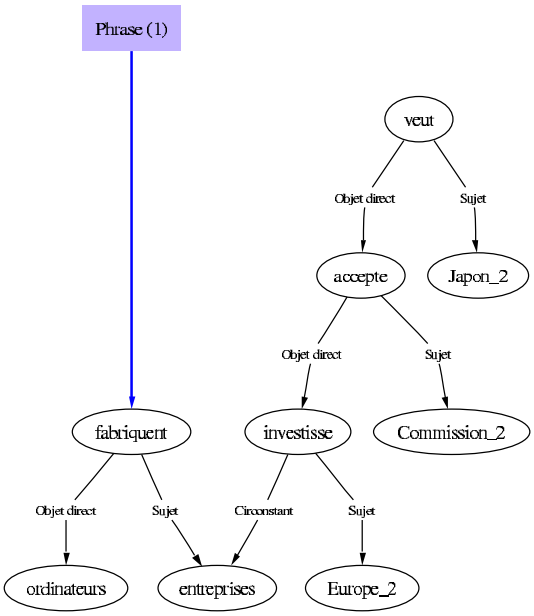


Figure 1. Graphe de dépendance produit par projection partielle de l'analyse de (1).

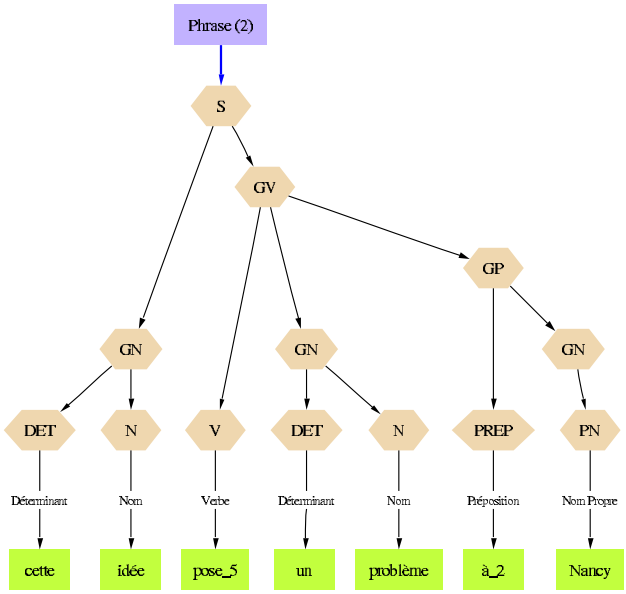


Figure 2. Arbre des constituants produit par projection partielle de l'analyse de (2).

5 Conclusion

Dans cet article, nous avons montré l'importance du concept de non-linéarité pour la description formelle des langues. Nous avons introduit en particulier un formalisme, les Méta-RCG, fondé sur les Grammaires à Concaténation d'Intervalles (RCG). Nous avons ensuite décrit très rapidement la grammaire du français que nous développons, montrant ainsi la pertinence de la prise en compte simultanée de faits linguistiques de natures différentes, et le fait que les analyses classiques peuvent être obtenues à partir de nos analyses par simple projection.

Nous envisageons désormais d'augmenter la couverture de la grammaire, de développer des méthodes automatiques pour la construction du lexique très riche sous-jacent à un tel type de grammaires, et d'étendre la portée de notre grammaire en y incluant des faits linguistiques généralement négligés dans les grammaires, tels que la structuration du discours.

Références

BOULLIER P. (2003). Counting with range concatenation grammars. *Theoretical Computer Science*, **293**, 391–416.

BOULLIER P. (2004). Range concatenation grammars. In *New developments in parsing technology*, p. 269–289. Kluwer Academic Publishers.

GROENINK A. (1996). Mild context-sensitivity and tuple-based generalizations of context-free grammar. In D. JOHNSON & L. MOSS, Eds., *Actes de MoL 4* : Linguistics and Philosophy.

SAGOT B. (2005). Linguistic facts as predicates over ranges of the sentence. In *Proceedings of LACL 05*, Bordeaux, France. To appear.

SAGOT B. & BOULLIER P. (2004). Les RCG comme formalisme grammatical pour la linguistique. In *Actes de TALN 04*, p. 403–412, Fès, Maroc.

Pauses and punctuation marks in Brazilian Portuguese read speech

Izabel C. Seara, Fernando S. Pacheco, Rui Seara Jr., Sandra Kafka, Rui Seara, Simone Klein

LINSE – Circuits and Signal Processing Laboratory
Department of Electrical Engineering
Federal University of Santa Catarina
88040-900 – Florianópolis – SC – Brazil
E-mails: {izabels, fernando, ruijr, kafka, seara, klein}@linse.ufsc.br

Mots-clés : pauses, ponctuations, lecture à haute voix.

Keywords: pauses, punctuation marks, read speech.

Résumé Dans cet article, nous avons examiné la relation entre pause et ponctuation (virgule, point et virgule, deux-points). Toutes ces pauses sont internes aux phrases. À l'aide de l'analyse de plusieurs milliers de pauses dans un corpus de presque 17 heures d'enregistrement réalisé par une locutrice professionnelle native du portugais brésilien, nous avons vérifié une proportion importante des pauses hors ponctuations (61,3%). Les données renforcent aussi la présence des structures topique/commentaire dans la lecture à haute voix. Les résultats des durées de pause correspondantes aux ponctuations sont consistants avec les données présentées dans les grammaires.

Abstract In this paper we assess pause effects corresponding to comma, semicolon, colon and the ones that are not related to any punctuation marks, all of them within sentences. Thus, through the analysis of a corpus of approximately 17 hours of recording, carried out by a female professional speaker (native) of the Brazilian Portuguese language, we observe a large proportion of pauses without punctuation (61.3%). Besides, our data reinforce the presence of topic-comment structures in reading. The results here presented with respect to pause and punctuation are consistent with several studies about this theme.

1 Introduction

Several research work about pause (Vannier *et al.*, 1999; Marin *et al.*, 2002; Campione & Véronis, 2002) have shown the existence of an inaccurate concept about the relationship between pause and punctuation. This concept states that there exists a one-to-one relationship between pause and punctuation; that is, if there is a pause there exists punctuation, no pause, no punctuation. Other works also have shown that several structures, which seem inadequate in reading, reflect usual situations of the spoken language, such as the topic-comment structures (Cagliari, 1992; Pontes, 1987).

The aim of this study is to evaluate situations in which a Brazilian Portuguese speaker inserts pauses in read speech. For such we analyze a Brazilian Portuguese speech corpus and also

investigate the relationship between pause and punctuation marks. The used material consists of recorded read texts by a female professional speaker. This paper is organized as follows. Section 2 presents our speech material and the analysis methodology. Section 3 shows how grammar books have considered a relationship between pause and punctuation. Experimental analyses are carried out in Section 4, investigating the occurrence of pauses, its duration, and the association pause-punctuation. Finally, conclusions and remarks are presented in Section 5.

2 Corpus and Methodology

In the open literature we have found some research works about pause dynamics in natural language which include: (a) a large number of small speech corpora (considering different speakers), according to Campione & Véronis (2002); or (b) a large corpus of a single speaker, according to Marin *et al.* (2002); Vannier *et al.* (1999). By considering that our goal is to obtain a pause model for a Brazilian Portuguese speech synthesis system, we believe that the investigation allowing for a single speaker not only could give rise to the required model, but also to determine some important characteristics (concerning pauses) for the language in question. In this way, we have based our research on a speech corpus with approximately 17 hours of recording, accomplished by a female professional speaker (native) of the Brazilian Portuguese language. This speaker read aloud texts at a normal speaking rate. The data were recorded in a studio and care was taken to avoid environmental noise. The recorded corpus includes different kinds of text, taken from newspapers and magazines, reports, extract of contemporaneous novels, short tales, and academic texts all transcribed and labeled in terms of its morphosyntactic classification. For this task we have used an ad hoc parser conceived for a speech synthesis system (Seara *et al.*, 2002). After that, an expert linguist corrected manually the labeling.

Our analysis is restricted to pauses associated with silent intervals. We have not considered filled pauses (related to hesitations) and pauses created by the lengthening of phonetic segments. To investigate the occurrence, position, and duration of pauses we have considered two main classes: pauses with a silent interval larger than 300 ms, termed long pauses, and pauses with a silent interval between 90 and 299 ms, named short pauses. Seara (2000) has shown that silent intervals associated with the stop closure interval have an average duration of 45 ms for the Brazilian Portuguese language. In this way, if we consider a lower limit equal to 90 ms, it is possible to assure that there will always be a silent interval associated with a pause, even if the stop closure occurs within this interval.

In order to study the duration and types of acoustic pauses, the speech data have been processed automatically by a pause detector. This detector retrieves each context with pauses from the corpus. In this way, we obtain a total of 9,985 pauses from the 17 hours of recording of which 3,858 are associated with punctuation marks. From those, 3,633 are accompanying commas; 52, semicolon; and 173, colon. The other 6,127 pauses are not related to punctuation marks.

3 Pause and Punctuation Relationship in Grammar Books

Usually, Brazilian Portuguese teachers complain about mistakes associated with punctuation made by their students. Such mistakes refer to the idea that it is imperative to observe the pauses to assign a comma within a sentence. Thus, we perceive how difficult it is to teach the students that the punctuation marks are based on syntactic structures of the sentences and not on the carried out pauses.

If we observe Brazilian Portuguese grammar books, almost all of them mention pauses when they discuss punctuation marks (Almeida, 1988; Faraco & Moura, 1991, among others). Grammarians state that punctuation marks, such as comma, period, and semicolon are used for pause marking. They also explain that the comma, period, and semicolon represent, respectively, pauses of short, long and intermediate duration. However, they emphasize that the comma should not be used either between the subject and verb or between the verb and object. Thus, we can assume that pauses would not occur in reading aloud in these situations, since the punctuation marks, which indicate such pauses, would not be included. Nevertheless, our speaker inserts pauses in these structures. The following example presents pauses between verb-object and subject-verb, respectively:

> *Os pesquisadores **afirmam** [short pause] **que os resultados** são a primeira evidência de que os **transgênicos** [long pause] **podem** gerar conseqüências*
>
> (*The researchers affirm [short pause] that the results are the first evidence that the transgenic organisms [long pause] may produce consequences.*)

Almeida (1988) states that a comma is never found where no pause is placed. Despite our analysis data we have found 1% of commas with no correspondence to pauses. Thus, it would be more accurate to state that frequently if there is no pause there is no punctuation. Remark that Vannier *et al.* (1999) also have found 4.6% of pauses without punctuation in a French corpus.

To obtain a notion of the insertion dynamics of long and short pauses, mainly for the cases in which there is no association with punctuation marks, we achieve an analysis of the syntactic sequences previously mentioned.

4 Data Analysis and Discussion

To evaluate pauses between both subject-predicate and verb-object (structures in which the comma is forbidden), we divide the data that present such sequences into two groups: subject and verb (or predicate) termed Group 1, and verb and object, Group 2. Results are shown in Tables 1 and 2.

Table 1: Number and occurrence frequency of the Group 1 sequences with and without pauses

Group 1	Without pauses		Long pauses		Short pauses	
	Number of pauses	%	Number of pauses	%	Number of pauses	%
Subject-verb	1658	51.88	459	14.40	1079	33.80
			459/1538	29.84	1079/1538	70.16
			1538 (48.12%)			

According to Tables 1 and 2, we can verify that there exists a larger tendency of occurring pauses between subject and predicate than between verb and object. However, in both cases the prohibition of the grammar books of inserting punctuation in these sequences has not inhibited the presence of long pauses, albeit less frequent than short pauses. Besides the pauses observed in Groups 1 and 2, we also verify that the pauses that are not associated with punctuation precede syntactic boundaries (29.23% correspond to conjunctions, 17.48% prepositions, and 4.13% adverbs). In these cases 81.22% are short pauses.

Table 2: Number and occurrence frequency of the Group 2 sequences with and without pauses

Group 2	Without pauses		Long pauses		Short pauses		Total
	Number of pauses	%	Number of pauses	%	Number of pauses	%	
Verb-object	1489	98.30	11	0.70	15	1	1515
Total	1489 (98.30%)		25 (1.70%)				100%

Examining the data that present pause between subject and verb, we notice that they seem to be associated with topic-comment structures. However, Cagliari (1992) points out that in written texts topic-comment structures should contain comma. As in the written texts there is no comma, we expect that in reading aloud the topic-comment structure has not occurred. In such a structure the speaker divides the statement into two tonal groups. A pause could (or not) be placed between a tonal group and another. Our speaker reinforces such a structure, since her reading presents the two tonal groups and pauses.

Topic-comment structure is a characteristic inherent to the spoken language. In Brazil the spontaneous speech presents a large quantity and diversity of clauses with topic-comment structures (Pontes, 1987). Nevertheless, as the grammar disapproves such structures, they do not appear widely in written language, as occurs in spontaneous speech.

Our data show the presence of topic-comment structures in the text reading, since we notice that 48.12% of the subject-verb sequences are interrupted by pauses. On the other hand, the large tendency of short pauses (70.16%) between subject and verb shows that in a certain way in reading, the grammatical objection inhibits the presence of long pauses. Thus, based on the results here shown we can verify the presence of topic-comment structures in reading, ratifying the data obtained by other authors (Pontes, 1987; Cagliari, 1992; Mollica, 1993).

In Campione & Véronis (2002) a multilingual study about pauses in read speech and its relationship with punctuation is presented. They show that 11.9% of pauses are not associated with punctuation in the French language. The largest percentage of this kind of occurrence is found in the Italian language, in which 33% of pauses are not associated with punctuation. However, Vanier (1999 *apud* Campione & Véronis, 2002) presents a study about the same theme, in which 36% of pauses are not associated with punctuation in the French language. Our data seems to indicate that Brazilian Portuguese is the language that presents a larger occurrence of pauses not associated with punctuation (61.36%), and in general the results have shown that pauses that are not related to punctuation marks are mostly short pauses (see Fig. 1 and Table 3).

The pauses related to punctuation marks (whose plots are also shown in Fig. 1) include: (a) when concerning commas, there is a slight tendency for them to be short; (b) when referring to semicolons and colons there is a strong tendency to be long (with semicolon presenting an intermediate duration) (see Table 3). However, we have not observed distinct clusters in our data, since overlapping occurs between these three classes. The total average duration of the pauses analyzed within the sentences is 224 ms. Data that present punctuation marks which do not correspond to a pause are few (less than 1%).

Data related to the topic-comment structures in the written text do not present continuity violation. This fact means that the speaker produces entire constituents without interrupting them, as described in Strangert (2004). These pauses are inserted before syntactic boundaries. However, in the reading it is possible to perceive a few cases of boundaries in syntactically unmotivated positions. We have found in our corpus less than 1% of cases involving pauses without syntactic

motivation. This occurrence shows the continuity violation defined by Strangert (2004), as can be verified in the following example:

... *dentre* ***outros*** *[short pause]* ***agentes*** *de doenças.* (*...among other [short pause] disease agents.*)

In this example, the pronoun *outros* and the noun *agentes* form a syntactic constituent (noun phrase), which is interrupted by a short pause, representing a continuity violation.

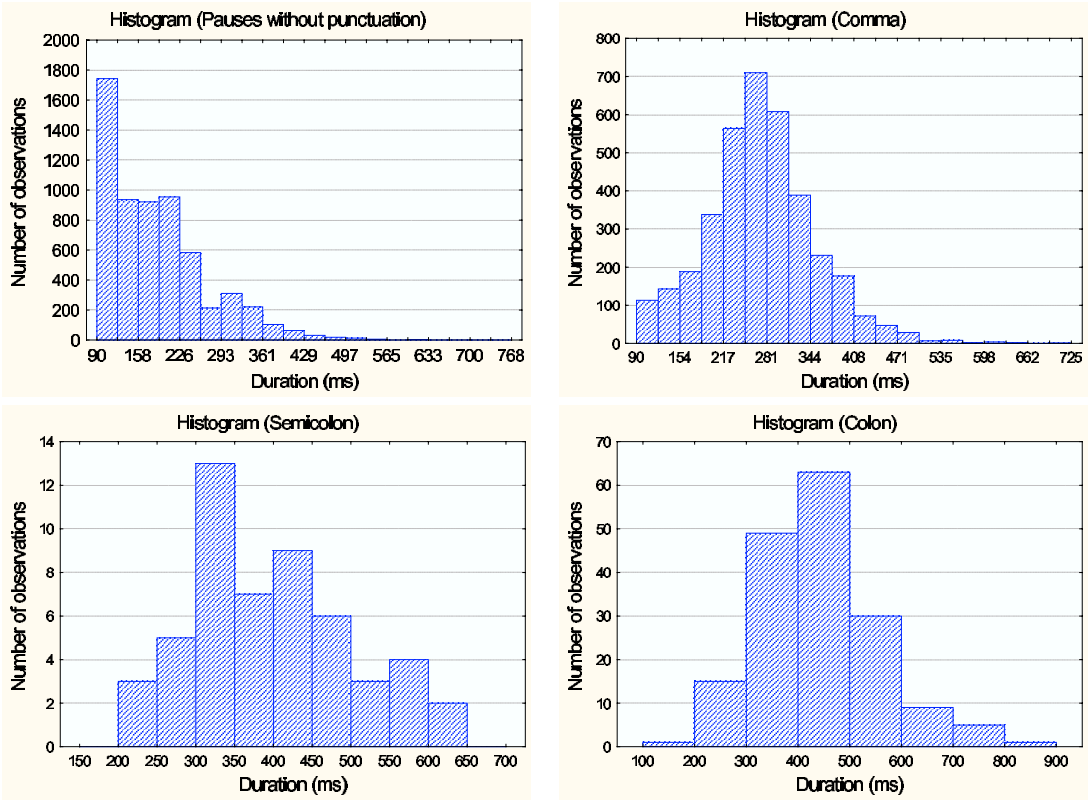


Fig. 1: Duration histogram of pauses: (top left) pauses not associated with punctuation; (top right) associated with comma; (bottom left) with semicolon; (bottom right) with colon.

Table 3: Average duration of pauses for the analyzed data

Pause Classification	Occurrence		Duration (ms)	
	Number	%	Average	Std. Dev.
[long pause]	759	7.60	356	58
[short pause]	5368	53.76	163	54
Comma [long pause]	1194	11.96	357	54
Comma [short pause]	2439	24.43	231	51
Semicolon [long pause]	44	0.44	426	91
Semicolon [short pause]	8	0.08	259	23
Colon [long pause]	157	1.57	478	103
Colon [short pause]	16	0.16	252	32
Total	9985	100		

5 Conclusions and Remarks

In this paper we have discussed the relationship between pauses and punctuation in the Brazilian Portuguese language. We have shown that the claim where there is pause there is also punctuation is not accurate, since we notice a large proportion of pauses without punctuation (61.3%). Also the categorical claim where there is no pause there is no punctuation is not quite accurate, because in 1% of the cases punctuation without pause has occurred. In addition, our data reinforce the presence of topic-comment structures in reading. This fact was not expected due to the objection to that structure. The results here presented with respect to pause and punctuation are consistent with several studies about this theme. However, other analyses using more speakers of Brazilian Portuguese are needed for obtaining a generalization of these findings as inherent to such a language.

References

ALMEIDA N. M. DE (1988), *Methodic Grammar of the Portuguese Language* (in Portuguese), São Paulo, Brazil, Saraiva.

CAGLIARI L. C. (1992), On the importance of the prosody in the description of grammatical events (in Portuguese), In: ILARI, R. (org.) *Grammar of the spoken Portuguese* (in Portuguese), v.2. Campinas (SP), Brazil, Unicamp.

CAMPIONE E., VERONIS J. (2002), Etude des relations entre pauses et pontuations pour la synthèse de la parole à partir de texte. Proceedings of *Traitement Automatique des Langues Natureles* (TALN 2002), Nancy, France, 1-10.

FARACO C. E., MOURA F. M. DE. (1991), *Grammar: phonetic and phonology, morphology, syntax, stylistic* (in Portuguese), São Paulo, Brazil, Ática.

MARIN R., AGUILAR L., CASACUBERTA D. (2002) Placing pauses in read Spanish : a model and an algorithm.*Language Design,* 4, 46-66.

MOLLICA C. (1993). Intervals between the silence and the speech and their representations in written text. *Cadernos de Letras* (in Portuguese), no. 9, Rio de Janeiro, 143-149.

PONTES E. (1987) *The topic in the Brazilian Portuguese* (in Portuguese), São Paulo, Brazil, Pontes.

SEARA I. C. (2000), Analysis acoustic-perceptual of nasality of the vowels in the Brazilian Portuguese. *PhD. Thesis* (in Portuguese), Federal University of Santa Catarina, Florianópolis, Brazil.

SEARA I. C., KAFKA S. G., KLEIN S., SEARA R. (2002), Vowel sound alternation of verbs and nouns of the Portuguese spoken in Brazil for application in text-to-speech synthesis. *Journal of the Brazilian Telecommunication Society* (in Portuguese), vol. 17, no. 1, 79-85.

STRANGERT E. (2004), Speech chunks in conversation: Syntactic and prosodic aspects. Proceedings of *Speech Prosody*, Nara, Japan.

VANNIER G., LACHERET-DUJOUR A., VERGNE J. (1999), Pauses location and duration calculated with syntactic dependencies and textual considerations for T.T.S. system. Proceedings of *XIV International Congress of Phonetics Sciences (ICPhS)*, San Francisco, USA.

Segmentation thématique par chaînes lexicales pondérées

Laurianne Sitbon, Patrice Bellot
Laboratoire d'Informatique d'Avignon - Université d'Avignon
339, chemin des Meinajaries - Agroparc BP 1228
84911 AVIGNON Cedex 9 - FRANCE
Tél : +33 (0) 4 90 84 35 09
{laurianne.sitbon, patrice.bellot}@univ-avignon.fr

Mots-clefs : segmentation thématique, chaînes lexicales, entités nommées

Keywords: topic segmentation, lexical chains, named entities

Résumé Cet article propose une méthode innovante et efficace pour segmenter un texte en parties thématiquement cohérentes, en utilisant des chaînes lexicales pondérées. Les chaînes lexicales sont construites en fonction de hiatus variables, ou bien sans hiatus, ou encore pondérées en fonction de la densité des occurrences du terme dans la chaîne. D'autre part, nous avons constaté que la prise en compte du repérage d'entités nommées dans la chaîne de traitement, du moins sans résolution des anaphores, n'améliore pas significativement les performances. Enfin, la qualité de la segmentation proposée est stable sur différentes thématiques, ce qui montre une indépendance par rapport au type de document.

Abstract This paper presents an innovative and efficient topic segmentation method based on weighted lexical chains. This method comes from a study of different existing tools, and experiments where we considered the influence of a term at each precise place in the text. We build lexical chains with different kinds of hiatus (varying, none or density weighted). We demonstrate good results on a manually built french news corpus. We show that using named entities does not improve results. Finally, we show that our method tends to be domain-independent because results are similar on various topics.

1 Introduction

La segmentation thématique intervient dans différents domaines en organisation de l'information, tels que déterminer les limites entre des dépêches dans un flux d'informations (*broadcast news*, TDT (*Topic Detection and Tracking*)), ou créer un résumé automatique de textes pour lequel la segmentation sert à isoler les thématiques et les parties les plus représentatives (McDonald &

Chen, 2002). Enfin, du point de vue d'un utilisateur, la segmentation thématique a également des avantages en terme de facilités de lecture.

Beaucoup de moyens ont été imaginés pour segmenter un texte en thèmes cohérents. La principale différence entre ces méthodes tient au fait qu'elles sont ou non supervisées. Parmi les méthodes supervisées on trouve par exemple PLSA (Brants *et al.*, 2002) qui apprend des probabilités d'appartenance des termes à des classes sémantiques. D'autres méthodes se basent sur un apprentissage comme (Amini *et al.*, 2000) qui s'appuient sur des modèles de Markov cachés, ou bien (Caillet *et al.*, 2004) qui propose une classification des termes de même que (Chuang & Chien, 2004) et (Mekhaldi *et al.*, 2004). Nous avons développé une méthode non supervisée, à base de matrice de similarités et de chaînes lexicales du type de celles utilisées par (Utiyama & Isahara, 2001) ou (Galley *et al.*, 2003). Ce choix est lié à leurs capacités naturelles d'adaptation à de nouvelles thématiques, et à leur relative indépendance vis à vis de la langue des textes.

2 Méthodologie

Avant de concevoir une nouvelle méthode, nous avons fait une évaluation de l'état de l'art pour des textes en français (Sitbon & Bellot, 2004). Cette étude préliminaire a notamment permis d'analyser les différentes mesures d'évaluation de la segmentation, et de créer un corpus de test en français. Les outils que nous avons étudiés ont été comparées pour l'anglais par (Choi, 2000). Nos expériences ont montré que dans les conditions où le texte à segmenter est une suite d'articles bien distincts, et où la qualité des outils est évaluée automatiquement en fonction de la distance entre les frontières trouvées et celles à trouver, l'outil le plus efficace est C99 (Choi, 2000), qui ordonne localement les similarités entre chaque paire de phrases, puis fait des regroupements par maximum de densité. Nous avons montré que le type de document que l'on segmente, son thème, la taille et la variation de taille des segments à repérer, sont autant de caractéristiques influençant le travail des segmenteurs.

2.1 Construction et pondération des chaînes lexicales

Une chaîne lexicale relie des termes de manière linéaire dans un texte. Les méthodes actuelles de segmentation les utilisent pour relier les occurrences d'un même terme (ou lemme) qui sont "proches". Une chaîne est rompue lorsque le nombre de termes qui séparent deux occurrences dépasse une valeur fixée appelée hiatus. On peut alors recenser pour chaque phrase les chaînes actives.

Les applications des chaînes lexicales utilisent actuellement des hiatus définis de manière empirique, et la notion d'activité d'une chaîne est binaire (elle est active ou non active). Notre premier objectif est d'éliminer le caractère empirique du hiatus, afin que notre outil puisse s'adapter à n'importe quel type de texte sans intervention de l'utilisateur. Pour cela on peut

imaginer tout simplement la **suppression du hiatus**, ce qui revient à relier toutes les répétitions de termes. Nous avons également envisagé l'utilisation de **hiatus locaux** : le hiatus moyen est calculé pour chaque lemme. Ainsi si un mot est fortement répété à deux endroits distincts du texte, il y aura automatiquement la création de deux chaines. De plus, s'il est répété trois fois en début de texte, puis une fois à la fin, il n'y aura qu'une seule chaîne comprenant les occurrences de début de texte.

Ces techniques créent un déséquilibre dans la significativité de l'activité des chaînes en jeu dans le calcul des frontières. Il faut alors pondérer les chaînes, en fonction de leur compacité (ratio entre leur taille et le nombre d'occurrences). (Galley *et al.*, 2003) propose une pondération des chaînes en fonction de la compacité et de la fréquence du terme considéré, et obtient de bons résultats, malgré un hiatus déterminé de manière empirique. La catégorie des lemmes a été intégrée à cette pondération. Le poids d'une chaîne associée à un terme m est défini par :

$$score(Chaine, m) = max(Chaine, cat(m) \times freq(Chaine, m) \times log(\frac{L_{texte}}{L_{chaine}}) \quad (1)$$

où freq(Chaine, m) est le nombre d'occurrences du terme m dans la chaîne, L_{texte} la longueur du texte, L_{chaine} la longueur de la chaîne (on est alors indépendant de la taille des textes à segmenter), et max(Chaine,cat(m)) le poids de la forme grammaticale la plus importante parmi les occurrences du terme dans la chaîne.

Puis on calcule les similarités à chaque fin de phrase, qui est une rupture thématique potentielle. La similarité est calculée avec :

$$sim(A, B) = \frac{\sum_m score(A, m) \times score(B, m)}{\sqrt{\sum_m score(A, m) \times \sum_m score(B, m)}} \quad (2)$$

où A et B sont les ensembles de vecteurs représentant les poids des chaînes lexicales actives dans les n phrases avant et après (nous avons choisi n=2), score(X, m) étant le poids maximal du terme m dans l'ensemble des vecteurs X.

Les frontières retenues sont alors celles pour lesquelles la similarité est en dessous d'un seuil déterminé par $sim_{limit} = \mu + \frac{\sigma}{2}$ où μ et σ sont la moyenne et la variance de toutes les similarités calculées (Galley *et al.*, 2003).

2.2 Evaluation

Afin de constituer un corpus de grande taille aisément, on compose un corpus de test à partir d'articles journalistiques de thème globalement éloignés car classés manuellement dans différentes rubriques. Le corpus de test est composé de 4 séries de 100 documents, chaque série

correspondant à une taille moyenne pré-définie des segments. Chaque document est composé de 10 segments qui sont autant d'extraits d'articles du journal Le Monde, choisis aléatoirement. Ce journal proposant des thématiques très variées, on suppose alors les segments thématiquement cohérents et différents.

Pour évaluer l'efficacité de nos nouveaux algorithmes, nous avons utilisé la mesure Window Diff proposée par (Pevzner & Hearst, 2002), présentée et analysée dans (Sitbon & Bellot, 2004).

Nous avons testé les différentes approches pour le calcul des chaînes lexicales. L'utilisation d'un hiatus fixe de 11 est le paramètre préconisé par (Galley *et al.*, 2003) avec l'outil *LCseg*. Les résultats montrés dans cet article et rappelés ici dans le tableau 1 affichent de meilleures performances pour cette approche que C99 sur le Brown corpus, ainsi que sur le corpus TDT.

	Brown Corpus	TDT Corpus
LCseg	0,1137	0,0909
C99	0,1457	0,1272

Table 1: comparaison de LCseg et C99 pour un nombre de segments inconnu, selon la mesure WindowDiff

Taille	LCseg	hiatus 120	hiatus locaux
9-11	0,3272	**0,3187**	0,3454
3-11	0,3837	**0,3685**	0,4016
3-5	0,4344	0,4309	**0,4204**

Table 2: Comparaison de LCseg et de notre méthode pour des segments de différentes tailles (en nombre de phrases)

Nous avons donc décidé de comparer notre approche à celle de (Galley *et al.*, 2003). Les résultats sont présentés dans le tableau 2. Etant donné que les textes ont tous moins de 110 phrases (le maximum étant 10 segments de 11 phrases chacun), le hiatus 120 correspond à une absence de hiatus. Pour ces tests, les lemmes n'ont pas été pondérés en fonction de leur catégorie.

Pour les segments de grande taille (9-11), ou de tailles variables (3-11) , la meilleure méthode est finalement celle qui ne coupe pas les chaînes lexicales (hiatus 120). Pour des segments de petite taille, on observe une très faible amélioration lorsqu'on utilise des hiatus différents pour chaque lemme (hiatus locaux).

3 Exploitation d'une détection d'entités nommées

Si nous avons pris jusqu'ici les termes et leur fonction syntaxique comme seuls critères, nous pensons par ailleurs que la variation des noms propres est un indice intéressant.

On appelle entité nommée dans un texte tout ce qui fait référence à un identifiant unique (Chinchor, 1997). Il peut s'agit un mot ou d'un groupe nominal. Nous avons utilisé trois types d'entités nommées, repérées à partir d'un lexique fermé : listes de noms de personne, noms de lieux et noms d'organisations. Etant un identifiant unique, une entité nommée a tendance à moins se répéter d'un thème à l'autre. De plus, il est moins malvenu en français de les répéter que les noms communs ou adjectifs, même si le problème des anaphores reste à résoudre comme

on le verra.

Nous avons conduit plusieurs types d'expériences (table 3), en fonction de poids plus ou moins élevés attribués aux entités, ou en n'utilisant que les entités. Dans un premier temps nous avons multiplié le poids des chaînes contenant des entités nommées, par deux (ENx2) puis par dix (ENx10). Ensuite nous avons testé la méthode en n'utilisant que les chaînes lexicales correspondant à des entités nommées (EN seules).

Segments 9-11		
Méthode	hiatus 120	hiatus locaux
classique	**0,3187**	0,3454
ENx2	0,3211	0,3536
ENx10	0,3521	0,3888
EN seules	0,4235	0,4975

Segments 3-5		
Méthode	hiatus 120	hiatus locaux
classique	0,4309	0,4204
ENx2	0,4291	**0,4128**
ENx10	0,4315	0,4202
EN seules	0,4228	0,4291

Table 3: Evaluation sur un corpus journalistique avec différentes pondérations des entités nommées pour 2 tailles moyennes des segments

Les résultats présentés dans la table 3 montrent que l'amélioration avec une utilisation des entités nommées est très peu significative d'une part, et qu'il faut doser cet usage d'autre part. En effet on observe une perte de qualité lorsqu'on leur accorde un poids trop important ou lorsque on ne considère qu'elles.

Nous avons ensuite refait les mêmes tests sur un corpus journalistique composé uniquement d'articles traitant de sport, et sur lequel les méthodes testées dans (Sitbon & Bellot, 2004) donnaient les résultats les plus médiocres.

Segments 9-11		
Méthode	hiatus 120	hiatus locaux
classique	**0,3202**	0,3463
ENx2	0,3255	0,3321
ENx10	0,3561	0,3695
EN seules	0,3976	0,4621

Segments 3-5		
Méthode	hiatus 120	hiatus locaux
classique	0,4375	**0,4179**
ENx2	0,4359	0,4183
ENx10	0,4393	0,4265
EN seules	0,4430	0,4634

Table 4: Evaluation sur un corpus journalistique avec différentes pondérations des entités nommées pour 2 tailles moyennes des segments

Les résultats présentés sur la table 4 montrent que l'on n'obtient pas l'amélioration attendue par l'utilisation des entités nommées. Cela peut s'expliquer par une trop fréquente utilisation des anaphores qui limite la répétition des entités. De plus la reconnaissance à l'aide de listes limite le nombre d'entités utilisées, et il faudra recommencer cette étude avec un outil de détection automatique des entités nommées, afin de pouvoir en utiliser un plus grand nombre. On peut également utiliser des cooccurrences d'entités pour créer des chaînes "multi-lexicales".

On constate que les résultats pour un corpus thématiquement cohérent (sport), sont du même ordre que ceux pour un corpus généraliste. Cela tend à montrer que ces méthodes sont indépendantes du type de lexique utilisé dans les documents segmentés, ce qui apporte une forme d'indépendance dans le type de document, et qui était un de nos objectifs initiaux.

4 Conclusion

Les techniques que nous avons imaginées pour s'affranchir du paramètre du hiatus dans l'emploi des chaînes lexicales pour la segmentation thématique sont efficaces. Nous pensons pouvoir encore améliorer la qualité de la segmentation en calculant les probabilités de rupture thématique à partir des similarités, en utilisant des ordonnancement locaux à chaque frontière candidate, comme cela est fait dans C99. Le développement sous forme d'API est en cours, l'outil sera distribué prochainement dans le cadre du projet technolangue AGILE/OURAL (`http://www.technolangue.net/article79.html`).

Références

AMINI M., ZARAGOZA H. & GALLINARI P. (2000). Learning for sequence extraction tasks. In *Proceedings RIAO'2000*, Paris, France.

BRANTS T., CHEN F. & TSOCHANTARIDIS I. (2002). Topic-based document segmentation with probabilistic latent semantic anaysis. In *Proceedings of CIKM'02*, McLean, Virginia, USA.

CAILLET M., PESSIOT J.-F., AMINI M. & GALLINARI P. (2004). Unsupervised learning with term clustering for thematic segmentation of texts. In *Proceedings RIAO'04*, Avignon, France.

CHINCHOR N. (1997). Muc-7 named entity task definition. in http://www.itl.nist.gov/iaui/894.02/related projects/muc/proceedings/ne task.html.

CHOI F. Y. Y. (2000). Advances in domain independent linear text segmentation. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, USA.

CHUANG S.-L. & CHIEN L.-F. (2004). A practical web-based approach to generating topic hierarchy for text segments. In *Proceedings of the Thirteenth ACM conference on Information and knowledge management table of contents*, p. 127–136, Washington, D.C, USA.

GALLEY M., MCKEOWN K., FOLSER-LUSSIER E. & JING H. (2003). Discourse segmentation of multi-party conversation. In *Proceedings of ACL'03*, Sapporo, Japan.

MCDONALD D. & CHEN H. (2002). Using sentence selection heuristics to rank text segments in txtractor. In *Proceedings of the 2nd ACM/IEEE Joint Conference on Digital Libraries*, p. 28–35.

MEKHALDI D., LALANNE D. & INGOLD R. (2004). Using bi-modal alignment and clustering techniques for documents and speech thematic segmentations. In *Proceedings of the Thirteenth ACM conference on Information and knowledge management table of contents*, p. 69–77, Washington, D.C, USA.

PEVZNER L. & HEARST M. A. (2002). A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, p. 19–36.

SITBON L. & BELLOT P. (2004). Adapting and comparig linear segmentation methods for french. In *Proceedings RIAO'04*, Avignon, France.

UTIYAMA M. & ISAHARA H. (2001). A statistical model for domain-independent text segmentation. In *in Meeting of the Association for Computational Linguistics*, p. 491–498.

Une plateforme pour l'acquisition, la maintenance et la validation de ressources lexicales

VanRullen T. , Blache P. , Portes C. , Rauzy S. , Maeyhieux J.-F. , Guénot M.-L. ,Balfourier J.-M. , Bellengier E.
Laboratoire Parole et Langage - CNRS - Université de Provence
29, Avenue Robert Schuman - 13100 Aix-en-Provence
{tristan,pb}@lpl.univ-aix.fr

Mots-clefs : dictionnaire, lexique, lexique noyau

Keywords: dictionary, lexicon, kernel lexicon

Résumé Nous présentons une plateforme de développement de lexique offrant une base lexicale accompagnée d'un certain nombre d'outils de maintenance et d'utilisation. Cette base, qui comporte aujourd'hui 440.000 formes du Français contemporain, est destinée à être diffusée et remise à jour régulièrement. Nous exposons d'abord les outils et les techniques employées pour sa constitution et son enrichissement, notamment la technique de calcul des fréquences lexicales par catégorie morphosyntaxique. Nous décrivons ensuite différentes approches pour constituer un sous-lexique de taille réduite, dont la particularité est de couvrir plus de 90% de l'usage. Un tel *lexique noyau* offre en outre la possibilité d'être réellement complété manuellement avec des informations sémantiques, de valence, pragmatiques etc.

Abstract We present a lexical development platform which comprises a lexical database of 440.000 lemmatized words of contemporary French, plus a set of maintenance tools. The lexical database is intended to be distributed and updated regularly. We present in this paper tools and techniques applied for the lexicon constitution and its enrichment, in particular the computation of lexical frequencies by morphosyntactic category. Then we describe various approaches to build an under-lexicon of reduced size, whose characteristic is to cover more than 90% of the use. Such a *kernel lexicon* makes it moreover possible to be really enriched by hand with semantic, valence, pragmatic information, etc.

1 Introduction

L'élaboration d'un lexique électronique peut sembler une tâche obsolète, de nombreux lexiques du français étant référencés. Cependant, force est de constater que cette affirmation doit être modulée. La première constatation est que seul un petit nombre d'entre eux est effectivement accessible. Il faut de ce point de vue souligner le rôle considérable joué par Bdlex (cf. [de Calmes98]) qui, dans le cadre des activités du GdR-PRC Communication Homme-Machine, a longtemps été le lexique le plus largement diffusé en contribuant ainsi puissamment à l'évolution du domaine en France. Le mode de diffusion constitue évidemment un aspect critique [1]. Un rapide survol des ressources lexicales libres d'accès pour le français permet d'en identifier deux :

- *Lexique* : il s'agit d'un lexique comportant 130.000 formes et comportant des informations morphosyntaxiques, phonologiques et des indications de fréquence (cf. [New01], `http://www.lexique.org/`).
- *ABU* : contient 300.000 formes avec indications morphosyntaxiques (cf. [ABU], `http://abu.cnam.fr/`).

On peut par ailleurs trouver quelques ressources verbales, par exemple :

- *Lefff* : il contient 200.000 formes verbales, avec les informations de base (temps, nombre, personne) (cf. [Clément04], `http://www.lefff.net/`);
- *Litote* : c'est une base contenant les formes conjuguées de 6.500 verbes. (`http://www.loria.fr/equipes/calligramme/litote/`)

Par ailleurs, il faut également signaler la démarche initiée par le Loria dans le cadre du projet *Morphalou* (cf. [Romary04], `http://loreley.loria.fr/morphalou/`). Ce projet fournira également à terme un lexique morphologique de 540.000 formes. Son intérêt tient d'une part au fait qu'il est collaboratif mais également qu'il s'inscrit dans le cadre du projet LMF (*Lexical Markup Framework*), proposant la normalisation du codage des informations linguistiques.

Il reste donc un travail important pour parvenir à un lexique de qualité. Pour cela, une base lexicale doit avant tout être nettoyée de façon à proposer une couverture adéquate du français. Il ne sert à rien de constituer une ressource de 400 ou 500.000 formes si la plupart d'entre elles ne sont pas attestées. Le second aspect concerne le type d'informations contenu dans le lexique. Il est en effet nécessaire qu'un lexique contienne pour une même entrée autant d'informations que possible concernant ses propriétés morphologiques, syntaxiques, bien entendu, mais également sémantiques, phonétiques ou phonologiques. La forme phonétisée de l'entrée, la syllabation ou la fréquence sont par exemple autant d'informations précieuses pour la description.

Nous décrivons dans cet article la base lexicale développée au LPL. Cette base, construite autour d'un lexique morphologique, présente la particularité d'être couvrante, de contenir des information variées et d'avoir été validée sur corpus. Cette base est associée à une véritable *plateforme* de développement lexical, munie de divers outils de maintenance et d'accès. Après une présentation des principales caractéristiques de cette plateforme, nous en proposons une évaluation se fondant sur différents corpus. Nous décrivons de plus l'exploitation de cette base dans la perspective d'une étude lexicale du français contemporain.

[1] Nous nous associons de ce point de vue à la démarche aujourd'hui proposée par le projet Morphalou et nos ressources seront distribuées dans ce cadre

2 Le lexique complet

Le lexique que nous avons mis au point a fait l'objet de beaucoup d'études et de travaux d'amélioration. Nous aboutissons actuellement à un lexique défactorisé de plus de 444.000 entrées correspondant à environ 320.000 formes orthographiques différentes. Ce lexique est associé à un ensemble d'outils permettant sa maintenance, sa sécurisation et son interrogation. Ce projet est la base nécessaire à des applications du TALN qui auront besoin d'une ressource fiable, c'est pourquoi l'accent a été mis sur la maintenabilité de la ressource.

Les entrées du lexique *DicoLPL* sont basées sur des ressources libres et une acquisition semi-automatique. Comme le montre la figure 1, nous avons au départ recensé et incorporé des lexiques libres, tels *ABU* ou *Lexique.org*. Le formatage de notre lexique a nécessité un travail de transformation, de catégorisation, de phonétisation etc., afin de faire correspondre les entrées acquises. L'étape importante que constitue le calcul des fréquences lexicales est abordé dans la prochaine section.

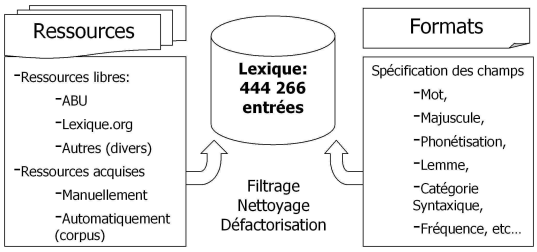


Figure 1: Conception du dictionnaire

Forme	Phonetisation	Frequence	Categorie	Lemme
de_traviole	d@_tRavjOl@	9	Rgp	de_traviole
dealer	dil@R	188	Ncms–	dealer
dealers	dil@R	368	Ncmp–	dealer
déambula	dea~byla	5	Vmis3s–	déambuler
déambulai	dea~bylE	0	Vmis1s–	déambuler
déambulaient	dea~bylE	23	Vmii3p–	déambuler
déambulais	dea~bylE	1	Vmii1s–	déambuler
déambulais	dea~bylE	0	Vmii2s–	déambuler
déambulait	dea~bylE	17	Vmii3s–	déambuler

Figure 2: Extrait du lexique

Le format du lexique, son codage, et son stockage ont été pensés afin d'accélérer son chargement dans les applications qui le requièrent. Ce lexique est en effet actuellement embarqué dans des applications de communication sur des machines ayant de petites capacités. D'autre part, il s'agit de permettre avec le même stockage un développement et des modifications manuels. C'est pourquoi un format ASCII, structuré en CSV tabulé classique a été choisi, plutôt qu'un standard XML ou qu'une forme binaire de type *base de données*. Ce choix a répondu à nos attentes et permet une transformation rapide dans d'autres formats tels que le XML répondant aux normes ISO (TC37/SC4) utilisées par le projet MORPHALOU par exemple.

Notre lexique se structure sous une forme défactorisée (une ligne par quadruplet *[Mot, Phonetisation, Categorie, Lemme]* par opposition à d'autres lexiques pour lesquels une seule ligne est réservée pour chaque forme orthographique.

L'extrait de lexique donné dans la table 2 met en évidence les caractéristiques de son format. On y observe la défactorisation du mot *déambulais*.

Certaines colonnes ont été réservées pour un usage ultérieur; les mots acceptés dans ce lexique ne doivent pas être des affixes, mais toujours des mots (simples ou composés) du langage courant. Ainsi, les préfixes et suffixes tels que *anti, hecto, isme* ou *able* en sont rejetés.

Le codage des champs du lexique est lui aussi contraint: les fréquences correspondent au nombre d'occurences de chaque entrée mesurée sur les corpus d'apprentissage. Les valeurs sont des entiers et ne représentent pas des pourcentages. Les valeurs de traits des catégories de chaque entrée sont formatées selon un codage dérivé de Multext et de Grace. La forme phonétisée est exprimée à l'aide de l'alphabet standard Sampa, qui permet un codage phonétique en texte brut sans faire appel à des polices de caractères spécifiques.

3 Plateforme d'enrichissement du lexique

Le lexique *DicoLPL* est une ressource en évolution. Nous présentons ici quelques uns des outils qui permettent son enrichissement.

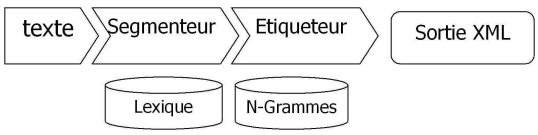


Figure 3: segmenteur et étiqueteur

Deux outils - un segmenteur et un étiqueteur- mettent en relation les mots d'un texte fourni en entrée avec les mots du lexique. La figure 3 illustre leur usage. Le segmenteur, basé sur des automates simples, effectue un découpage du texte en *tokens*. C'est à partir de ces informations que l'étiqueteur effectuera la désambiguisation en contexte des catégories à attribuer à chaque token.

La technique de désambiguisation que nous utilisons s'inspire des techniques stochastiques existantes. Nous avons cependant préféré développer notre propre étiqueteur afin de correspondre au mieux avec la précision des traits morphosyntaxiques que nous employons. Une première évaluation de l'étiqueteur sur le corpus du projet *Multitag* (cf. [Paroubek00]) a donné des résultats par catégorie variant de 60% à 99%. Le score moyen calculé sur le corpus de référence Multitag est de 95%.

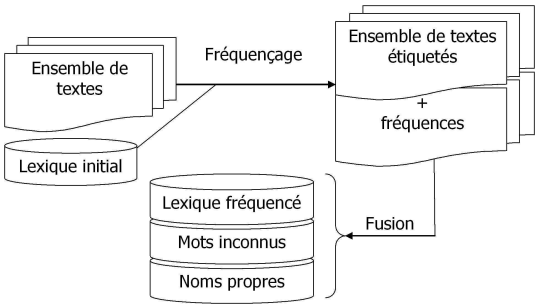


Figure 4: calcul des fréquences lexicales

Afin d'enrichir le lexique et de calculer au besoin les fréquences lexicales spécifiques à un ensemble de corpus, nous avons développé un outil de fréquençage. Comme l'indique la figure 4, cet outil fait appel aux résultats de l'étiquetage pour en déduire les fréquences des entrées du lexique, pour chaque couple *(mot, catégorie)*. A partir d'un lexique initial, étant donné un ensemble de textes, nous obtenons en sortie du fréquenceur un lexique des mots inconnus, un lexique des noms-propres et une nouvelle version du lexique initial, dont les champs *fréquence* sont mis à jour.

La version actuelle de *DicoLPL* dispose des fréquences acquises sur 153 millions de mots tirés du journal Le Monde, de ressources littéraires gratuites, de transcriptions de corpus oraux et des textes spécifiques (domaine médical, corpus de mails etc.).

D'autre part, la forme phonétisée des entrées est obtenue grâce à un phonétiseur inspiré du projet *Syntaix* (cf. [Di Cristo01]) pour la conception d'un système de synthèse vocale. D'autres champs (sémantique, valence verbale etc.) nécessitent toujours une validation manuelle.

L'évaluation des outils et du lexique est réalisée avec les techniques suivantes: Nous pouvons mesurer la **couverture** du lexique pour un corpus donné (calculer le quotient *nombre de mots reconnus / nombre total de mots*). La couverture actuelle du lexique représente 96% des corpus analysés (153 millions de mots). Lorsque nous souhaitons une information plus fine concernant l'étiquetage, il faut alors disposer d'un corpus de référence, pour lequel chaque mot est associé à une catégorie morphosyntaxique certifiée. Il est alors possible de mesurer les **scores de rappel et précision** pour chaque catégorie. C'est dans ce cadre que nous avons pu calculer un score de 95% sur le corpus de référence Multitag.

4 Un lexique noyau du français contemporain

Une fois le lexique constitué, il est nécessaire de vérifier sa couverture. Par ailleurs, l'analyse des corpus décrits plus haut permet de fournir des indications pour la constitution d'un dictionnaire minimal (ou *lexique noyau*) du français ayant une couverture maximale (un tel sous-lexique est toujours spécifique à un ensemble de corpus). Cette ressource est d'une grande importance pour le futur: Il n'est pas possible d'enrichir un grand lexique manuellement. Or, nombre d'informations ne peuvent aujourd'hui être acquises totalement automatiquement, notamment les informations sémantiques. Un lexique noyau permet d'identifier un nombre limité d'entrées lexicales qu'il est possible d'enrichir y compris manuellement. L'objectif est à terme de disposer d'une ressource lexicale très complète, comportant des informations syntaxiques, sémantiques, voire pragmatiques. Un lexique limité aux 10.000 formes les plus fréquentes couvre en moyenne 90% du français. Il s'avère donc intéressant de sélectionner un lexique noyau du Français contemporain avoisinant cette taille. La qualité de l'information concernant la fréquence de chacune des entrées du lexique complet permet de concevoir un lexique noyau (dorénavant LN) des mots les plus fréquents. C'est aussi l'occasion d'évaluer diachroniquement l'évolution du lexique de base du français depuis "Le Français Fondamental" (cf. [Gougenheim64] et [Blache05]). Nous avons sélectionné les formes pertinentes du LN grâce à une méthode simple utilisant une fréquence seuil (une autre méthode basée sur une réflexion à propos des types de catégories à conserver indépendamment de leur fréquence s'est révélée moins efficace et a été abandonnée). Ainsi, pour obtenir un dictionnaire de 10.000 formes (LN10) avons nous sélectionné les 10.000 entrées les plus fréquentes du lexique général DicoLPL, c'est-à-dire toutes les formes dont la fréquence est supérieure à 1091. Différentes versions de LN de taille croissante ont été produites suivant la même méthode : LN15 (fréquence>613, 15.017 formes), LN20 (fréquence>389,19.990 formes) et LN30 (fréquence>193, 30.018 formes) afin de comparer leurs couvertures et choisir le meilleur rendement taille/couverture.

DicoNoyau	Corpus écrit	Corpus oral
LN10 (f>1091)	88,63%	91,56%
LN15 (f>613)	91,07%	93,60%
LN20 (f>389)	92,50%	94,60%
LN30 (f>193)	94,08%	96,46%
DicoLPL	96,21%	99,02%

Figure 5: couvertures par taille de lexique et par type de corpus

Nous avons soumis les différentes versions du LN à un test de couverture sur deux types différents de corpus: un corpus écrit (580.000 mots extraits d'articles publiés dans le journal Le Monde) et un corpus oral (435.000 mots et regroupe le *Bristol Corpus*, un ensemble de 95 entretiens enregistrés et transcrits par Kate Beeching (1988-1990), ainsi que des corpus de parole recueillis au LPL).

Les résultats présentés dans la figure 5 appellent plusieurs commentaires: nous constatons d'abord que la couverture du lexique général DicoLPL (dernière ligne) n'est pas totale et qu'elle est meilleure pour le corpus oral que pour le corpus écrit, remarque qui vaut aussi pour les autres dictionnaires. Ceci s'explique selon nous par le fait que l'écrit utilise un vocabulaire beaucoup plus étendu et varié que l'oral. On constate aussi que les performances de couverture s'améliorent régulièrement au fur et à mesure que le LN contient plus de formes, ce qui est bien sûr attendu. Il faut néanmoins noter qu'il existe un saut qualitatif plus important entre LN10 et LN15 qu'entre LN15 et LN20 ou LN20 et LN30 alors même que l'écart de taille entre ces deux derniers est plus important. Le dictionnaire noyau de 15000 formes apparaît donc comme la version optimale pour obtenir la plus grande couverture avec un nombre réduit de formes.

5 Conclusion

La plateforme de développement de lexique décrite dans cet article répond à un certain nombre de besoins à la fois en termes de richesse d'informations, mais également de développements de lexiques spécialisés en produisant des fréquences spécifiques. Notre approche permet de rationaliser le choix des entrées sur lesquelles travailler en proposant la construction d'un lexique noyau élaboré sur la base d'une véritable analyse de la langue. L'enrichissement manuel de petits lexiques avec des informations sémantiques, pragmatiques etc. s'en trouve facilité. C'est pourquoi nous défendons la démarche qui consiste à concentrer les efforts sur un sous-lexique dont la couverture a été vérifiée sur corpus. D'autre part, un lexique de petite taille offre de nombreuses possibilités d'études sur l'usage avec notamment les *réseaux sémantiques*, les *petits mondes* etc. (voir à ce propos [Ferrer01]).

Le fait de disposer d'un grand lexique de formes n'en reste pas moins un atout, puisque c'est à partir d'une telle ressource que peuvent être extraits des sous-lexiques *ad hoc* couvrant des types de texte de domaines divers que le *fréquençage* permet d'isoler.

Enfin, la tâche de constituer une telle ressource est immense. Nous souhaitons la voir s'améliorer avec le temps, ce qui suppose sa diffusion, sa confrontation à l'usage et un retour de la communauté. La plateforme décrite ici comportant une série d'outils de maintenance, il est ainsi possible d'envisager une mise à jour régulière des informations. Au total, notre contribution viendrait s'inscire dans le mouvement de mise à disposition de ressources du français initié par les différents projets signalés plus haut.

Références

Association des Bibliophiles Universels, "ABU. Dictionnaire des mots communs", in La Bibliothèque Universelle, http://abu.cnam.fr/DICO/mots-communs.html. CNAM.

Blache P., M.-L. Guénot & C. Portes (2005), "Outils et ressources pour la mise à jour du Français Fondamental", in Proceedings of Français Fondamental: 50 ans de travaux et d'enjeux.

Clément L., B. Sagot & B. Lang (2004), "Morphology-Based Automatic Acquisition", in proceedings of LREC-04.

de Calmès M. & G. Pérennou (1998), "BDLEX : a Lexicon for Spoken and Written French", in proceedings of LREC-98

Di Cristo & P. Di Cristo (2001), "Syntaix : une approche métrique-autosegmentale de la prosodie", in revue TAL, 42:1

Ferrer R., Cancho I. & Sole R. (2001), "The small-world of human language", Proceedings of the Royal Society of London, B 268, 2261– 2266 url = "citeseer.ist.psu.edu/ferrer01small.html"

Gougenheim, G. ; Rivenc, P. ; Michéa, R. & Sauvageot, A. (1964), "L'élaboration du Français Fondamental", 1er degré, Didier : New B.

Pallier C., L. Ferrand & R.Matos (2001), "Une base de données lexicales du Français contemporain sur Internet : Lexique ", in L'Année Pschologique, 101

Paroubek P. & M. Rajman (2000), "MULTITAG, une ressource linguistique produit du paradigme d'évaluation", in Actes de la conférence TALN-2000

Romary L., S. Salmon-Alt & G. Francopoulo (2004), "Standards going concrete: from LMF to Morphalou", in Workshop on Electronic Dictionaries, COLING-04.

La plate-forme LinguaStream : un outil d'exploration linguistique sur corpus

WIDLÖCHER Antoine, BILHAUT Frédérik
GREYC - CNRS UMR 6072 - Université de Caen
Campus II, Sciences 3, B.P. 5186, 14032 Caen Cedex, France
{awidloch,fbilhaut}@info.unicaen.fr

Mots-clefs : Linguistique de corpus, TAL, plate-forme logicielle

Keywords: Corpus linguistics, NLP, software framework

Résumé À travers la présentation de la plate-forme LinguaStream, nous présentons certains principes méthodologiques et différents modèles d'analyse pouvant permettre l'articulation de traitements sur corpus. Nous envisageons en particulier les besoins nés de perspectives émergentes en TAL telles que l'analyse du discours.

Abstract By presenting the LinguaStream platform, we introduce different methodological principles and analysis models, which make it possible to articulate corpus processing tasks. More especially, we consider emerging approaches in NLP, such as discourse analysis.

1 Introduction

Par delà la diversité des domaines d'investigation et des objets d'étude, un certain nombre de tendances communes se confirment peu à peu au sein de la communauté TAL. Se manifeste tout d'abord, désormais distinctement, la généralisation du travail sur corpus, mouvement qui constitue d'ailleurs un point de convergence fécond entre les travaux spécifiquement dédiés au TAL et les démarches plus immédiatemment linguistiques. Les modèles théoriques proposés doivent désormais trouver leur justification et prouver leur validité « en corpus », et leur pertinence sera jugée, tantôt sur leur capacité à rendre compte de la diversité dudit corpus (dans une perspective descriptive), tantôt à la lumière de leur capacité à l'explorer « efficacement » (dans une perspective d'ingénierie documentaire). Se pose alors inévitablement la question de la méthode et des outils à adopter pour travailler ainsi « sur corpus ».

Il devient en effet difficilement envisageable de considérer celui-ci comme une matière brute à laquelle devraient se référer *immédiatement* les différents modèles et traitements. Au contraire, la multiplication des *points de vue* sur le corpus, qu'ils soient morphologiques, syntaxiques, sémantiques, rhétoriques ou pragmatiques, qu'ils ne visent que l'une de ces dimensions ou qu'ils les croisent, rend pressante la question des interdépendances entre ces vues possibles, interdépendances qui seront d'autant plus nombreuses que des résultats satisfaisants seront obtenus par chacune des approches. Une récente Journée d'Étude de l'ATALA a d'ailleurs permis de poser très frontalement la question désormais centrale de l'articulation des traitements sur corpus. Or, si l'articulation des traitements rend indispensable une réflexion sur leur modularité, elle conduit également à réinterroger l'ensemble de leur processus d'élaboration, de la prise en

charge du corpus, jusqu'à l'*évaluation* des résulats, en imposant que soient repensées les notions même d'*observation* et d'*expérimentation*, à travers, en particulier, une réflexion sur les cycles d'*évaluation/validation* puis d'ajustement des méthodes d'analyse.

Enfin, de nouvelles perspectives en TAL confirment ces nouveaux besoins. Si nous considérons l'intérêt récent accordé à l'analyse automatique de l'organisation discursive, par exemple en termes *thématiques* (Bilhaut, 2004) ou *rhétoriques* (Widlöcher, 2004), il apparaît clairement que ces investigations sont rendues possibles par la préexistence de résultats satisfaisants aux niveaux de granularité inférieurs, en matière d'analyse morpho-syntaxique et sémantique, aux niveaux lexicaux et syntagmatiques. Ces travaux se trouvent consubstantiellement liés à des stratégies d'*empilement* de traitements successifs permettant l'*enrichissement incrémental* des vues sur le corpus et l'abstraction progressive des formes de surface par l'utilisation des analyses préalables. À travers la présentation de LinguaStream[1], nous envisageons ici différents éléments méthodologiques et techniques pouvant permettre d'assumer ces nouvelles orientations.

2 La plate-forme LinguaStream

LinguaStream (Bilhaut & Widlöcher, 2005) a pour principale ambition de faciliter la réalisation d'expériences sur corpus non triviales en TAL, ainsi que le cycle d'évaluation/ajustement qui en découle. Sans outil adapté, le coût de développement induit par chaque nouvelle expérience devient en effet un frein considérable à l'approche expérimentale. Pour répondre à ce problème, LinguaStream facilite la mise en œuvre de procédés complexes tout en requérant des compétences informatiques minimales. Plate-forme générique fondée sur le principe d'**enrichissement incrémental** des documents électroniques, elle facilite la conception et l'évaluation de chaînes de traitements complexes, par assemblage visuel de modules d'analyse de types et de niveaux variés : morphologique, syntaxique, sémantique, discursif... Chaque palier de la chaîne de traitement se traduit par la découverte et le marquage de nouvelles informations, sur lesquelles pourront s'appuyer les analyseurs subséquents.

Un environnement de développement intégré permet de construire visuellement ces chaînes de traitement, à partir d'une « palette » de composants (une cinquantaine est intégrée en standard) facilement extensible grâce une API Java, un système de macro-composants, et des *templates*. Certains sont spécifiquement dédiés au TAL, et d'autres permettent de résoudre différents problèmes liés à la gestion des documents électroniques (traitements XML en particulier). D'autres peuvent être utilisés pour effectuer des calculs sur les annotations produites par les analyseurs, ou encore générer des diagrammes. Chacun dispose d'un ou plusieurs points d'entrée et/ou de sortie que l'on relie pour obtenir la chaîne voulue, celle-ci étant représentée par un graphe où les divers composants apparaissent sous forme de « boîtes » reliées entre elles. Chaque composant propose un nombre variable de paramètres permettant d'adapter leur comportement. Les marquages produits sur un même document sont organisés en couches indépendantes, supportant enchâssements et chevauchements. La plate-forme se base systématiquement sur les **standards et outils** XML, et peut traiter tout fichier de ce type en préservant sa structure orginelle. À l'exécution, elle se charge de l'ordonnancement des sous-tâches, et différents outils permettent *in fine* de visualiser les documents analysés et leurs annotations.

Principes fondamentaux

En premier lieu, la plate-forme recourt systématiquement à des **représentations déclaratives** pour spécifier les différents traitements, ainsi que leur enchaînement sous forme de graphe. Les différents formalismes disponibles permettent ainsi de transcrire directement l'expertise

[1] On trouvera une présentation complète de la plate-forme à l'adresse `http://www.linguastream.org`.

linguistique à mettre en œuvre, l'appareil procédural qui en résulte étant pris en charge par la plate-forme. Les règles données ont donc une valeur tant **descriptive**, en tant que représentations formelles d'un phénomène linguistique, que **prescriptive**, en tant qu'instructions de traitement fournies à un processus informatique.

De plus, la plate-forme exploite la **complémentarité des modèles d'analyse**, plutôt que de privilégier un hypothétique modèle « omnipotent » capable d'exprimer efficacement tout type de contrainte. Nous faisons en effet l'hypothèse qu'un analyseur complexe doit adopter successivement plusieurs regards sur le même matériau linguistique, auxquels répondront des formalismes distincts. On pourra combiner, au sein d'un même traitement, des expressions régulières au niveau morphologique, une grammaire locale au niveau syntagmatique, un transducteur au niveau phrastique et une grammaire DSDL (cf. *infra*) au niveau discursif. L'interopérabilité des différents modèles d'analyse proposés est garantie par l'usage d'une **représentation unifiée** des marquages et des annotations. Ces dernières sont uniformément représentées par des **structures de traits**, modèle communément utilisé en TAL et en linguistique, et permettant de représenter des annotations riches et structurées. Tout composant d'analyse pourra produire son propre marquage en s'appuyant sur les analyses précédentes : les formalismes proposés permettent de spécifier des contraintes sur les annotations existantes par **unification.** La plate-forme favorise ainsi l'**abstraction progressive des formes de surface.** Chaque palier d'analyse pouvant accéder simultanément aux annotations produites par tous les paliers antérieurs, les analyseurs de plus haut niveau sont généralement conduits à s'abstraire progressivement du matériau textuel pour ne plus reposer que sur des représentations symboliques antérieurement calculées.

Un autre aspect important concerne la **variabilité du grain d'analyse** au cours du traitement. De nombreux modèles d'analyse imposent la définition d'un grain d'analyse minimal, dit « jeton » ou *token*. C'est par exemple le cas de toute grammaire ou transducteur : ces formalismes supposent l'existence d'une unité textuelle (comme le caractère ou le mot) à laquelle s'appliquent les patrons. Quand la définition de ce grain minimal est nécessaire au fonctionnement d'un composant, la plate-forme permet de spécifier **localement** le ou les types d'unités à considérer comme jetons. Toute unité préalablement délimitée peut jouer ce rôle : il pourra s'agir du découpage habituel en mots, mais aussi de toute autre unité ayant été préalablement marquée : syntagmes, phrases, cadres du discours, etc. Le grain minimal peut donc être différent pour chaque palier de l'analyse, ce qui augmente considérablement la portée des différents modèles d'analyses utilisables dans la plate-forme. D'autre part, chaque module d'analyse spécifie les marquages antérieurs auxquels il souhaite faire référence, pouvant ainsi ne tenir compte que des marquages qu'il estime pertinents, et donc s'affranchir partiellement de la linéarité du texte. La combinaison de ces fonctionnalités permet d'adopter un **point de vue** sur le document spécifique à chaque étape d'une chaîne de traitement.

La **modularité** des chaînes de traitements favorise quant à elle la **réutilisabilité** des composants dans des contextes différents : un module d'analyse développé au sein d'une première chaîne pourra être réutilisé dans d'autres chaînes. De façon similaire, toute chaîne pourra être réutilisée en tant que constituant d'une chaîne de plus haut niveau, sous forme de « macro-composant ». Réciproquement, pour une chaîne donnée, on pourra **substituer** à un composant tout autre composant **fonctionnellement équivalent.** Pour une sous-tâche donnée, un prototype rudimentaire pourra être remplacé *in fine* par un équivalent pleinement opérationnel. Ceci rend possible la mise en comparaison des traitements, en soumettant ces derniers à des contextes rigoureusement identiques, condition *sine qua non* d'une comparaison pertinente.

Modèles d'analyse

Nous avons évoqué plus haut quelques-uns des composants susceptibles de prendre part à une chaîne de traitement. Parmi ceux spécifiquement dédiés au TAL, on peut distinguer deux familles. La première regroupe les analyseurs « prêts à l'emploi », dédiés à une tâche précise.

Il s'agira par exemple de l'étiquetage morpho-syntaxique, une interface avec TreeTagger étant intégrée par défaut. Ces composants sont paramétrables, mais il n'est pas possible de modifier fondamentalement leur fonctionnement. D'autres au contraire proposent un *modèle d'analyse*, c'est-à-dire un formalisme de représentation de contraintes linguistiques, éventuellement associé à un modèle opératoire, par lequel l'utilisateur peut spécifier intégralement le traitement à opérer. Ils permettent d'exprimer des contraintes tant sur les formes de surface que sur les annotations insérées par les analyseurs précédents. Toutes les annotations sont représentées sous forme de structures de traits, et les contraintes sont systématiquement spécifiées par unification sur ces structures. Quelques-uns des systèmes proposés sont :

- Un système appelé EDCG (pour *Extended*-DCG), permettant de décrire des **grammaires locales d'unification** en se basant sur la syntaxe DCG (*Definite Clause Grammars*) de Prolog. Une telle grammaire peut être décrite dans le plus pur style déclaratif, bien que les spécificités du langage logique restent accessibles aux utilisateurs expérimentés.
- Un système, nommé MRE (pour *Macro-Regular-Expressions*), permettant de décrire des patrons par **transducteurs**, s'appliquant aussi bien aux formes de surface qu'aux annotations préalablement calculées. Sa syntaxe est similaire à celle des expressions régulières communément utilisées en TAL et en linguistique sur corpus. Mais à la différence de ces dernières, ce formalisme ne s'applique pas spécifiquement aux caractères ni aux mots, et peut porter sur toute unité textuelle préalablement analysée.
- Un formalisme d'expression de **contraintes au niveau discursif.** En cours d'élaboration, DSDL (*Discourse Structure Description Language*), que nous décrirons plus loin, permet l'exploration des organisations discursives par l'expression et la satisfaction de contraintes, pouvant être non séquentielles exprimées à l'aide d'un ensemble de fonctions discursives primitives (présence/absence, cohérence sémantique...), et pouvant porter en particulier sur les annotations produites en amont et sur des relations entre ces dernières.
- Un système d'annotation à partir de **lexiques sémantiques**, un système de **tokenisation** basé sur des expressions régulières (au niveau caractère), un système permettant de délimiter des objets linguistiques en se basant sur le balisage XML du document, etc.

3 Analyse du discours

Voyons quels avantages l'analyse automatique du discours peut tirer des principes proposés. Un premier apport significatif résulte de l'approche par enrichissement incrémental et par abstraction progressive des formes de surface. S'il est naturel d'opérer au niveau de ces dernières pour une analyse par exemple morphologique ou syntaxique, il va sans dire que l'analyse discursive ne peut s'accomoder de la diversité combinatoire apparaissant à ce niveau, et qu'un filtrage s'impose. En plus de la possibilité d'opérer la pure et simple *occultation* d'éléments peu pertinents pour tel ou tel besoin interprétatif, la plate-forme permet d'opérer ladite abstraction de deux manières complémentaires. En premier lieu, l'unicité du modèle de marquage et d'annotation donne à chaque étape d'analyse l'accès aux représentations symboliques produites en amont, et permet ainsi de ramener la diversité combinatoire de surface à celle des valeurs interprétées, généralement moins nombreuses. En second lieu, le principe de variabilité du grain d'analyse déjà évoqué permet d'exploiter au niveau discursif des modèles d'analyse habituellement dédiés à des niveaux de granularité inférieurs. Par exemple, des règles EDCG pourront aussi bien décrire des patrons syntagmatiques qu'une grammaire textuelle, selon le grain choisi.

Par ailleurs, la plate-forme propose des modèles d'analyse spécifiquement adaptés au niveau discursif. Le langage DSDL en particulier, s'écarte des paradigmes généralement adoptés par les autres formalismes (y compris MRE ou EDCG), qui reposent fondamentalement sur des principes de *linéarité* (on tient compte de tous les éléments successifs) et de *séquentialité* (un ordre est imposé), principes souvent inadaptés au niveau discursif. En permettant l'expression de

contraintes *non séquentielles* et *non linéaires*, le formalisme DSDL autorise l'expression et la détection de motifs pouvant porter sur des éléments distants dans le texte, sans faire d'hypothèse sur leur ordre, ce qui s'avère particulièrement adapté à l'analyse du discours.

Afin de donner une idée plus concrète des principes méthodologiques présentés, envisageons à présent une configuration linguistique particulière, assez représentative des problèmes posés par l'analyse discursive, en abordant le problème de l'encadrement du discours (Charolles, 1997), et plus particulièrement de la détection automatique des *cadres temporels*. Rappelons que cette théorie qualifie ainsi des segments textuels homogènes du point de vue d'un critère d'interprétation fixé dans une expression en position détachée en début de phrase, dite *introducteur de cadre*. L'opérationnalisation en TAL de ce modèle psycho-linguistique impose la résolution de deux problèmes principaux : détection des introducteurs, puis évaluation de leur *portée*, c'est-à-dire détermination de la borne droite du cadre introduit. Bien que cette dernière tâche soit très problématique dans la mesure où les critères formels de clôture des cadres sont difficiles à établir, un certain nombre d'indices ont toutefois pu être dégagés dans le cas précis des cadres temporels (Bilhaut *et al.*, 2003), que nous évoquerons ci-après.

Le problème de la détection des introducteurs temporels se décline lui-même en deux sous-problèmes : l'analyse des expressions temporelles, et celle des introducteurs s'appuyant sur elles. Les principes de modularité évoqués trouvent ici leur justification, puisque nous souhaiterons généralement traiter ces problèmes indépendamment. L'analyse sémantique des expressions temporelles fait l'objet d'une grammaire EDCG, exprimant des contraintes sur les résultats d'une analyse morpho-syntaxique préliminaire, et associant aux expressions reconnues une représentation de leur « sens » sous forme de structures de traits. Sur cette base, la détection des introducteurs peut être mise en place à l'aide de critères essentiellement positionnels. Les contraintes exprimées sont fondamentalement séquentielles : nous recherchons des zones de texte vérifiant des motifs imposant la présence, dans un ordre fixé, d'éléments immédiatement successifs. Ces règles sont donc simplement exprimables à l'aide du formalisme MRE (outre les expressions temporelles, nous exploitons ici le marquage des phrases et des connecteurs de discours) :

```
{type : phrase, anchor : start}
<introducteur>
{type : connecteur}? {tag : pre} {type : temporel} /as $t
</introducteur> /sem {axe : temps, valeur : $t} ","
```

Les contraintes sur les structures de traits produites en amont (ici en gras), ainsi que sur les formes de surface (ici, la virgule en fin de motif) permettent de délimiter l'introducteur. Nous recherchons les éléments précédés d'un début de phrase et composés, d'un éventuel connecteur de discours, d'une préposition et d'une expression temporelle. Le reste de l'expression correspond au marquage et à l'annotation produits en sortie. L'élément reconnu aura le type « introducteur » et sera associé à l'annotation sémantique qui lui fait suite. Précisons que la variable `$t` permet de faire « remonter » l'information contenue dans la structure de traits associée à l'expression temporelle, pour un usage ultérieur.

Pour la détermination de la portée de l'introducteur, la méthode présentée dans (Bilhaut *et al.*, 2003) s'appuie sur des critères énonciatifs tels que la cohésion des temps verbaux, sur la structuration en paragraphes, et sur des calculs sémantiques de cohérence entre l'introducteur et les autres expressions temporelles. La nature de ces contraintes diffère radicalement des précédentes. D'une part, nous pouvons désormais nous abstraire de la linéarité du texte : contrairement à une approche par expressions régulières, nous pouvons ici ignorer un certain nombre d'éléments du flot textuel. D'autre part, s'il existe bien des contraintes interprétatives entre l'introducteur et certains éléments de la zone introduite, il n'est pas souhaitable de concevoir ces contraintes comme imposant un ordre strict entre ces éléments. Pour l'expression de telles contraintes à la fois *non linéaires* et *non séquentielles*, nous disposons du formalisme DSDL

et pouvons formuler la « grammaire » ci-dessous. Nous recherchons une unité textuelle composée de phrases complètes, commençant par un élément identifié comme introducteur et ne comportant pas d'autre élément de ce type, dont tous le verbes sont au même temps, et au sein de laquelle les expressions temporelles portent sur une plage comprise dans l'intervalle fixé par l'introducteur, en ne retenant que le plus long des candidats partageant un même introducteur.

```
Rule {type : "cadre"} :
        start({type : "introducteur"})
        end({type : "phrase"})
        homogeneity(comparator : portée)
        not presence(pattern : {type : "intro"}, amount : 2)
        size(mode : #LONGEST)

Comparator portée ({type : "verbe"} as $v1, {type : "verbe"} as $v2) :
        $v1/temps = $v2/temps

Comparator portée ({type : "intro"} as $i, {type : "tempo"} as $t) :
        ($t/debut >= $i/debut) and ($t/fin <= $i/fin)
```

Il est ainsi possible, à l'aide des principes méthodologiques promus par la plate-forme, et en nous appuyant sur la complémentarité des modèles d'analyse, de mettre en place un analyseur de cadres temporels, certes encore imparfait, mais ne faisant usage que de formalismes purement déclaratifs propices à la capitalisation de l'expertise linguistique mise en œuvre.

4 Conclusion

Initialement développée dans le cadre du projet GeoSem[2], la plate-forme évolue maintenant indépendamment. Elle est aujourd'hui utilisée dans le cadre d'un projet TCAN[3], de différents travaux de recherche en TAL, notamment en analyse sémantique du discours : organisation thématique (Bilhaut, 2004), ou rhétorique (Widlöcher, 2004). Le logiciel est également utilisé à des fins pédagogiques au GREYC et à l'ERSS, et a été mis à la disposition de laboratoires tels que LIUPPA ou LATTICE. La plate-forme reste en elle-même indépendante des modèles d'analyse utilisés, pour peu qu'ils partagent le même système de marquage et d'annotation, et il est donc envisageable d'intégrer des modules exploitant d'autres modèles d'analyse.

Références

BILHAUT F. (2004). Analyse automatique de la structure thématique du discours pour la navigation documentaire. In *Journée ATALA « Modéliser et décrire l'organisation discursive à l'heure du document numérique »*.

BILHAUT F., HO-DAC M., BORILLO A., CHARNOIS T., ENJALBERT P., DRAOULEC A. L., MATHET Y., MIGUET H., PÉRY-WOODLEY M.-P. & SARDA L. (2003). Indexation discursive pour la navigation intradocumentaire : cadres temporels et spatiaux dans l'information géographique. In *Actes de Traitement Automatique du Langage Naturel (TALN)*, Batz-sur-Mer, France.

BILHAUT F. & WIDLÖCHER A. (2005). La plate-forme LinguaStream. In *Journée ATALA « Architectures logicielles pour articuler les traitements sur corpus »*.

CHAROLLES M. (1997). L'encadrement du discours : Univers, champs, domaines et espaces. Cahier de Recherche Linguistique no 6. Université de Nancy 2.

WIDLÖCHER A. (2004). Analyse macro-sémantique : vers une analyse rhétorique du discours. In *Actes de RECITAL 2004*, p. 183–188, Fès, Maroc.

[2] « Traitement sémantique de l'information géographique », programme CNRS « Société de l'information ».
[3] « Intervalles temporels et applications à la linguistique textuelle », projet interdisciplinaire du CNRS.

RECITAL 2005

9ème Rencontre des Étudiants Chercheurs
en Informatique pour le
Traitement Automatique des Langues Naturelles

CONFÉRENCE PRINCIPALE

How semantic is Latent Semantic Analysis?

Tonio Wandmacher

Laboratoire d'Informatique – E.A. 2101 – Equipe BdTln
Université François Rabelais de Tours
Site de Blois – 3 place Jean Jaurès – F-41000 Blois
tonio.wandmacher@etu.univ-tours.fr

Mots-clefs - Keywords

Analyse de la sémantique latente, analyse de collocations, relations lexicales, sémantique computationelle.

Latent Semantic Analysis, Collocation Analysis, lexical relations, computational semantics.

Résumé - Abstract

Au cours des dix dernières années, l'analyse de la sémantique latente (LSA) a été utilisée dans de nombreuses approches TAL avec parfois de remarquables succès. Cependant, ses capacités à exprimer des ressemblances sémantiques n'ont pas été réellement recherchées de façon systématique. C'est l'objectif de ce travail, où la LSA est appliquée à un corpus de textes de langue courante (journal allemand). Les relations lexicales entre un mot et ses termes les plus proches sont analysés pour un test de vocabulaire. Ces résultats sont alors comparés avec les résultats obtenus lors d'une analyse des collocations.

In the past decade, Latent Semantic Analysis (LSA) was used in many NLP approaches with sometimes remarkable success. However, its abilities to express semantic relatedness were not yet systematically investigated. This is the aim of our work, where LSA is applied to a general text corpus (German newspaper), and for a test vocabulary, the lexical relations between a test word and its closest neighbours are analysed. These results are compared to the results from a collocation analysis.

1 General introduction

In its beginnings, Latent Semantic Analysis aimed at improving the vector space model in information retrieval. Its abilities to enhance retrieval performance were remarkable; results could be improved by up to 30%, compared to a standard vector space technique (Dumais, 1995). It was further found that LSA was able to retrieve documents that did not even share a single word with the query but were rather semantically related.

This finding was the headstone for many subsequent researches. It was tried to apply the LSA approach to other areas, such as automated evaluation of student essays (Landauer et al., 1997) or automated summarization (Wade-Stein & Kintsch, 2003). In (Landauer & Dumais, 1997), even an LSA-based theory of knowledge acquisition was presented.

In these works, many claims on the analytic power of LSA were made. It is asserted that LSA does not return superficial events such as co-occurrence relations, but is able to describe semantic similarity between two words.[1] The extracted word relations are referred to as latent, hidden or deep[2], however, none of these papers addresses the nature of this deepness. LSA is called "semantic", but a thorough evaluation of its abilities to extract the semantics of a word or a phrase is missing.[3] One work that takes a little step in this direction, was done by Landauer & Dumais (1997). They use LSA-based similarities to solve a synonym test taken from the *TOEFL* (Test Of English as a Foreign Language) They found that the abilities of LSA to assign the right synonym (out of 4 test words) to the target word are comparable to those of human non-native speakers of English (mean LSA: 64,4%; mean humans: 64,5%).

However, this result can only be seen as a first indication for the capacity of LSA; it is neither a systematic assessment, nor a comparison to similar techniques. This is what we try to achieve in the following. Our aim is therefore not improvement, but evaluation and a better understanding of the method.

2 Presentation of LSA

LSA, as presented by (Deerwester et al. 1990) and others, is based on the vector space model of information retrieval (Salton & McGill, 1983). First, a given corpus of text is transformed into a term×context-matrix, displaying the occurrences of each word in each context. A context can be only a 2-word window, a sentence, a paragraph or a full text. For LSA, a paragraph window is normally assumed (cf. (Dumais, 1995), (Landauer et al, 1997)).

In a second step, this matrix is weighted by one of the weighting methods used in IR (c.f. (Salton & McGill, 1983)). For LSA, a log-entropy scheme showed the best results (Dumais, 1990). The decisive step in the LSA process is then a *Singular Value Decomposition* (*SVD*) of the weighted matrix. Thereby the original matrix A is decomposed as follows:

$$\mathrm{SVD}(A) = U \Sigma V^{\mathrm{T}} \qquad (1)$$

The matrices U and V consist of the eigenvectors of the columns and rows of A. Σ is a diagonal matrix containing in descending order the singular values of A. By only keeping the k

---

[1] Cf. (Wade-Stein & Kintsch, 2003), p. 10: "*LSA does not reflect the co-occurrence among words but rather their semantic relatedness.*"

[2] Cf. (Landauer et al., 1998), p. 4: "*It is important to note from the start that the similarity estimates derived by LSA are not simple contiguity frequencies, co-occurrence counts, or correlations in usage, but depend on a powerful mathematical analysis that is capable of correctly inferring much deeper relations.*"

[3] The 'latency' of LSA was indeed assessed by Wiemer-Hastings (1999).

strongest (k usually being around 300) singular values and remultiplying Σ_k with either U or V, one can construct a so-called *semantic space* for the terms or the contexts, respectively. Each term or each context then corresponds to a vector of k dimensions, whose distance to others can be compared by a standard vector distance measure. In most LSA approaches the cosine measure is used. By calculating the cosine of the angle between one term vector and all the others, a ranked list of next neighbours can be obtained for a given word. From the LSA point of view, these neighbours should be semantically related to the test word.

3 Method

To assess the abilities of LSA to generate semantic similarity, we applied it to a large corpus of German newspaper text. We used a random sample of 120.000 paragraphs (app. 20 mio. words) of the *Tageszeitung* (*TAZ*) from 1989 to 1998, which was stoplisted for frequent words and lemmatized by the *DMOR* package (Schiller, 1995). Words having a corpus frequency of less than 5 were also removed. This reduced the vocabulary size from 385.344 to 63.651 types. This was necessary, since the calculation of the SVD is heavily constrained by complexity matters.

After transforming the corpus into a term×context matrix (having the size 63.651 × 120.000), we applied a log-entropy weighting scheme. Using Michael Berry's GTP package v. 3.0 for Linux, we calculated the SVD for the above matrix up to 400 dimensions. To find the optimal factor k, we conducted some preliminary tests with various dimensionalities (250 – 400). A k of 309 gave the best results (in terms of the percentage of meaningful relations), even though the results for each of the samples were very close.

For a random sample of 400 words (nouns, verbs and adjectives only), their 20 next neighbours (= words having the highest cosine score with the centroid, see 2.) were extracted. We considered a fixed number of neighbours, since the usage of a threshold distance (e.g. *cos* = 0,5) proved not to be practical (the cluster size varied strongly).

The relations between the centroid (test word) and each of the 20 neighbours were then manually categorized in one of eight relation classes. The classes were the following:

- Synonymy
- Antonymy
- Hypo-/Hypernymy
- Co-Hyponymy
- Mero-/Holonymy

} truly semantic relations

- Loose association
- Morphological relation
- Erroneous relation

The notion of semanticity described by this classification can be questioned. However, our selection of semantic relations seems to be widely accepted in lexical semantics (cf. (Cruse, 1986)) and precise definitions exist to determine if a relation holds between two words X and Y (e.g. for meronymy: X IS-PART-OF Y). The same is true for the class of morphological

relations. A derivational or inflectional relation between two terms can be recognized easily most of the time.

Still, we admit that our classification is neither exhaustive, nor always clear-cut. Especially the class of "loose association" is rather intuitive. It was assumed as a collection class for all term pairs that were not related by definition of a semantic or a morphological relation, but still were somehow connected. Typical examples for this class might be '*Flugzeug*' ('airplane') / '*landen*' ('to land') or '*Katze*' ('cat') / '*Milch*' ('milk').

To balance out doubtful cases, we set the size of our test sample sufficiently large (20 neighbours for 400 words = 8000 categorized relations[4]) and had the classification task done independently by two German native speakers (including the author).

For each of the neighbours, additional information, such as its corpus frequency, context frequency and entropy value, was also determined.

4 Results

4.1 Quantitative Analysis

Results were calculated for the first 5, 10, 15 and 20 neighbours, respectively. As the fractions for each of the semantic classes were all quite low (0-5%), only the total of semantic relations is displayed here:

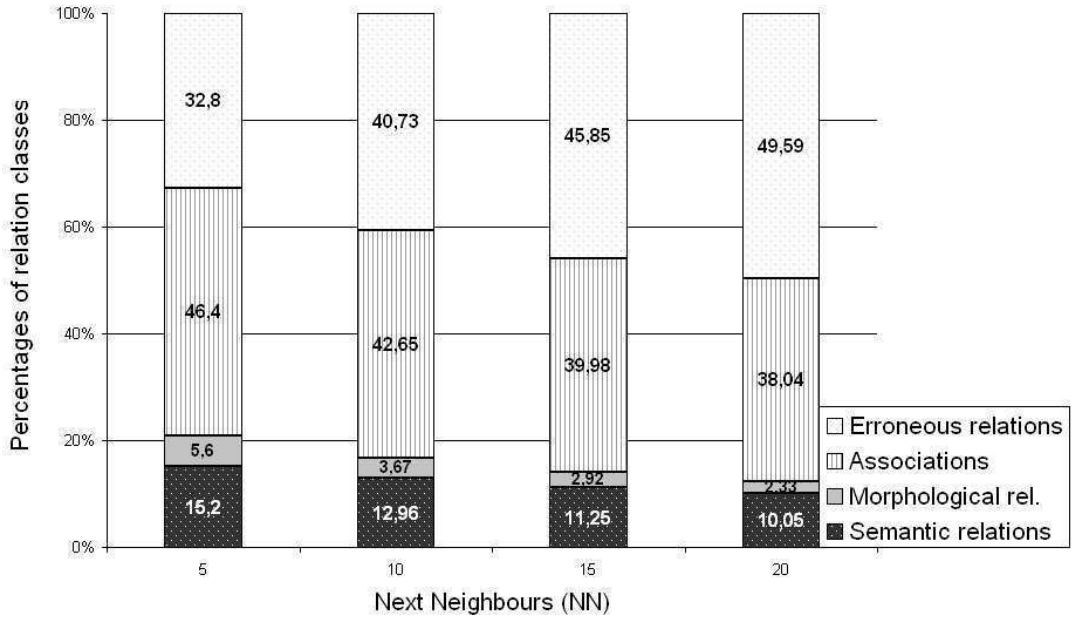


Figure 1: Percentages of relation classes resulting from LSA for a sample of 400 test words.

[4] For a sample size n = 400 the 95% confidence interval is maximally ± 4,9.

As the number of neighbours under consideration rises, the results get generally worse. Considering only the first five neighbours for every test word, we find nearly a third of erroneous relations (32,8%). Almost half of them are loose associations, whereas the truly semantic relations make up only 15%. Moreover, 5% of morphologically related word pairs were found.

When we consider the relations for 20 neighbours, the part of erroneous relations rises to nearly 50%, whereas the class of associations falls to 38%. Only 10% of the relations can be classified as semantic and app. 2% as morphological.

Splitting up the sample into parts of speech, we get the following picture:

Relation class	**Nouns**	**Adjectives**	**Verbs**
Semantic relations:	15,1%	7,8%	3,8%
Morphological relations:	2,4%	1,8%	2,1%
Associations:	39,3%	30,5%	41,0%
Erroneous relations:	43,2%	59,9%	53,1%

Table 1: Percentages of relations for the three parts of speech.

Table 1 shows a clear distinction. Nouns have far more meaningful relations (56,8%) than adjectives (40,1%) or verbs (46,9%). The difference becomes even clearer if only the class of semantic relations is considered. Opposing verbs and adjectives, another remarkable difference can be found: Verbs have much more (10,5%) associations than adjectives, but only less than half of semantic relations.

4.2 Qualitative Analysis

If one opposes the words with the lowest and the highest fractions of meaningful relations, a difference in usage of the two groups can be observed:

Lowest fractions (0-5%)		**Highest fractions (95-100%)**	
Ansehen (‚image')	natürlich (‚natural')	Mediziner (‚health prof.')	singen (‚to sing')
Aufbruch (‚breakup')	teilen (‚divide')	Reporter (‚reporter')	sterben (‚to die')
Beispiel (‚example')	zahlreich (‚numerous')	Therapie (‚therapy')	studieren (‚to study')
Kasten (‚box')	zumuten (‚to expect of')	Wohnraum (‚living space')	Luftwaffe (‚air force')
Umstand (‚circumstance')	überstehen (‚to overcome')	Zuhörer (‚auditor')	Malerei (‚painting')
Unsinn (‚nonsense')	Auflösung (‚resolution')	deportieren (‚to deport')	Religion (‚religion')
aufrecht (‚upright')	Rücksicht (‚consideration')	gesund (‚healthy')	Uniform (‚uniform')
automatisch (‚automatic)	bescheren (‚to bring')	kochen (‚to cook')	Wirtschaft (‚economy')
glatt (‚flat'/‚smooth')	denken (‚to think')	lernen (‚to learn')	lesen (‚to read')
intensiv (‚intensive')	einfallen (‚to occur')	operieren (‚to do a surgery')	orthodox (‚orthodox')

Table 2: Words of the sample having the lowest and highest fractions of meaningful relations with the test word.

Regarding the two groups shown in table 2, it appears that the words with the worst results can occur in every context. Words like '*Beispiel*' ('example'), '*Unsinn*' ('nonsense') or '*denken*' ('to think') are not connected to a certain theme or a typical context. On the other

hand, the words having the highest scores are rather specific. This group comprises terms such as '*Mediziner*' ('health professional'), '*Malerei*' ('painting') or '*kochen*' ('to cook'). These terms have a typical context; they are bound to a particular topic.

To get a clearer picture of this kind of specificity, it seems reasonable to further analyse the distribution of the words in the corpus. From the research on information retrieval it is known that specific terms are better predictors and get therefore a higher weight (Spärck-Jones, 1972), (Salton & McGill, 1983). The relevant values for the weighting schemes in IR are normally the *term frequency* (*tf*), the *corpus frequency* (*cf*) and the *context* (or *document*) *frequency* (*df*). A combination of these values forms the base for nearly all weighting schemes of the so-called *tf*idf*-family (s. (Salton & McGill, 1983)). Could these values be good predictors for our purposes?

We calculated the correlation between several of these values (as well as some combinations) and the fraction of meaningful relations among 20 next neighbours. The results were disappointing. None of the pure frequencies had a significant correlation with the fraction of meaningful relations. While trying out several combinations of the values, we found only one that showed a slight correlation (*Pearson*-Coefficient = 0,32, significance level <0,001), namely the quotient of *cf* and *df*.

In addition, we calculated the correlation between the mean distance of the 20 neighbours and the percentage of meaningful relations for the whole test set. We observed a medium correlation (*Pearson*-Coeff. = 0,56 at a significance level of <0,001). We therefore can conclude that medium distance and relation quality are related, although not too strongly.

5 Comparison with collocation analysis

5.1 Method

To obtain a contrastive example, we did the same experiment using collocation analysis (CA). We hereby used a formula presented by Quasthoff & Wolff (1998), (2002). They calculate the collocative significance between two words *A* and *B* as follows:

$$sig(A,B)=\frac{C_A C_B}{n}-k\cdot\log\frac{C_A C_B}{n}+\log k! \qquad (2)$$

where C_A (C_B) is a context, in which *A* (*B*) occurs, *n* is the amount of all contexts and *k* the number of all contexts containing *A* and *B*.[5]

To ensure comparability, we used the same corpus and did the same pre-processing (i.e. stoplisting, removal of low frequency words etc.). We then calculated for our sample of 400 test words the 20 words having the highest collocative significance with the test word. This gave us again 400 word clusters, for which every relation was categorized as above.

[5] This measure is obviously related to the one given by Dunning (1993). Both measures appear to give rather similar results (cf. (Quasthoff & Wolff, 2002)).

5.2 Results

Figure 2 shows the fractions of the different relation classes obtained by collocation analysis:

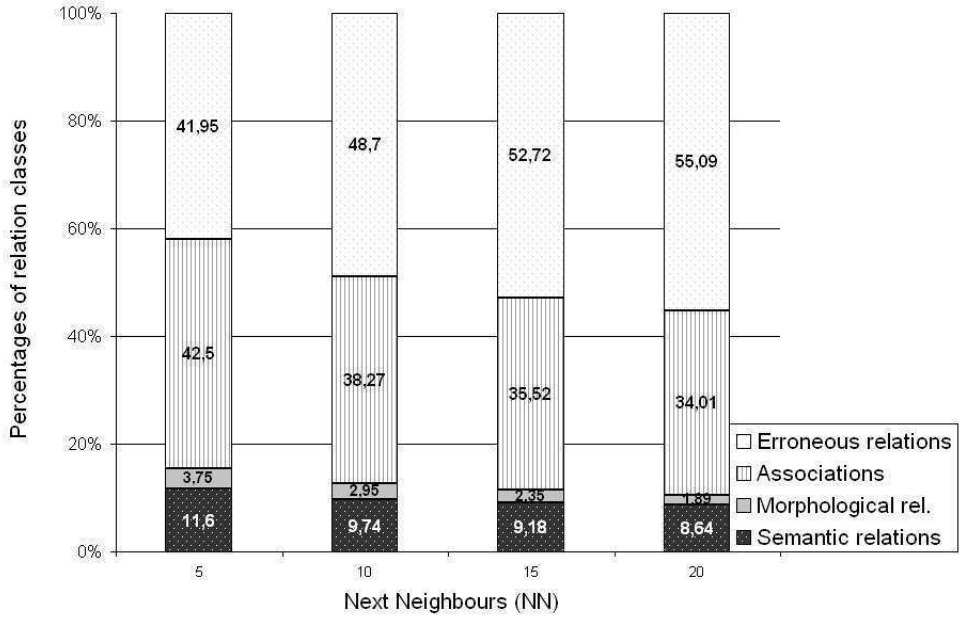


Figure 2: Percentages of relation classes resulting from CA for a sample of 400 test words.

Comparing figures 1 and 2, the results seem quite close. The differences between the fractions for the meaningful classes do not exceed 4,5%. This is remarkable, keeping in mind that the collocation analysis exploits only co-occurrence relations in the text.

As a whole, the percentage of meaningful relations is a little lower for CA than for LSA. Still we find 8% (20 NN) to 12% (5 NN) of semantic relations, not much less regarding the LSA results (10%-15%). Still, all LSA scores were significantly higher than the CA scores in a Student's T test (significance level: <0,001).

Regarding the words and their neighbours, we get a similar picture as with LSA: terms being bound to a particular theme get better results than those that are context-independent. Moreover, the results of CA and LSA show a high correlation (*Pearson*-Coeff. = 0,72 at a significance level of <0,001) for our sample. It seems therefore that LSA and CA make use of the same properties of a word and its context.

However, one difference can be observed with regard to ambiguous words. CA returns neighbours that still belong to both meanings, LSA however seems to mask out one of them. This behaviour can be seen in the following examples for '*Bach*' (Meanings: J. S. Bach, the composer / 'creek') and '*Schlange*' (Meanings: 'queue' / 'snake'):

	Next neighbours for '*Bach*' ('creek'/J.S. Bach)		Next neighbours for '*Schlange*' ('queue'/'snake')	
Rank	**LSA results**	**CA results**	**LSA results**	**CA results**
1.	Musik (‚music')	Sebastian (‚Sebastian')	stehen (‚to stand')	stehen (‚to stand')
2.	Beethoven (‚Beethoven')	Johann (‚John')	warten (‚to wait')	warten (‚to wait')
3.	musizieren (‚to make music')	Musik (‚music')	Schild (‚sign')	Reptil (‚reptile')
4.	klanglich (‚sonorical')	Emanuel (‚Emanuel')	Wartezeit (‚waiting time')	lang (‚long')
5.	musikalisch (‚musical')	Elvira (‚Elvira')	Kaufhaus (‚department store')	bilden (‚to form')
6.	Klang (‚sound')	Mozart (‚Mozart')	Supermarkt (‚supermarket')	Schalter (‚counter')
7.	Gesang (‚chant')	Artist (‚artist')	Vesna (‚name')	Kaninchen (‚rabbit')
8.	rhythmisch (‚rhythmical')	runtergehen (‚to flow down')	lang (‚long')	Buchstabe (‚letter')
9.	komponieren (‚to compose')	verunreinigen (‚to pollute')	Straßenrand (‚roadside')	Mensch (‚human')
10.	Improvisation (‚improvisation')	Brahms (‚Brahms')	Bäckerei (‚bakery')	giftig (‚poisonous')
11.	Mozart (‚Mozart')	Fluss (‚river')	tagsüber (‚in the day')	auftauchen (‚to emerge')
12.	virtuos (‚'virtuoso)	Ton (‚sound/note')	Matte (‚mat')	einreihen (‚to queue')
13.	Komposition (‚composition')	hinunter (‚down')	Stau (‚holdup')	Warteschlange (‚queue')
14.	Rhythmus (‚rhythm')	Gewässer (‚water')	Warteschlange (‚queue')	lange (‚long')
15.	Saxophon (‚saxophone')	Geige (‚violin')	Einlass (‚entry')	Australien (‚Australia')
16.	Geige (‚violin')	Oboe (‚oboe')	Brot (‚bread')	Stau (‚holdup')
17.	Komponist (‚composer')	Flussufer (‚river bank')	Greenfield (‚Greenfield')	Tag (‚day')
18.	akustisch (‚acustical')	Aufführung (‚performance')	lange (‚long')	Käfig (‚cage')
19.	klassisch (‚classical')	rauschen (‚rush')	Auslage (‚display')	öffnen (‚to open')
20.	Cello (‚cello')	Philipp (‚Philipp')	Mittelpunkt (‚center')	Wartezeit (‚waiting time')

Table 3: Two examples of ambiguous words and their next neighbours. The neighbours belonging to the prominent meaning are in blank, the ones of the non-prominent meaning are in black. Terms that cannot be assigned are shaded in grey.

The difference in the analyses is obvious: in both examples, LSA generates neighbours of the prominent meaning only, whereas the NN produced by CA are of both domains. However, this is only a first observation; we did not yet assess this masking-out property of LSA in a systematic way. It should be subject to further research.

6 Discussion

To take up our initial question: how semantic is LSA? The conclusion that we draw from our results, is: not as much as its name might suggest. The fractions of truly semantic relations were not very high (10% at the 20-NN level), and a big part (38% at 20 NN) of the relations, however, could rather be described as associative. These words are conceptually related, but not necessarily in a narrow semantic or morphological sense.

The biggest part however, nearly half of the relations generated at the 20-NN level, are erroneous, i.e. there is no apparent relation between the test word and its neighbour.

Comparing the LSA approach to other procedures exploiting co-occurrence information, one reason for the high percentage of error relations may be found: Techniques such as *HAL* (Lund & Burgess, 1996) or the approach by Rapp (2002), (2003) use a co-occurrence window of a few (3-40) words only. LSA however relies on full paragraphs (average length in our case: 102 words). And even though a paragraph can be regarded as a semantically coherent unit, many of the inter-word relations in it may already be too weak. This may cause arbitrary relations.

Another questionable point about LSA arises from the modelling itself. The term-by-context-matrix is extremely sparse. In our experiments, the matrix had only maximally 0,08% nonzero elements. This is by itself of course not harmful, but recalling that the complexity of the SVD process constrains the overall size of the matrix, a different modelling seems more reasonable. Again, the approaches of Lund and Burgess (1996) and Rapp (2003) may give the answer: A term-by-term-matrix[6] can model the same amount of text in a smaller and less sparse format. Using this kind of matrix, much larger corpora can be used for the analysis; Rapp (2003) was able to analyse the full *British National Corpus* comprising more than 100 million words.

With respect to the results obtained from a collocation analysis of the same corpus, the LSA results do not show big differences. In general, they are significantly better, but none of the classes differs more than 4,5% from the CA. This is surprising, since a technique like CA relies on co-occurrence information only and does not make use of complex matrix calculations.

Still, the two analyses seem to show a different behaviour with regard to ambiguous words. Whereas CA finds for a given ambiguous test word neighbours from both conceptual domains, LSA seems to mask out one of the meanings.

We hope to have given a deeper understanding of what LSA can and cannot do. Regarding our results, it is not much more semantic than a simple technique like CA, and some of its modelling aspects, such as the optimal context size or the kind of co-occurrence matrix still leave space for improvement and further research. Particularly the effects of the *SVD* on word similarity should be further investigated, before LSA is used as a general tool to derive semantic relations from text.

References

CRUSE, D.A. (1986), *Lexical Semantics*, Cambridge University Press, Cambridge.

DEERWESTER, S. C., DUMAIS, S.T., LANDAUER, T.K., FURNAS, G.W., HARSHMAN, R.A. (1990), "Indexing by Latent Semantic Analysis", *Journal of the American Society of Information Science*, 41 (6), pp. 391 – 407.

DUMAIS, S. (1990), *Enhancing Performance in Latent Semantic Indexing*, Technical Report TM-ARH-017527, Bellcore.

[6] A term-by-term matrix for a vocabulary V is of the size $|V|^2$ and reflects the frequency of co-occurrence of two terms within a certain text window (e.g. ± 5 words).

DUMAIS, S. T. (1995), "Latent Semantic Indexing (LSI): TREC-3 Report", in D. Harman (Ed.), *Proceedings of the 3rd Text REtrieval Conference (TREC-3)*, Vol. 500-226, pp. 219-230, NIST Special Publication.

DUNNING, T. (1993), "Accurate Methods for the Statistics of Surprise and Coincidence", *Computational Linguistics* **19(1)**, pp. 61-74.

LANDAUER, T. K. und DUMAIS, S. T. (1997), "A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge", *Psychological Review* **104**, pp. 211-240.

LANDAUER, T. K., LAHAM, D., REHDER, B., und SCHREINER, M. E. (1997), "How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans", in M. G. Shafto und P. Langley (Eds.), *Proceedings of the 19th annual meeting of the Cognitive Science Society*, pp. 412-417, Erlbaum, Mawhwah, NJ.

LANDAUER, T. K., FOLTZ, P. W., und LAHAM, D. (1998), "Introduction to Latent Semantic Analysis", *Discourse Processes* **25**, pp. 259-284.

LUND, K. and BURGESS, C. (1996), "Producing high-dimensional semantic spaces from lexical co-occurrence", *Behaviour Research Methods, Instruments and Computers* **28**(2), pp. 159-165.

QUASTHOFF, U. (1998), „Deutscher Wortschatz im Internet", in *Proceedings des LDV-Forum 2/98*.

QUASTHOFF, U. und WOLFF, C. (2002), "The Poisson Collocation Measure and its Applications", in *Proceedings of the 2nd International Workshop on Computational Approaches to Collocations*, Wien.

RAPP, R. (2002), "The Computation of Word Associations: Comparing Syntagmatic and Paradigmatic Approaches", *Proceedings of the COLING 02*, Taipei.

RAPP, R. (2003), "Word Sense Discovery Based on Sense Descriptor Dissimilarity", *Proceedings of the 9th Machine Translation Summit*, New Orleans.

SALTON, G. und MCGILL, M. (1983), *Introduction to Modern Information Retrieval*, McGraw-Hill, New York.

SCHILLER, A. (1995), *DMOR: Benutzeranleitung*, Technical report, IMS Stuttgart, Draft.

SPÄRCK-JONES, K. (1972), "A statistical interpretation of term specificity and its application in retrieval", *Journal of Documentation* **28(1)**, pp. 11-20.

WADE-STEIN, D. und KINTSCH, E. (2003), *Summary Street: Interactive Computer Support for Writing*, Technical report, University of Colorado.

WIEMER-HASTINGS, P. (1999) "How latent is Latent Semantic Analysis?", *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, Stockholm, Sweden, Aug 1999, pp. 932-937. San Francisco: Morgan Kaufmann.

Quels types de connaissance sémantique pour Questions-Réponses ?

Vincent Barbier
LIMSI, UPR CNRS 3251, Bât. 508
Université de Paris XI, 91403 Orsay
barbier@limsi.fr

Mots-clefs – Keywords

Système de Question Réponse, ressources sémantiques, évaluation des reformulations
Question Answering system, semantic resources, evaluation of reformulations

Résumé - Abstract

Les systèmes de Questions Réponse ont besoin de connaissances sémantiques pour trouver dans les documents des termes susceptibles d'être des reformulations des termes de la question. Cependant, l'utilisation de ressources sémantiques peut apporter un bruit important et altérer la précision du système. ne fournit qu'une partie des reformulations possibles.

Cet article présente un cadre d'évaluation pour les ressources sémantiques dans les systèmes de question-réponse. Il décrit la fabrication semi-automatique d'un corpus de questions et de réponses destiné à étudier les reformulations présentes entre termes de la question et termes de la réponse. Il étudie la fréquence et la fiabilité des reformulations extraites de l'ontologie WordNet.

Question Answering systems need semantic knowledge to find in the documents terms that are reformulations of the question's terms. However, the use of semantic resources brings an important noise and a system's precision might get depreciated.

This article presents a framework for evaluating semantic resources in question-answering systems. It describes the semi-automated construction of a corpus containing questions and answers. This corpus can be used to study the various reformulations between the terms of the questions and the terms of the answers. This article studies the frequency and reliability of reformulations given by the WordNet ontology.

1 Introduction

La problématique questions-réponses (QR) consiste à concevoir un programme qui reçoit de l'utilisateur une question formulée en langue naturelle et recherche dans une collection de documents un fragment de texte, typiquement un court passage, répondant à cette question. Les campagnes d'évaluation TREC, dont l'enjeu est de répondre à des questions factuelles par de courts passages d'articles de journaux ont encouragé la réalisation de nombreux systèmes de QR et donné une évaluation globale de ces systèmes.

La plupart des systèmes sont constitués de trois étapes principales (Grau, 2004). Tout d'abord, l'analyse de la question qui permet de découper la question en termes et de trouver le type attendu de la réponse. Ensuite, la recherche des documents dans la collection, au moyen d'un moteur de recherche. On effectue pour cela une ou plusieurs requêtes successives composées des termes de la question et éventuellement de reformulations de ceux-ci. Enfin, la sélection des réponses mettant en jeu des critères de pertinence faisant intervenir la syntaxe et la sémantique.

Pour trouver des documents pertinents, les systèmes ont besoin de reconnaître dans les documents des reformulations de termes de la question. Par exemple, dans la collection d'articles de journaux Aquaint (TREC11), la question "What is the name of the volcano that destroyed Pompeii ?" , ne possède pas de réponse réutilisant le verbe *to destroy*. Par contre, des équivalents comme *to devastate*, *to bury under ashes*... permettent de retrouver la réponse.

Les reformulations peuvent intervenir à deux endroits. Premièrement, au niveau de la requête, pour permettre de retrouver les documents qui n'utilisent pas la même formulation de la question, c'est-à-dire, augmenter le rappel de l'étape de recherche des documents. Ensuite au niveau de la sélection des documents, des passages ou des phrases retournés, afin d'améliorer leur classement et sélectionner ceux dont la formulation approche le mieux possible celle de la question. Plus les reformulations sont complexes et plus elles sont susceptibles d'engendrer du bruit, c'est-à-dire d'amener le système à juger pertinents des documents ou passages qui ne le sont pas. Les systèmes de questions-réponses doivent donc utiliser les ressources linguistiques avec précaution, et pour cela, doivent estimer la fiabilité des reformulations utilisées.

Notre but est d'étudier le besoin de reformulations des systèmes de questions réponses, de chercher quels types de reformulations seraient utiles à ces systèmes et quel est le degré de fiabilité de ces reformulations. Cette étude se fonde sur un corpus de questions réponses constitué automatiquement. Après un état de l'art des sytèmes de questions réponses et de la façon dont ils utilisent les ressources sémantique, nous décrirons la construction du corpus, puis nous étudierons les reformulations obtenues grâce à l'ontologie WordNet, et tenterons d'évaluer leur fiabilité.

2 La sémantique dans les travaux actuels

Lors du filtrage des documents ou des passages, les systèmes utilisent différents critères pour évaluer la présence d'une réponse. Ceux-ci se répartissent en critères syntagmatiques et paradigmatiques. Les critères paradigmatiques sont ceux qui déterminent une proximité entre un élément de la question et un élément de la réponse. Les critères syntagmatiques sont ceux qui tiennent compte de la disposition de ces éléments dans les entités textuelles où ils apparaissent, ou de leurs dispositions relatives. Les systèmes allient ces deux types de contraintes sous

diverses formes.

Des techniques les plus restrictives aux plus souples, la contrainte syntagmatique peut être implémentée par :

- L'utilisation de patrons d'extraction de la réponse, par exemple une date de naissance peut être repérée par "X was born in <date-naissance>" ou "X (<date-naissance> - <date-décès>) ". (Jijkoun & de Rijke, 2004) montre que cette méthode apporte des résultats pour des questions simples (TREC8). Elle est de plus bien adaptée à des collections de documents de taille très importante, comme celle rendue disponible par la toile.

- Une preuve logique visant à démontrer une implication possible entre deux graphes, représentant respectivement la question et un passage candidat à contenir la réponse (Moldovan *et al.*, 2002). Cette technique consiste à écrire la question et un passage candidat sous forme logique. Par exemple "How did Hitler die?" est retranscrit par :

  manner_at(e) & hitler_nn(x) & die_vb(e,x,y).

  La preuve peut utiliser des axiomes linguistiques et semantiques. Un axiome sémantique, tiré de WordNet, pourrait être "*to kill* cause *to die*" ce qui a été retranscrit par :

  kill(e1,x1,x2) -> die(e1,x2,$c) ou encore - kill(e1,x1,x2) | die(e1,x2,$c)

  où $c est un argument indéterminé, et '-' l'opérateur de négation.

- Une mesure de la proximité spatiale des éléments de la question dans le passage candidat.

En ce qui concerne la contrainte paradigmatique, le but est d'associer les élément de la question à des éléments correspondants dans la réponse. Il peut s'agir de repérer dans le passage candidat un terme correspondant au type attendu de la réponse (entités nommées trouvées dans une taxonomie ou concept d'une ontologie), ou de repérer des reformulations sémantiques des termes de la question.

La mise en correspondance des termes de la question et de la réponse nécessite des ressources. Le type de ressource majoritairement utilisé sont les dictionnaires de synonymes (Magnini *et al.*, 2002; Ferret *et al.*, 2001). Mais il est possible d'utiliser les différentes relations d'une ontologie telle que WordNet (Moldovan *et al.*, 2002). L'ontologie WordNet (Fellbaum, 1998) est intéressante pour sa couverture générale de la langue. C'est sur elle que s'axera en premier lieu notre travail.

WordNet est un lexique dont les entrées, les synsets, sont assimilables à des concepts. Les catégories grammaticales traitées sont les noms, les verbes, les adjectifs et les adverbes. La plupart des travaux concernant WordNet se concentrent sur l'ontologie des noms. Les entrées de WordNet sont décrites par leurs connexions avec les autres entrées du lexique. Ces liens sont de plusieurs types : l'hyponymie, qui relie un concept général à des concepts plus spécifiques, et son inverse, l'hypéronymie. Cette paire de relation constitue la principale structuration de l'ontologie noms. Les noms sont aussi organisés selon la relation d'holonymie qui relie un élément à un ensemble le contenant (par exemple *argentin* est un méronyme de *Argentine*). Les relations structurant les verbes sont le couple hyponymie/hypéronymie, ainsi que les relations de cause (ex : tuer cause le fait de mourir) et d'implication (ex : ronfler entraîne que l'on dort). Les adjectifs, peu structurés, sont définis par leurs synonymes et antonymes, et surtout par les noms auxquels ils se rapportent.

A la structuration des liens, il faut ajouter les définitions (glosses), qui peuvent elles aussi apporter des liens utiles qui ne seraient pas codés explicitement.

LCC fait un usage important de ces liens grâce à une version étendue de WordNet où les définitions sont codées sous forme de graphes et les mots de ces définitions ont été désambiguïsés manuellement, c'est-à-dire rattachés au synset leur correspondant le mieux.

Cependant, les relations codées dans WordNet ne sont qu'une partie des relations existantes. Par exemple, les liens de WordNet donnent principalement des relations entre mots d'une même catégorie grammaticale. Or certaines reformulations ne s'effectuent pas mot à mot mais sur une combinaison de deux ou plusieurs mots (Jacquemin, 1999). Certaines constituent une transformation complète de la phrase. Ces reformulations complexes peuvent remplacer un mot par un mot d'une catégorie grammaticale différente. Par exemple la question :

"What is the democratic party *symbol*?"

possède une réponse de la forme :

"The democratic party was *represented* as a donkey."

Pour remédier aux manques de WordNet, on peut utiliser les relations morphologiques qui relient par exemple le verbe *to make* au nom *maker*. Cependant ces relations morphologiques couvrent le lexique de façon lacunaire et hasardeuse. Ainsi, en anglais, les termes décrivant les relations conjugales : "husband, wife, married" ne sont pas morphologiquement reliés. Pour combler ces manques, (Claveau & Sebillot, 2004; de Chalendar G., 2001) acquièrent automatiquement des relations nom-verbe sur corpus.

Plusieurs sources de connaissances sémantiques sont donc disponibles, synonymes, ontologies, ou relations apprises sur corpus. L'architecture générale des systèmes de QR est définie et relativement homogène d'un système à l'autre. Dans ce cadre, il est maintenant nécessaire d'examiner plus en détail l'apport des différentes étapes, ainsi que d'évaluer l'apport des ressources linguistiques utilisées.

3 Construction du corpus

Afin d'étudier l'utilité et d'évaluer la fiabilité des reformulations, nous sommes passés par la construction d'un corpus. Nous voulons pouvoir utiliser ce corpus pour étudier, pour chaque terme d'une question, les termes qui lui correspondent dans les réponses. Ce corpus pourra aussi servir pour apprendre automatiquement des mesures de proximités sémantiques.

Actuellement, ce corpus est fondé sur 54 questions de la campagne d'évaluation TREC11. Nous l'étendrons à plus de questions. La collection de documents utilisée est la collection Aquaint fournie pour cette campagne.

A chaque question est associé un ensemble de documents pertinents et non pertinents. Un passage est jugé pertinent s'il contient la réponse à la question, laquelle est repérée par une expression régulière, telle que "John(F.)? Kerry", qui indique que les caractères " F." sont facultatifs. Cette méthode laisse passer du bruit dans la liste de documents pertinents, mais elle permet d'automatiser la sélection de ces documents.

3.1 Eviter les biais

Pour la constitution du corpus, il est nécessaire d'effectuer une requête sur un ensemble de documents en évitant que cette requête n'introduise un biais par rapport au sujet de l'étude. Dans notre cas l'étude porte sur les reformulations des mots de la question, et notamment, les reformulations obtenues en suivant les liens proposés par WordNet. Son but est de mesurer la fiabilité de certaines reformulations par rapport à d'autres, ou plus précisément de certaines caractéristiques de ces reformulations.

Nous avons opté pour l'utilisation de requêtes booléennes plutôt que vectorielles car ces premières permettent de mieux contrôler le contenu des documents rapportés.

Or, si la requête se présente sous la forme d'une conjonction de termes, ces termes apparaîtront nécessairement dans les documents pertinents et les documents non pertinents. Ceci rendrait inutilisable une mesure fondée sur la comparaison des fréquences d'un terme entre documents pertinents et non pertinents.

On peut tenter de pallier ce problème en présentant les requêtes sous la forme d'une conjonction de disjonctions où chaque terme de la question est remplacé dans la requête par une disjonction de variantes de ce terme. Cependant, les variantes choisies sont favorisées par rapport aux autres. De plus, si ces variantes sont extraites d'une ressource telle que WordNet, les reformulations apparaissant dans les documents collectés seront principalement celles de cette ressource, ce qui réduirait la portée de l'étude à un type de relations très particulières.

3.2 Requête et filtrage

La méthode que nous avons retenue est la création à partir des termes de la question d'une "requête à trou", c'est-à-dire une requête où au moins un des termes ne sera pas représenté. Le ou les termes omis seront l'objet de l'étude. Les autres termes sont reformulés grâce WordNet. Les variantes autorisées sont les synonymes des termes de la question, ainsi que les mots le plus courant des synsets situés à un lien de distance. L'étendue des reformulations est volontairement réduite afin que la requête engendrée ne rapporte pas trop de documents non pertinents. Une plus grande expansion de la requête augmenterait le bruit qui est déjà important (CF Table 2). Enfin, les termes les moins significatifs de la question ne sont pas utilisés dans les requêtes. Cette significativité est estimée par un expert humain qui classe les termes de la question par ordre décroissant. Par exemple, les termes de la question "What is the name of the volcano that destroyed the ancient city of Pompeii?" ont été classés dans l'ordre suivant :

Pompeii > volcano > destroy > ancient > city.

De manière systématique, les entités nommées sont considérés comme de meilleurs filtres que les noms communs et les autres catégories grammaticales décrites dans WordNet.

Cet ordre permet ensuite de créer des requêtes adéquates de façon automatique. Par exemple, si l'on veut étudier les reformulations possibles du terme "destroy", on utilisera pour la requête :

Pompeii & expansion(volcano) & expansion(ancient)
avec expansion(volcano) = (mountain | mount | crater | volcano)

Une fois les documents rapportés, on sépare les documents pertinents des non pertinents. Il faut noter que par cette méthode, les documents pertinents sont obtenus par une même requête, ce qui réduit les risques de biais.

Certains patrons de réponses sont des filtres très imprécis, par exemple une date ou un personnage souvent cité dans le corpus pour des raisons variées. Dans certains cas, la question contient des termes assez précis - par exemple, une entité nommée - pour permettre un filtrage efficace des documents. Mais si ce n'est pas le cas, il peut être nécessaire de rendre la requête plus précise en ajoutant des termes ou en réduisant l'étendue des reformulations. Dans le cas extrême, on pourra supprimer une question de l'étude.

3.3 Résultat

Le corpus obtenu contient 5952 documents jugés pertinents par le programme et 65419 documents jugés non-pertinents. Il a été testé manuellement pour 16 questions contenant 114 termes. Nous avons obtenu 295 documents pertinents, et 156 erreurs, ce qui correspond à un bruit de 35%. Un nettoyage manuel du corpus semble donc nécessaire.

patron +	patron -
5952	65419

Table1: taille du corpus brut.

	patron +	patron -
validé +	294(65%)	-
validé -	156(35%)	-

Table2: qualité du corpus validé.

Dans le sous-corpus validé, le nombre moyen de documents pertinents rapportés par question est de 10,6. Ce nombre est très variable d'une question à l'autre. Les extrêmes valent 1 et 43. Dans les deux quartiles centraux, ce nombre évolue de 3 à 11,5. Cette variabilité s'explique par le fait que certaines informations sont répétées (vie privée de stars, événements sportifs marquants) alors que d'autres sont très peu cités. D'autre part, on doit noter que les filtrages effectués pour réduire le bruit du corpus font nécessairement disparaître certains documents pertinents.

4 Quelle sémantique pour les systèmes de questions réponses

4.1 Les types de reformulations présentes.

A présent, analysons les types de reformulations présentes dans le corpus. Ceci peut s'effectuer manuellement en recensant et examinant les liens, ou de façon plus automatique, en mesurant pour différentes caractéristiques des liens, les fréquences d'apparition dans les phrases contenant la réponse, comparées aux fréquences d'apparition dans le reste du corpus.

4.1.1 Etude manuelle des liens

Nous avons examiné les liens trouvés par le programme, et nous avons noté ceux qui semblaient pertinents. Pour qu'un lien nous semble pertinent il faut qu'un trait sémantique se transmette d'un bout à l'autre de la chaîne. On peut avoir plusieurs niveaux de pertinence, selon que les traits sémantiques partagés sont plus ou moins spécifiques. Par exemple, le lien

volcano-(hyponyme)->Krakatau-(holonyme)->Indonésie-(hypéronyme)->country-(hyponyme)-> Philipines

possède une certaine pertinence, car il relie des noms géographiques, mais celle-ci est très faible.

Dans le corpus de test, on distingue deux catégories de liens intéressants : des liens fortement pertinents, pointant vers des mots susceptibles de faire partie de la réponse. Et d'autres plus distants, pointant vers des mots qui forment le thème général de l'article. Dans un article sur l'éruption du Vésuve on trouve des mots comme *earth*, *Naples*, *Campania*, qui viennent renforcer les champs thématiques de la question.

Cette observation, à vérifier sur un corpus plus important, nous indique que des relations assez distantes pourront peut-être aider au filtrage de la question.

Différentes catégories de lien Voici une courte description des types de liens trouvés dans le corpus. Aussi la majorité des liens trouvés concernent les noms. Les relations entre verbes sont moins fréquentes. Les relations liées aux adjectifs sont rarement pertinentes.

Nous notons une part importante de liens triviaux. Il s'agit de la reprise du mot de la question tel quel ou sous une forme fléchie. Les reformulations par synonyme sont également présentes. Pour certains mots, les reformulations morphologiques, permettant de changer de catégorie grammaticale sont importantes : par exemple pour la question

> "When did the *shooting* in Columbine happen?"
>
> Les reformulations les plus trouvées sont les verbes *to kill* et *to shoot* (260 occurences dans le corpus) devant le nom *shooting* (50 occurences).

Les relations d'hyponymie-hypéronymie et de méronymie-holonymie fournissent toutes deux des liens pertinents. Cette prépondérance des relations entre noms s'explique par le fait que les noms sont la catégorie grammaticale la plus présente dans la question, et que le réseau des noms contient plus de liens que celui des verbes.

Nous remarquons que les reformulations se font parfois par une succession d'une seule relation, par exemple, une succession d'hypéronymes :

> *political_party -(hypéronyme)-> organization -(hypéronyme)-> social_group -(hypéronyme)-> group*

Mais des reformulations de ce type sont rares par rapport à celles reliant deux concepts frères, c'est-à-dire deux concepts pointant vers un même concept par une même relation. La relation empruntée est le plus souvent celle d'hypéronymie-hyponymie, mais on peut aussi trouver les relations d'holonymie-méronymie et d'implication comme dans les exemples ci-dessous.

Relation de co-méronymie correspondant à une proximité géographique :

> *Vesuvius-(holonyme)->Italy<-(holonyme)-Pompeii*

Verbes reliés par une relation de co-implication :

> *destroy -(hypéronyme)-> smother -(implique)-> cover <-(implique)- entomb*

De nombreuses relations de co-hyponymie

> *score -(hypéronyme)-> achieve -(hypéronyme)-> succeed <-(hypéronyme)- hit*
>
> *city -(hypéronyme)-> municipality <-(hypéronyme)- town*

Ces observations sont encourageantes et une analyse détaillée sur un corpus plus fourni permettra d'y puiser les éléments d'une heuristique.

4.2 Fiabilité des reformulations.

4.2.1 Protocole d'évaluation d'une heuristique

Caractéristiques et heuristiques Les liens de WordNet permettent de constituer des chemins plus où moins longs pour relier des mots sensés posséder un lien sémantique. Cependant, plus on s'éloigne et plus on risque d'aboutir à un mot non pertinent pour la question posée. Afin de réduire le bruit engendré, on peut utiliser des heuristiques fondées sur les caractéristiques de la géométrie des chemins parcourus (Budanitsky, 1999).

Le corpus peut servir à évaluer la fiabilité d'heuristiques complexes ou de caractéristiques élémentaires des chemins, telle que leur longueur ou la présence d'une relation particulière. Ces mesures élémentaires peuvent permettre de justifier ou invalider des heuristiques plus complexes.

Nous décrivons ci-après un cadre commun pour l'évaluation d'heuristiques complexes et de caractéristiques unitaires.

Evaluer la fiabilité des caractéristiques On peut représenter heuristiques et caractéristiques sous la forme d'une fonction qui à chaque élément d'un ensemble (ici, l'ensemble des chemins) associe une valeur, soit booléenne, soit réelle.

Nous avons pour l'instant considéré des caractéristiques booléennes telles que *"est de longueur N"* ou *"Contient la relation R"*. On calcule parmi les chemins vérifiant cette caractéristique, combien sont "pertinents", c'est à dire pointent vers un document pertinent. On calcule la précision de cette caractéristique dans le corpus qui vaut :

$$\frac{Nb\ de\ liens\ pertinents\ verifiant\ la\ caracteristique}{Nb\ total\ de\ liens\ verifiant\ lacarcteristique}$$

La valeur de la précision calculée dépend du ratio entre nombre de documents pertinents et non-pertinents. Elle n'a donc pas de réalité absolue mais permet d'effectuer des comparaisons entre deux caractéristiques ou heuristiques obtenues par le même protocole.

Estimer la significativité des mesures Nous appliquerons à notre protocole une méthode de validation par rééchantillonnement utilisée par exemple dans (Voorhees, 2002) pour la campagne d'évaluation TREC11. Cette méthode permettra d'estimer la variance des mesures effectuées et de savoir si elles sont fiables. Pour l'instant nous n'avons pas effectué cette étape de validation.

4.2.2 Les caractéristiques testées

Nous avons commencé par tester des caractéristiques élémentaires des chemins.

Longueur La première est la longueur. Le résultat est prévisible. En effet, on s'attend à ce que plus un chemin est de longueur courte et plus il est fiable que le protocole d'évaluation.

longueur	0	1	2	3	4
précision(%)	17*	40	20	10	5

Table3: fiabilité des reformulations en fonction de la distance dans WordNet.

* Présence d'un biais dans le corpus pour les reformulations de longueur nulle.

Ce résultat est très intéressant car il montre bien que la fiabilité d'un lien décroît avec sa longueur.

Effets de la combinaison de plusieurs relations Il semble que les relations apparaissent préférentiellement accompagnées de leurs relation opposée, par exemple, la relation d'hypéronymie accompagnée d'une ou plusieurs relation d'hyponymie. Nous voulons vérifier si certaines relations "vont bien ensemble" en termes de fiabilité de la reformulation obtenue, ou si d'autres assortiments ont des effets négatifs. Ce problème revient à mesurer par le protocole que nous avons présenté la fiabilité des deux caractéristiques booléennes suivantes :

Cooccurrence exclusive : vraie s'il y a présence dans un chemin de deux types de relation à l'exclusion de tout autre type

Cooccurrence non exclusive : vraie en cas de présence dans un chemin de ces deux types de relation qu'il y ait ou non d'autres types de relations présents.

Nous nous sommes limité aux cooccurrences non exclusives de 2 relations :

	hype	hypo
hype	6,4%	7,8%
hypo	7,8%	4,9%

	mero	holo
mero	3.6%	6,3%
holo	6,3%	*40%

*non significatif

Table4: fiabilité des reformulations selon les liens qui les composent.

La cooccurrence d'un type de relation avec lui-même n'est vérifiée que si cette relation est présente deux fois dans le chemin.

En ce qui concerne les cooccurences d'hyponymies et d'hypéronymies, on constate que les chemins contenant uniquement l'une ou l'autre relation sont moins fiables que ceux contenant les deux. Une interpretation possible de ce résultat est que les chemins qui ne sont constitués que de l'une de ces deux relations causent un changement du niveau de généralité entre un terme et sa reformulation. Au contraire une reformulation mixte a plutôt tendance à conserver le niveau de généralité.

Ces résultats rejoignent les observations sur le corpus qui soulignaient la présence de reformulations mettant en relation des concepts "frères".

4.3 Utilité d'une catégorie de reformulation

Notre travail présente une méthode pour évaluer la fiabilité de différentes catégories de reformulations. Le problème suivant est d'évaluer l'utilité de ces catégories de reformulations. Il faut pour cette étude définir ce qu'est l'utilité d'une catégorie.

Une première piste est de considérer qu'une variante d'un terme est utile si le fait de la prendre en compte permet de trouver au moins un document pertinent alors qu'ignorer cette variante n'aurait pas permis d'en trouver. Le recensement de ces variantes utiles permettra d'évaluer dans quelle mesures des catégories de variantes sont utiles à la recherche des documents pertinents.

Conclusion

Cette étude montre la possibilité d'utiliser un corpus de questions réponses constitué semi-automatiquement pour étudier les reformulations dans le cadre des systèmes de questions réponses.

Les résultats obtenus peuvent nous guider dans le choix de reformulations fiables. Ils semblent notamment indiquer l'utilité et la fiabilité de reformulations par mots "frères". Un autre facteur important pour juger de l'utilité d'un type de reformulations est la fréquence avec laquelle celle-ci peut s'utiliser. Ces considérations nécessitent une étude de fréquence sur corpus que nous effectuerons dans nos travaux futurs.

Notre travail a porté sur WordNet car c'est une ressource de large couverture lexicale et fortement structurée. Cependant, le corpus produit n'est pas biaisé en faveur de WordNet, et permettra par conséquent l'étude d'autres ressources sémantiques. Une perspective de mon travail est d'étudier l'utilité de ressources sémantiques constituées de liens nom-verbe et acquises automatiquement sur corpus par (de Chalendar G., 2001).

Références

BUDANITSKY A. (1999). *Lexical semantic relatedness and its application in natural language processing*. Rapport interne CSRG-390, University of Toronto, Department of Computer Science.

CLAVEAU V. & SEBILLOT P. (2004). Extension de requêtes par lien sémantique nom-verbe acquis sur corpus. In *Actes de TALN 2004*.

DE CHALENDAR G. (2001). *SVETLAN', un système de structuration du lexique guidé par la détermination automatique du contexte thématique*. PhD thesis, Université Paris XI.

FELLBAUM C. (1998). *WordNet, an electronic lexical database*. Cambridge, MA: The MIT Press.

FERRET O., GRAU B., HURAULT-PLANTET M., ILLOUZ G. & JACQUEMIN C. (2001). Utilisation des entités nommées et des variantes terminologiques dans un système de question-réponse. In *Actes de TALN 2001*.

GRAU B. (2004). Les systèmes de question réponse. In M. IHADJADENE, Ed., *Méthodes avancées pour les systèmes de recherche d'informations*. Paris: Hermes Sciences.

JACQUEMIN C. (1999). Syntagmatic and paradigmatic representation of term variation. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, p. 341–348.

JIJKOUN V. & DE RIJKE M. (2004). Information extraction for question answering : Improving recall through syntactic patterns. In *Proceedings of ACL 2004*.

MAGNINI B., NEGRI M., PREVETE R. & TANEV H. (2002). Mining knowledge from repeated co-occurences: Diogene at trec 2002. In *Proceedings of the 11th Text REtrieval Conference, (TREC11)*.

MOLDOVAN D., HARABAGIU S., GIRJU R., MORARESCU P., LACATUSU F., NOVISCHI A., BADULESCU A. & BOLOHAN O. (2002). Lcc tools for question answering. In *Proceedings of the 11th Text REtrieval Conference (TREC 2002)*.

VOORHEES E. (2002). Overall view of the question answering track. in special publication. In *The 11th Text REtrieval Conference (TREC 2002)*.

Une plate-forme logicielle dédiée à la cartographie thématique de corpus

Thibault ROY
Laboratoire GREYC, Equipe ISLanD - Université de Caen / Basse-Normandie
Campus II - Côte de Nacre - Bd Maréchal Juin - 14032 Caen Cedex
troy@info.unicaen.fr
date de soutenance prévue fin 2007

Mots-clefs – Keywords

cartographie de corpus, analyse thématique, logiciel individu-centré, analyse des données textuelles
corpora cartography, thematic analysis, user-centered software, textual data analysis

Résumé - Abstract

Cet article présente les principes de fonctionnement et les intérêts d'une plate-forme logicielle centrée sur un utilisateur ou un groupe d'utilisateurs et dédiée à la visualisation de propriétés thématiques d'ensembles de documents électroniques. Cette plate-forme, appelée ProxiDocs, permet de dresser des représentations graphiques (des cartes) d'un ensemble de textes à partir de thèmes choisis et définis par un utilisateur ou un groupe d'utilisateurs. Ces cartes sont interactives et permettent de visualiser les proximités et les différences thématiques entre textes composant le corpus étudié. Selon le type d'analyse souhaitée par l'utilisateur, ces cartes peuvent également s'animer afin de représenter les changements thématiques d'un ensemble de textes au fil du temps.

This article presents a user-centered software dedicated to the visualization of thematic properties of sets of electronic documents. This software, called ProxiDocs, allows its users to realize thematic maps from a corpora and themes they choose and defined. These maps are interactive and reveal thematic proximities and differences between texts composing the studied corpus. According to the analysis wished by the user, maps can be animated in order to represent thematic changes of the analysed set of texts relating to the time.

1 Introduction

Cet article présente les principes de fonctionnement et les intérêts de la plate-forme logicielle ProxiDocs dédiée à des analyses thématiques de corpus de textes. Sur le modèle de (Pichon et Sébillot, 1999), nous entendons par *thèmes*, les sujets abordés dans un texte. Traiter la thématique d'un texte revient donc pour nous à mettre en évidence les principaux sujets abordés dans ce dernier. Dans un grand nombre de situations (extraction et recherche d'information, analyse de flux documentaires, etc.), l'appréhension des thèmes abordés dans des textes constitue une première analyse importante et délicate.

ProxiDocs a pour objectif d'aider ses utilisateurs dans de telles situations en leur fournissant des représentations graphiques (que nous appelons des *cartes thématiques*) d'un corpus de textes donné. Les cartes construites mettent en évidence la répartition des thèmes au sein des textes du corpus et révèlent des proximités et des différences de thèmes entre textes. Ces cartes sont construites à partir de thèmes choisis et définis par l'utilisateur en fonction de la tâche qu'il souhaite accomplir. En ce sens, c'est un système *anthropocentré* tel que le définit (Thlivitis, 1998) : son exécution n'est pas guidée par des ressources propres, les traitements réalisés sont personnalisés et intégralement conditionnés par les besoins et les choix de l'utilisateur.

La plate-forme ProxiDocs est *open-source*, développée en Java et disponible avec sa documentation sur le Web[1]. C'est un système développé à la façon d'un logiciel d'étude au sens de (Nicolle, 1996), c'est-à-dire qu'il est conçu dans le but de vérifier des hypothèses sur les langues en les expérimentant sur du matériau textuel attesté. ProxiDocs fait partie d'un ensemble de logiciels d'étude en constante évolution dédiée à l'analyse linguistique informatisée de corpus de documents électroniques[2] développés au sein de l'équipe ISLanD du laboratoire GREYC.

Dans cet article, nous présentons tout d'abord des outils utilisant des techniques de cartographie afin d'accéder aux informations contenues dans des collections de documents. Ensuite, nous abordons tout particulièrement la plate-forme ProxiDocs en détaillant ses principes de fonctionnement et en présentant les différentes types de cartes qu'elle permet de construire.

2 La cartographie d'ensembles de documents textuels

Le nombre de documents textuels produits et échangés chaque jour ne cesse de croître. Afin d'isoler les principales informations contenues dans des ensembles de documents textuels, il peut être intéressant de proposer des représentations graphiques de ces ensembles. De telles représentations permettent de faire intervenir une notion de proximité entre éléments. Selon le type d'analyse réalisée, il est alors possible d'observer des similarités et des différences de styles, de thèmes, de mises en forme entre documents d'un même ensemble.

Depuis quelques années, des outils d'analyse textuelle exploitent une technique de visualisation appelée *cartographie*. À la manière d'une carte routière mettant en évidence des villes et des routes les reliant, une carte d'un ensemble de données textuelles met en évidence des proximités et des liens entre entités textuelles, tels des mots, des textes, etc.

Dans une tâche d'extraction d'information, les auteurs de (Mokrane et al., 2004) propose d'utiliser une technique de cartographie afin de visualiser les liens entre les principaux termes présents

[1] http://www.info.unicaen.fr/~troy/proxidocs

[2] http://www.greyc.unicaen.fr/island/logiciel

dans un ensemble de dépêches d'agences de presse.

Depuis 2001, les deux métamoteurs de recherche cartographiques KartOO (Chung et al., 2002) et MapStan (Spinat, 2002) sont disponibles sur le Web[3]. En réponse à une requête de l'utilisateur, ces deux outils retournent des cartes représentant les sites proposés en réponse à cette requête. Les sites jugés similaires par le système sont alors situés à proximité sur les cartes et il est ainsi possible de distinguer les grandes catégories d'informations proposées en réponse à la requête de l'utilisateur.

Pour une tâche de parcours rapide d'un ensemble documentaire, le logiciel NeuroNav (Lelu et Aubin, 2001) de la société Diatopie[4] présente sur une carte des groupes de documents. Les différents groupes déterminés par le système et la disposition de ces groupes sur la carte peuvent ainsi indiquer des proximités de contenu entre documents et groupes représentés.

De nombreux logiciels dédiés à l'analyse de données textuelles proposent également des résultats d'analyses sous forme de cartes. Parmi ces logiciels, nous pouvons citer Hyperbase d'Etienne Brunet[5], Lexico3 de l'équipe CLA2T de Paris III[6] ou encore Lexica de la société Le Sphinx[7].

À la manière de la plupart des outils précédents, la plate-forme que nous proposons va utiliser une technique de cartographie afin de mettre en évidence des proximités et des liens entre textes d'un même ensemble. Contrairement à ces outils, les traitements réalisés par ProxiDocs prennent en considération les particularités de l'utilisateur et de sa tâche. À partir de ressources décrivant le point de vue de l'utilisateur sur des thèmes de son choix, ProxiDocs cherche à extraire les tendances thématiques des textes d'un corpus. Ces tendances sont ensuite mises en évidence sur des cartes.

3 La plate-forme logicielle ProxiDocs

3.1 Définitions et intérêts

La plate-forme ProxiDocs permet de construire différents types de cartes thématiques à partir d'un corpus de textes et de thèmes choisis et définis par l'utilisateur :

- des cartes en 2 ou 3 dimensions représentant chaque texte du corpus analysé par un point, les proximités entre points indiquent des similarités de thèmes abordés dans les textes représentés (figure 2) ;
- des cartes en 2 ou 3 dimensions mettant en évidence des groupes de textes abordant des thèmes proches, chaque groupe est alors représenté sur la carte par un disque ou une sphère de diamètre proportionnelle à nombre de textes qu'il contient (figure 3) ;
- et des cartes en 2 dimensions animées mettant en évidence l'évolution des thèmes abordés dans les textes du corpus au fil du temps (figure 4).

Les intérêts de ProxiDocs sont multiples. Les cartes thématiques construites par la plate-forme permettent d'observer des regroupements entre textes abordant des thèmes proches. De tels

[3] Respectivement disponibles aux adresses : `http://www.kartoo.fr` et `http://www.mapstan.net`
[4] `http://www.diatopie.com/`
[5] `http://ancilla.unice.fr/~brunet/pub/hyperbase.html`
[6] `http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicowww/lexico3.htm`
[7] `http://www.lesphinx-developpement.fr/`

regroupements sont très utiles dans des tâches de parcours rapide d'un grand ensemble de textes. Dans des tâches de recherche documentaire sur Internet, la cartographie thématique appliquée aux pages désignées par un moteur de recherche en réponse à une requête permet d'observer des regroupements entre pages abordant des thèmes proches. De cette manière, l'utilisateur peut orienter sa recherche selon les thèmes l'intéressant dans le cadre de sa recherche. Dans le cadre de la veille documentaire, la cartographie d'un ensemble de documents sur différentes périodes consécutives illustre des changements de thèmes au fil du temps.

3.2 Les thèmes utilisés

Afin de construire des cartes thématiques d'un corpus de textes, l'utilisateur doit tout d'abord choisir et définir les thèmes qu'ils souhaitent faire intervenir sur ses cartes. Chaque thème est représenté par une liste de lexies (mots simples ou mots composés) lui étant associées du point de vue de l'utilisateur. Afin de simplifier la phase de construction des thèmes, l'utilisateur n'indique que les formes lemmatisées des lexies. La plate-forme intégre une étape de génération de formes fléchies à l'aide d'une base de données lexicales. Un utilisateur peut par exemple associer les lexies suivantes au thème de l'aviation : avion, appareil, vol, pilote, pilotage, piloter, passager, Boeing, Air France, décollage, etc.

Deux logiciels sont proposés à l'utilisateur afin de l'aider à construire ses thèmes :
- l'outil Memlabor (Perlerin, 2002) permettant une analyse statistique des graphies répétées d'un corpus. En exploitant le principe de cohésion lexicale, MemLabor se fonde sur l'hypothèse que plus une graphie (hors mots d'un anti-dictionnaire contenant par exemple les mots grammaticaux) est répétée dans le corpus, plus elle est susceptible de pouvoir être associée à l'un des thèmes présents dans le corpus (Perlerin, 2004, p. 141). En présentant à l'utilisateur une liste des graphies classées par ordre décroissant de fréquence d'apparition, le logiciel permet une première assistance à l'extraction de graphies intéressantes pour l'utilisateur selon sa tâche à partir d'un corpus.
- l'outil ThemeEditor (Beust, 2002) permettant de composer des graphies en lexies et de les rassembler en thèmes. Ce rassemblement est non exclusif, une lexie pouvant être associée à plusieurs thèmes. Les ressources ainsi constituées sont projetées sur le corpus initial par une annotation XML. Un principe de surlignage avec différentes couleurs (une couleur correspondant à un thème) permet de mettre en évidence la répartition, l'alternance et les enchaînements au long d'un texte des thèmes ainsi créés.

Les thèmes construits à l'aide des logiciels précédents sont stockés en machine sous forme de listes au format XML afin de permettre une facile réutilisation.

3.3 Les traitements réalisés

La plateforme ProxiDocs prend en entrée un fichier XML contenant des thèmes construits par l'utilisateur et un corpus de documents électroniques au format texte ou HTML[8]. Les différents traitements réalisés par ProxiDocs sont présentés en figure 1. Les cartes thématiques produites en sortie de l'application sont représentées dans le format SVG (W3C, 2001), ceci afin de garan-

[8]Dans le cas des documents HTML, seules les parties textuelles sont traitées. Les informations concernant la structure du document et les éléments qu'il contient (telles les images) ne sont pas encore prises en considération dans nos analyses mais pourront l'être dans des prochaines versions de la plate-forme.

tir leur portabilité sur différents systèmes et de permettre à l'utilisateur d'effectuer facilement des zooms et des déplacements.

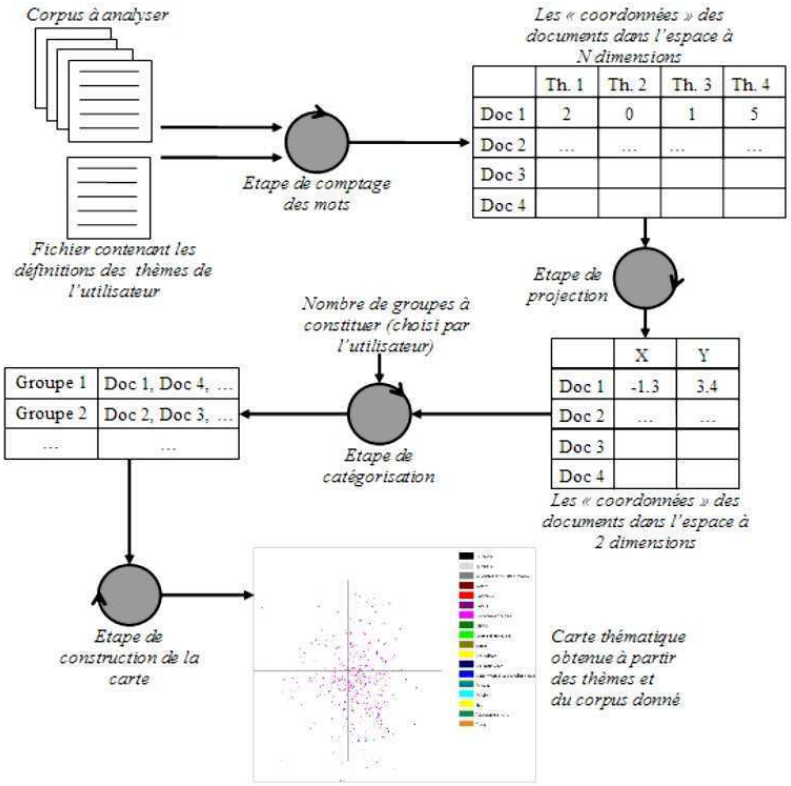


FIG. 1 – Chaîne de traitement de la plate-forme ProxiDocs

Le premier traitement réalisé par la plateforme est un comptage des occurrences de lexies de chaque thème et de leurs formes fléchies dans chaque texte. Un *vecteur* d'entiers de dimension égale au nombre de thèmes choisis et définis par l'utilisateur est alors associé à chaque texte. Supposons qu'un utilisateur fasse intervenir dans la construction de ses cartes thématiques les thèmes de la Bourse, de l'Économie, de la Météo et du Sport, les vecteurs représentant les textes sont de la forme :

Vecteur(Texte) = (nb_lexies(Bourse), nb_lexies(Économie), nb_lexies(Météo), nb_lexies(Sport))

Cette méthode de comptage est appelée *absolue*, du fait qu'elle ne fait pas intervenir la taille des textes lors du comptage. Une seconde méthode, appelée *relative*, détermine pour chaque texte du corpus, les pourcentages des lexies de chaque thème et de leurs formes fléchies par rapport au nombre total de mots du texte. Cette méthode est particulièrement intéressante lorsque la taille des textes du corpus varie significativement.

Les vecteurs obtenus à l'issue de l'étape de comptage prennent place dans des espaces de dimensions égales aux nombres de thèmes définis par l'utilisateur[9]. Pour visualiser ces vecteurs sur des cartes en 2 ou 3 dimensions, il faut réaliser une *projection* de ces vecteurs. Pour cela, nous proposons plusieurs méthodes d'analyse des données dont l'*Analyse en Composante Principales* (ACP) (Bouroche et Saporta, 1980) et l'*Analyse Factorielle des Correspondances* (Benzécri, 1980). À l'issue de cette étape de projection, nous proposons aux utilisateurs des regrou-

[9]Dans l'exemple précédent, les vecteurs représentant les textes prennent place dans un espace à 4 dimensions.

pements automatiques des textes sur les cartes. Pour cela, nous avons intégré une méthode de catégorisation, appelée *Catégorisation Hiérarchique Ascendante* (Bouroche et Saporta, 1980).

Afin de mettre en évidence les résultats produits par l'enchaînement de ces différents traitements, nous présentons dans la partie suivante des exemples de cartes thématiques construites par ProxiDocs à partir d'un corpus d'articles de presse et de thèmes que nous avons définis. Pour plus de détails sur ces traitements, nous renvoyons à (Roy et Beust, 2004).

3.4 Les cartes thématiques obtenues sur un exemple

Les cartes présentées dans cette section ont été construites à partir d'un corpus constitué de 789 articles de l'année 1989 du journal "Le Monde" totalisant environ 700 000 graphies. Le jeu de thèmes utilisé est généraliste et propose des descriptions des 18 thèmes suivants : la justice, la religion, la violence, l'éducation, l'agriculture, la sécurité routière, l'aviation, la navigation, le dopage, l'économie, la politique, l'aérospatial, la guerre, l'informatique, la pollution, le sport, la télévision et le travail . Nous avons construit un tel ensemble de thèmes dans un objectif de découverte des sujets abordés dans des corpus de textes nous étant peu connus. Les outils MemLabor et ThemeEditor ont été utilisés lors de la construction de ces thèmes. La carte présentée en figure 2 a été construite à partir de ces entrées[10].

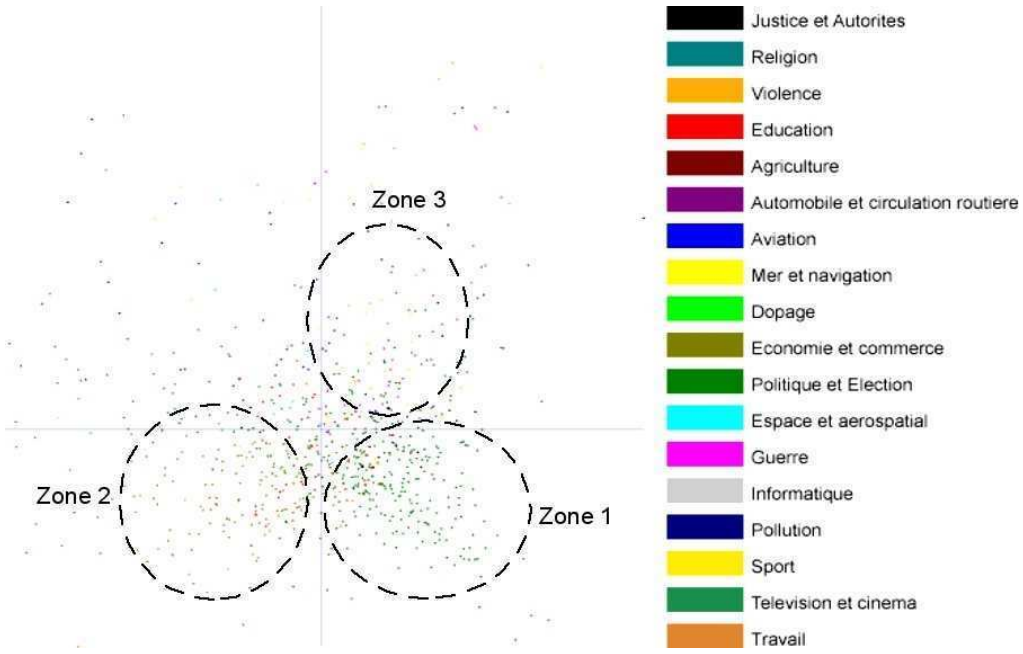


FIG. 2 – Carte thématique du corpus analysé représentant chaque article par un point. Version électronique : `http ://www.info.unicaen.fr/~troy/proxidocs/cartes_classiques/acp1.html`

Chaque point sur la carte représente un article du corpus analysé. La couleur d'un point correspond au thème majoritaire repéré dans l'article représenté[11]. Chaque point est un hyperlien vers l'article représenté. Les zones 1, 2 et 3 ont été marquées manuellement sur la carte afin d'en faciliter l'analyse. La plupart des articles de la zone 1 sont de thème majoritaire Politique et élection

[10] La méthode de comptage relative et la méthode de projection de l'ACP ont été utilisées lors de la construction des cartes présentées dans cet article.

[11] Une légende de couleurs est disponible sur la droite de la carte, l'association d'une couleur à un thème est réalisée par l'utilisateur lors de la construction des thèmes.

alors que la zone 2 contient plus particulièrement des articles de thème majoritaire Économie et commerce. La zone 3, présente en haut et à droite de la carte, contient un petit nombre d'articles de thème majoritaire Guerre.

À partir de la carte précédente, nous avons choisi d'aider l'utilisateur dans son analyse en appliquant une méthode de catégorisation automatique sur les textes de la carte. La carte présentée en figure 3 met en évidence les résultats de cette catégorisation[12]. Chaque groupe de textes est représenté sur la carte par un disque de taille proportionnelle à sa cardinalité. Chaque disque est centré sur le centre de gravité de l'ensemble des points représentant les documents du groupe. La couleur attribuée à ce disque correspond au thème majoritaire repéré dans les textes du groupe. Chaque disque est un hyperlien vers le texte *représentatif* du groupe, c'est-à-dire le texte étant le plus proche de son centre de gravité. Sur cette carte, chaque groupe est caractérisé par les cinq lexies des thèmes les plus fréquentes au sein des textes du groupe.

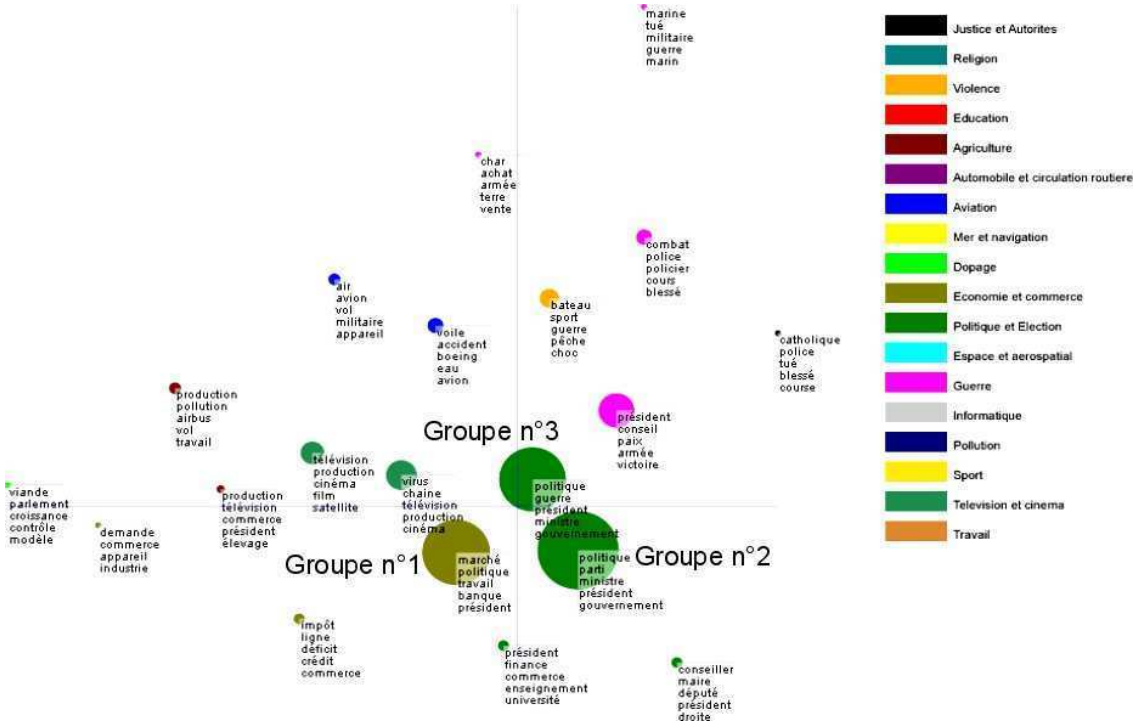


FIG. 3 – Carte thématique mettant en évidence des groupes d'articles. Version électronique : http ://www.info.unicaen.fr/~troy/proxidocs/cartes_categorisation/carte_avec_groupes.htm

La couleur, la taille et la disposition des groupes sur la carte donnent une idée sur les thèmes abordés dans les textes du corpus ainsi que sur leur répartition. En visualisant les textes représentatifs des groupes, l'utilisateur peut avoir une idée plus précise des thèmes abordés dans les textes de chaque groupe. Ainsi, les textes représentatifs des groupes 1 et 2 (groupes marqués manuellement sur la carte, tout comme le groupe 3), traitent respectivement du rachat des parts boursières d'une grande entreprise et des enjeux des futures élections européennes. Il est alors possible de déduire que les thèmes abordés dans ces articles se retrouvent dans les autres textes de leurs groupes respectifs.

Les lexies caractérisant les groupes de textes sur la carte peuvent également aider l'utilisateur dans l'appréhension des thèmes abordés au sein des groupes. Les cartes étant interactives, lorsque l'utilisateur passe avec sa souris sur l'une de ces lexies, celle-ci ainsi que les lexies

[12] Le nombre de groupes empiriquement choisi dans cet exemple est de 20.

identiques caractérisant les autres groupes se colorient en rouge. Cette opération permet ainsi d'observer des lexies communes à plusieurs groupes ou bien des lexies ne caractérisant qu'un seul groupe. Nous pouvons ainsi observer que la lexie politique est commune aux groupes 1, 2 et 3. Par contre, si nous souhaitons différencier ces trois groupes les uns des autres, nous pouvons remarquer que le groupe 3 est le seul à posséder la lexie guerre, le groupe 2 est le seul à posséder la lexie parti et le groupe 1 est le seul à posséder les lexies marché, travail et banque.

Afin d'offrir à l'utilisateur une vision encore plus précise et dynamique de son corpus, nous proposons de tenir compte du temps dans la construction de ses cartes. Pour cela, nous construisons des cartes thématiques à partir du corpus et des thèmes choisis par l'utilisateur sur différentes périodes[13]. Une carte dynamique proposant un enchaînement automatique de ces cartes peut alors mettre en évidence l'évolution des thèmes abordés dans les articles du corpus au fil du temps. A partir du corpus et des thèmes considérés précédemment, une carte dynamique a pu être construite, des extraits de cette carte sur deux périodes sont présentés en figure 4.

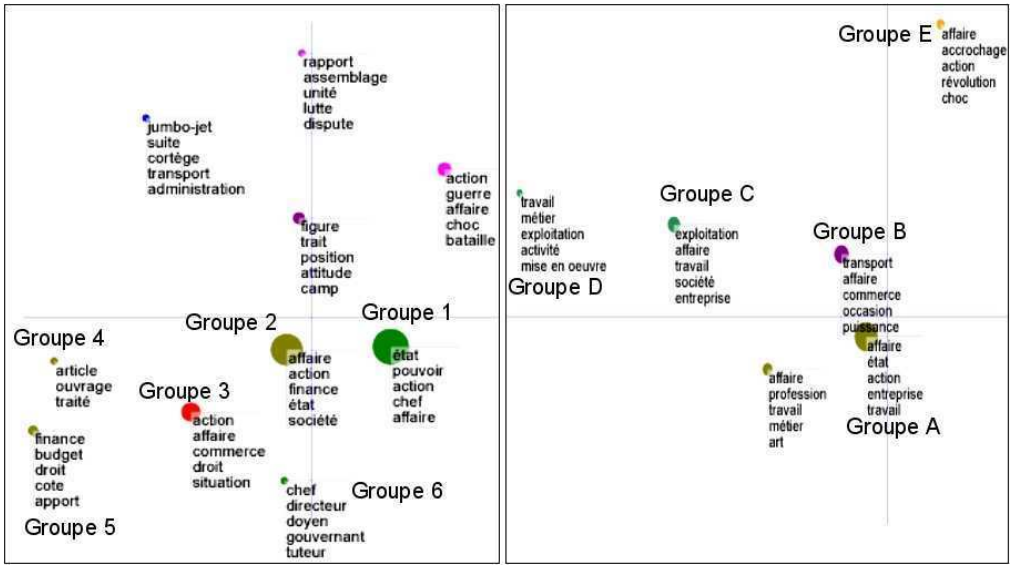


FIG. 4 – Extraits de la carte dynamique globale du corpus. Version électronique : http ://www.info.unicaen.fr/~troy/proxidocs/cartes_temps/carte_dyn_1.html

L'extrait situé sur la partie gauche de la figure met en évidence des groupes d'articles dont la date de publication est comprise entre le 28 janvier et le 27 février 1989. L'extrait de droite représente des groupes d'articles publiés entre le 28 février et le 27 mars. Pour des raisons de lisibilité, la légende de couleurs, identique à celles des cartes des figures 2 et 3, n'est pas rappelée. Pour ces mêmes raisons, certains groupes sont marqués manuellement sur les extraits.

L'extrait de gauche met en évidence deux importants groupes d'articles de thèmes majoritaires Politique et élection et Économie et commerce (groupes 1 et 2, contenant respectivement 28 et 24 articles). Sur la partie en bas et à gauche, un groupe de thème majoritaire Education est également présent (groupe 3). Ce groupe se situe à proximité de groupes de thèmes majoritaires Économie et commerce et Politique et élection (groupes 1, 4, 5 et 6), ce qui laisse penser que les articles contenus dans ce groupe abordent d'une certaine manière ces deux thèmes. Cette idée se confirme en visualisant le texte représentatif du groupe 3, ce dernier abordant des réformes budgétaires du gouvernement sur le système éducatif. La partie de l'extrait située au-dessus de l'axe des abscisses met en évidence des groupes de petite taille (contenant 1 ou 2 articles). En

[13]Dans l'exemple présenté ici, nous nous sommes basés sur la date de publication des articles.

visualisant les articles représentatifs de ces groupes, nous pouvons remarquer qu'ils abordent des sujets d'actualité très ponctuels, tels un crash d'avion et des actes terroristes.

L'extrait de droite met toujours en évidence un important groupe de thème majoritaire Économie et commerce (groupe A, contenant 22 articles). En visualisant la carte dynamique globale du corpus, nous pouvons remarquer qu'un tel groupe est présent tout au long de la période étudiée, ce qui peut laisser penser que le thème de l'économie est constamment abordé dans les articles du corpus analysé. Au-dessus du groupe A se situe un groupe de thème majoritaire Automobile et circulation routière (groupe B). La grande majorité des articles de ce groupe traitent des ventes de voitures en France. Nous pouvons également remarquer la présence de petits groupes (contenant 1 ou 2 articles) de thèmes majoritaires Télévision et cinéma (groupes C et D) et Violence (Groupe E). Les articles représentatifs de ces groupes traitent de sujets d'actualité ponctuels liés à la disparition d'un grand producteur de cinéma (groupes C et D) et au décès d'un jeune boxeur lors d'un combat (groupe E).

La figure 4 commentée précédemment présente deux extraits de la carte dynamique globale retournée à l'utilisateur. Cette carte globale permet entre autres de visualiser les thèmes constamment abordés dans les articles du corpus tout au long de la période étudiée, mais aussi d'observer des thèmes liés à l'actualité abordés de façon plus ponctuels dans les textes.

D'autres possibilités, non détaillées dans cet article, sont également offertes par la plate-forme, telle la possibilité de construire des cartes thématiques en 3 dimensions[14].

4 Conclusions et perspectives

Dans cet article, nous avons présenté les principes de fonctionnement de la plate-forme logicielle ProxiDocs dédiée à la cartographie thématique de corpus de textes. Nous avons illustré les intérêts de cette plateforme sur un exemple précis : la découverte des thèmes abordés dans un corpus d'articles d'un grand quotidien français. Les différentes cartes construites permettent de mettre en évidence les principaux sujets abordés dans les textes de cet ensemble, d'observer des groupes de textes abordant des thèmes proches et de visualiser l'évolution des sujets abordés dans ces articles au fil du temps.

Plusieurs améliorations de la plate-forme ProxiDocs sont actuellement envisagées. D'un point de vue théorique, nous souhaitons intégrer un modèle de représentation des thèmes beaucoup plus fin que celui utilisé jusqu'à présent (dépassant les simples listes de lexies). Le modèle de représentation lexicale envisagé (intitulé LUCIA) est expérimenté depuis plusieurs années au sein de notre équipe (Nicolle et al., 2002; Perlerin, 2004). L'intégration de ce modèle à notre plate-forme permettrait à l'utilisateur de préciser et de structurer les lexies relevant des thématiques de son choix en précisant, pour chacune d'elles, les significations qu'ils jugent pertinentes et appropriées à la tâche qu'il vise. Les cartes ainsi produites devraient révéler des informations plus précises sur le corpus analysé et surtout plus en rapport avec le point de vue de l'utilisateur ou du groupe d'utilisateurs destinataires des cartes.

D'un point de vue applicatif, nous avons commencé le développement de deux composants :
- le métamoteur de recherche ProxiDocs Web interrogeant des moteurs de recherche généralistes (tels Google, Yahoo, etc.) à partir de mots-clés saisis par l'utilisateur et retournant les

[14]Des exemples de cartes en 3 dimensions sont disponibles à l'adresse : http://www.info.unicaen.fr/~troy/proxidocs/cartes_3D

pages proposés par ces moteurs sous la forme de cartes thématiques construites à partir de thèmes choisis et définis par l'utilisateur ;
– et l'outil ProxiDocs Mail réalisant la cartographie dynamique d'un flux de courriers électroniques selon des thèmes choisis et définis par l'utilisateur.

Ces deux outils devraient nous permettre de traiter d'autres types de corpus et ainsi proposer de nouveaux services aux utilisateurs.

Afin de mettre en évidence les intérêts et les limites de notre plate-forme, une expérimentation avec un grand nombre d'utilisateurs nous semble incontournable. Cette expérimentation, et surtout son évaluation, n'est pas sans poser problèmes car il n'est pas question ici de juger (par exemple, en terme de rappel et de précision) si la plate-forme produit des résultats corrects ou non, mais plutôt d'évaluer la façon dont les utilisateurs s'approprient l'outil chacun à leur façon selon leurs buts. Ce n'est donc pas simplement le logiciel qu'il faut évaluer mais le couple *outil-utilisateur*. À travers une telle expérimentation, de nouveaux besoins devraient émerger, ceci nous persuadant à toujours mieux instrumentaliser la dimension intertextuelle de la sémantique des langues.

Références

Benzécri J.P. (1980), *L'analyse des données - tome 2 : l'analyse des correspondances*, Éditions Bordas.

Beust P. (2002), Un outil de coloriage de corpus pour la représentation de thèmes, Actes des *6èmes Journées internationales de l'Analyse statistique de Données Textuelles*.

Bouroche J.M. et Saporta G. (1980) *L'analyse des données*, Collection Que sais-je ?, PUF.

Chung W., Chen H. et Numaker J.F.Jr. (2002) Business Intelligence Explorer : A Knowledge Map Framework for Discovering Business Intelligence on the Web, Actes de la *36ème HICSS*.

Lelu A. et Aubin S. (2001) Vers un environnement complet de synthèse statistique de contenus textuels, Présentation au séminaire *Association pour la mesure des sciences et des techniques* du 13/11/2001.

Mokrane A., Arezki R., Dray G. et Poncelet P. (2004) Cartographie automatique du contenu d'un corpus de documents textuels, Actes des *7èmes Journées internationales de l'Analyse statistique de Données Textuelles*, pages 816-823.

Nicolle A. (1996), L'expérimentation et l'intelligence artifi cielle, *Intellectica*, numéro 22, pages 9-19, Association pour la Recherche Cognitive (ARC).

Nicolle A., Beust P. et Perlerin V. (2002), Un analogue de la mémoire pour un agent logiciel interactif, *In Cognito*, numéro 21, pages 37-66.

Perlerin V. (2002), MemLabor, un environnement de création, de gestion et de manipulation de corpus de textes, Actes de *TALN / RECITAL 2002*, pages 507 à 516.

Perlerin V. (2004), *Sémantique légère pour le document : assistance personnalisée pour l'accès au document et l'exploration de son contenu*, Thèse d'informatique de l'Université de Caen.

Pichon R. et Sébillot P. (1999), Différencier les sens des mots à l'aide du thème et du contexte de leurs occurrences : une expérience, Actes de *TALN 1999*, pages 279-288.

Roy T. et Beust P. (2004), ProxiDocs, un outil de cartographie et de catégorisation thématique de corpus, Actes des *7èmes Journées internationales de l'Analyse statistique de Données Textuelles*, pages 978-987.

Spinat E. (2002), Pourquoi intégrer des outils de cartographie au sein des systèmes d'information de l'entreprise ?, Colloque *Cartographie de l'Information*, Paris.

Thlivitis T. (1998), *Sémantique interprétative intertextuelle : assistance anthropocentrée à la compréhension des textes*, Thèse d'informatique de l'Université de Rennes I.

W3C (2001), *Scalable Vector Graphics (SVG)*, `http://www.w3.org/TR/SVG/`.

Segmentation morphologique à partir de corpus

Delphine Bernhard

Laboratoire TIMC-IMAG
Institut de l'Ingénierie et de l'Information de Santé
Pavillon Le Taillefer – Faculté de Médecine
F-38706 LA TRONCHE cedex
Delphine.Bernhard@imag.fr

Mots-clefs – Keywords

Segmentation morphologique, alignement de segments de mots, corpus.

Morphological segmentation, word segments alignment, corpus.

Résumé – Abstract

Nous décrivons une méthode de segmentation morphologique automatique. L'algorithme utilise uniquement une liste des mots d'un corpus et tire parti des probabilités conditionnelles observées entre les sous-chaînes extraites de ce lexique. La méthode est également fondée sur l'utilisation de graphes d'alignement de segments de mots. Le résultat est un découpage de chaque mot sous la forme (préfixe*) + base + (suffixe*). Nous évaluons la pertinence des familles morphologiques découvertes par l'algorithme sur un corpus de textes médicaux français contenant des mots à la structure morphologique complexe.

We describe a method that automatically segments words into morphs. The algorithm only uses a list of words collected in a corpus. It is based on the conditional probabilities between the substrings extracted from this lexicon. The method also makes use of word segments alignment graphs. As a result, all words are segmented into a sequence of morphs which has the following pattern: (prefix*) + base + (suffix*). We evaluate the morphological families discovered by the algorithm using a corpus of French medical texts containing words whose morphological structure is complex.

1 Introduction

L'analyse des mots en morphèmes, qui sont les plus petites unités porteuses de sens, facilite l'exécution de diverses tâches telles que la recherche d'informations (Hahn et al., 2003), la construction de dictionnaires (Lovis et al., 1995) ou de terminologies (Zweigenbaum, Grabar, 2000). Les méthodes existantes permettent de découvrir des suffixes flexionnels ou dérivationnels (Gaussier, 1999), voire également des préfixes (Déjean, 1998; Schone, Jurafsky, 2001; Goldsmith, 2001; Creutz, Lagus, 2002). Cependant, l'analyse finale d'un mot se limite généralement à 3 unités morphologiques au plus ((préfixe?) + base + (suffixe?)).

Certaines langues, comme l'allemand, et langues de spécialité, comme le vocabulaire médical, présentent des caractéristiques nécessitant une segmentation plus fine, notamment en raison du procédé de composition à la base de la formation des mots. La procédure de segmentation que nous proposons permet d'obtenir un découpage de chaque mot sous la forme (préfixe*) + base + (suffixe*) sans imposer de limite au nombre d'affixes. Pour cela, nous avons combiné l'utilisation de trois propriétés caractérisant les morphèmes et leurs frontières :

- Il existe une frontière morphémique lorsqu'il est difficile de prédire le segment suivant en fonction des segments précédents. Par exemple, la méthode proposée par (Harris, 1955) pour les phonèmes et appliquée par (Hafer, Weiss, 1974) et (Déjean, 1998) à la langue écrite utilise le nombre de phonèmes différents qui peuvent suivre une suite de phonèmes. Un nombre élevé de phonèmes indique une frontière entre deux morphèmes. Ainsi, dans notre corpus, seule 1 lettre, le "o", peut suivre la séquence initiale "micr" en français tandis que la séquence "micro" peut être suivie par 10 lettres différentes, marquant ainsi une frontière morphémique. Nous avons expérimenté une méthode similaire utilisant les probabilités conditionnelles observées entre les sous-chaînes extraites d'une liste de mots pour prédire les frontières morphémiques d'un mot (section 2.1).

- La similitude graphique est un indice de lien morphologique. En effet, les mots morphologiquement liés partagent une base identique et diffèrent par leurs affixes. La découverte des bases et des affixes peut donc se faire en comparant la graphie des mots, par exemple en recherchant la plus longue chaîne initiale commune (Gaussier, 1999; Zweigenbaum, Grabar, 2000). Dans de nombreux cas cependant la base ne correspond pas à la chaîne initiale, notamment pour les formes préfixées comme "antihormone" ou "précancéreux " par rapport aux formes non préfixées "hormone" et "cancéreux". Nous avons donc choisi de dissocier les phases d'apprentissage des affixes (section 2.2) et des bases (section 2.3). Nous utilisons également une structure de données (graphe) permettant d'aligner les mots à partir d'une base qui peut apparaître à toute position dans les mots comparés.

- Si l'on remplace les mots par la liste des morphèmes qu'ils contiennent, il est possible de comprimer les données du lexique. La meilleure segmentation d'un ensemble de mots est alors celle qui donne la représentation la plus compacte des données et qui réutilise un maximum de morphèmes. Ce principe est notamment utilisé par (Goldsmith, 2001) et (Creutz, Lagus, 2002). Ainsi, notre algorithme privilégie la réutilisation d'unités morphologiques apprises lors de la première phase de l'apprentissage (section 2.2).

Nous présentons dans un premier temps notre méthode de segmentation morphologique (section 2). Nous décrivons ensuite l'évaluation effectuée à partir d'une liste de mots extraits d'un corpus médical français (section 3). Enfin, nous discutons les résultats et proposons des possibilités d'évolution de l'algorithme (section 4).

2 Présentation de la méthode

Nous détaillons tout d'abord notre méthode de segmentation morphologique basée sur les probabilités conditionnelles (section 2.1) permettant de découvrir une liste d'affixes initiaux (section 2.2). Cette liste d'affixes est ensuite utilisée pour la découverte des bases et la segmentation finale (section 2.3).

2.1 Segmentation basée sur les probabilités conditionnelles

Nous pouvons à partir de chaque mot former un graphe directionnel dont les nœuds sont étiquetés par une sous-chaîne du mot. Il existe un arc entre deux nœuds si les sous-chaînes correspondantes se suivent immédiatement dans le mot. Pour tout chemin du graphe tel que le premier nœud n'a pas prédécesseur, la concaténation des sous-chaînes associées aux nœuds du chemin est équivalente au mot. Afin de faire la distinction entre les sous-chaînes apparaissant en début, milieu et fin de mot, les frontières de mots sont marquées par le caractère "#". Tout mot de longueur N est ainsi décomposable en un ensemble de sous-chaînes de longueur 1 (caractères) à N+2 (mot complet auquel s'ajoutent les marques de début et de fin "#"). La figure 1 représente une sous-partie du graphe des sous-chaînes du mot "microcalcifications".

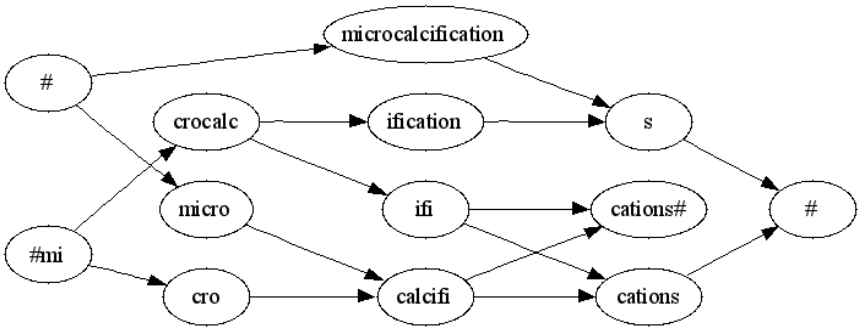


Figure 1 : Sous-partie du graphe des sous-chaînes du mot "microcalcifications".

Pour un arc du graphe reliant deux nœuds étiquetés par les sous-chaînes s_1 et s_2, nous pouvons estimer les probabilités conditionnelles $p(s_1|s_2)$ et $p(s_2|s_1)$ à partir du nombre d'occurrences dans le lexique de s_1, s_2 et de leur concaténation notée $s_1.s_2$:

$$p(s_1 | s_2)=\frac{f(s_1.s_2)}{f(s_2)} \quad \text{et} \quad p(s_2 | s_1)=\frac{f(s_1.s_2)}{f(s_1)}$$

Si les sous-chaînes s_1 et s_2 appartiennent à des morphèmes différents, $p(s_1|s_2)$ et $p(s_2|s_1)$ sont faibles car il est difficile de prédire la succession de s_1 et s_2 . A l'inverse, si elles appartiennent au même morphème, les probabilités conditionnelles sont plus fortes. Nous conservons le maximum de $p(s_1|s_2)$ et $p(s_2|s_1)$.

Pour chaque position dans le mot, nous obtenons ainsi un ensemble de valeurs correspondant aux maximums des probabilités conditionnelles entre les sous-chaînes se terminant à cette position et celles commençant à cette position. Le nombre de valeurs n'est pas le même pour toutes les positions. Afin d'analyser ces données et de les comparer, nous calculons à chaque position dans le mot le maximum, le quartile supérieur, la médiane, la moyenne, le quartile inférieur et le minimum[1] de ces valeurs. La figure 2 représente graphiquement ces données pour le mot "microcalcifications".

Nous pouvons observer des minimums locaux sur ces courbes à diverses positions dans le mot. Ces minimums indiquent des frontières morphémiques potentielles. Un minimum local constitue un point de segmentation potentiel si la différence avec le maximum précédent et le maximum suivant est au moins égale à un écart-type des valeurs de la courbe. Nous appliquons un deuxième critère afin de garantir la précision de la segmentation : une frontière morphémique est validée si elle correspond à des points de segmentation potentiels sur au moins la moitié des courbes. Les résultats ont montré qu'une courbe ne pouvait être

[1] La médiane et les quartiles inférieur et supérieur permettent de partager l'ensemble des valeurs en 4 sous-groupes de valeurs de même taille : 25 % des valeurs sont inférieures au quartile inférieur, 50 % sont inférieures à la médiane et 75 % sont inférieures au quartile supérieur.

privilégiée par rapport à une autre : nous avons donc appliqué la même pondération à chacune d'elles. Sur la figure 2, les frontières morphémiques validées sont indiquées par des flèches. Pour le mot "microcalcifications", la segmentation proposée est donc "micro + calcification + s".

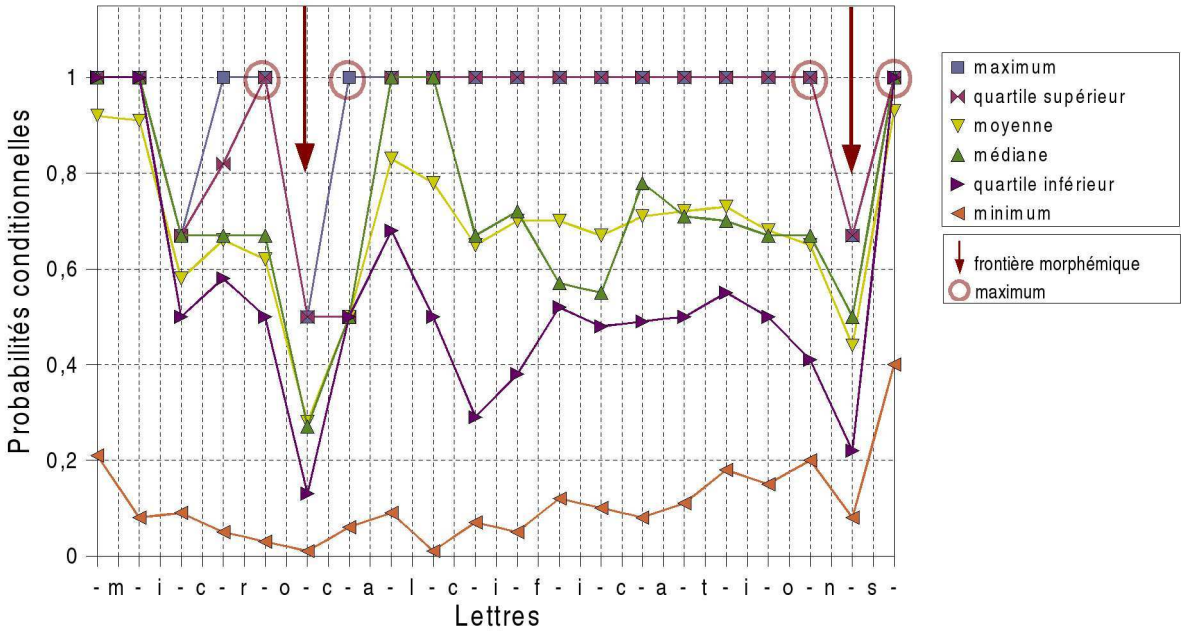


Figure 2 : Variation des probabilités conditionnelles pour le mot "microcalcifications".

Cette méthode permet d'identifier des frontières morphémiques mais n'est pas suffisante pour reconstituer des familles morphologiques. Elles peut donner lieu à des segmentations erronées notamment pour les mots courts ("foyer" donne "fo + yer", "mêlé" : "mê + lé") ou lorsque l'on trouve peu de mots de la même famille dans le corpus ("incubées" : "incu + bées"). Le tableau 1 donne les segmentations obtenues pour un ensemble de mots appartenant à la famille de "microcalcifications". On constate l'absence de segmentation des mots les plus courts et du segment "calcification", ainsi qu'une segmentation erronée du préfixe "macro" en "ma + cro". Cette dernière peut s'expliquer par une prépondérance du préfixe "micro" (25 occurrences) par rapport au préfixe "macro" (13 occurrences). De plus, en l'absence de base commune, il n'est pas possible de reconstituer la famille morphologique. Cette méthode de segmentation est donc utilisée uniquement pour obtenir une liste d'affixes initiaux qui sera réutilisée pour la découverte des bases et la segmentation finale des mots.

Mots	Segmentation
calcifier	calcifier
calcifiée	calcifiée
calcifiés	calcifi + és
calcifiées	calcifiées
calcifications	calc + ification + s
microcalcification	micro + calcification
micro-calcifications	micro- + calcification + s
macrocalcifications	ma + cro + calcification + s

Tableau 1 : Exemples de segmentations obtenues en utilisant les probabilités conditionnelles.

2.2 Découverte des affixes initiaux

La sélection des affixes valides peut se faire en utilisant un critère de fréquence (Déjean, 1998) : dans ce cas, seuls sont conservés les affixes qui dépassent un certain nombre d'occurrences après la phase d'apprentissage. Nous proposons un autre critère de sélection des affixes initiaux. En effet, les mots morphologiquement complexes sont généralement plus longs que les mots morphologiquement simples. L'apprentissage des affixes initiaux n'est donc effectué que sur les mots les plus longs du lexique et non pas sur l'ensemble des mots. Le nombre de mots du lexique d'apprentissage est paramétrable (seules quelques centaines de mots sont nécessaires).

Nous segmentons chacun des mots du lexique d'apprentissage en utilisant la méthode de segmentation basée sur les probabilités conditionnelles décrite dans la section 2.1. Dans la mesure où la segmentation met à jour aussi bien des préfixes que des suffixes, il n'est pas possible de déterminer le type (préfixe, suffixe ou base) d'un segment uniquement en utilisant des critères positionnels. (Vergne, 2003) utilise les différences de fréquence et de longueur pour distinguer mots vides (fréquents et courts) et mots pleins (rares et longs) dans un énoncé. Nous pouvons établir un parallèle entre mots vides et affixes d'une part et mots pleins et bases d'autre part. En effet, une base est généralement moins fréquente et plus longue qu'un affixe : la combinaison de ces deux critères nous permet de repérer une pseudo-base[2] parmi les segments proposés. Les pseudo-bases identifiées par cette méthode correspondent aux segments soulignés dans le tableau 1. Le tableau 2 donne quelques exemples de mise en oeuvre de cette méthode. Dans le cas du mot "chimio-hormonothérapie", effectif minimal (2 : "chimio") et longueur maximale (9 : "othérapie") ne correspondent pas, la segmentation est alors considérée comme non valide.

Segments	micro	calcification	s
Effectifs	37	6	8289
Longueurs	5	13	1
Segments	multi	dimension	nelle
Effectifs	32	5	41
Longueurs	5	9	5
Segments	chimio-	hormon	othérapie
Effectifs	2	29	25
Longueurs	7	6	9

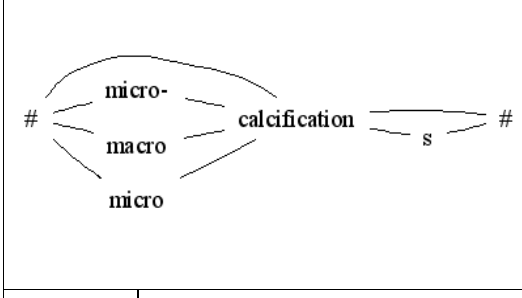
Tableau 2 : Repérage de la base parmi des segments. Les cases grisées correspondent respectivement à l'effectif minimal et à la longueur maximale.

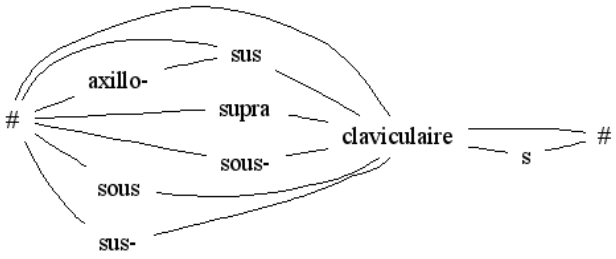
Afin d'augmenter le nombre d'affixes extraits nous recherchons la pseudo-base dans l'ensemble des mots du lexique et nous procédons à l'alignement de ces mots en les insérant dans un graphe. La pseudo-base sert de point d'ancrage. Nous appliquons des critères supplémentaires pour vérifier la validité de la pseudo-base en fonction des mots dans lesquels elle apparaît :

[2] Nous réutilisons ici le terme "pseudo-base" employé par (Schone, Jurafsky, 2001). Il peut en effet s'agir d'une base non valide ou contenant des affixes.

- Le nombre de mots de cette liste doit être supérieur ou égal à 2.
- La pseudo-base doit débuter un de ces mots. Par exemple, la segmentation du mot "radiopharmaceutiques" est "radio + pharmac + eutiques". Le segment "eutiques" est reconnu comme étant la pseudo-base car il est à la fois le plus long et le moins fréquent. Les 5 mots contenant "eutiques" sont : "chimiothérapeutiques", "postthérapeutiques", "pharmaceutiques", "radiopharmaceutiques", et "thérapeutiques". Aucun des mots du corpus ne commence par "eutiques", la pseudo-base est donc considérée comme non valide.

La figure 3 donne deux exemples d'alignement à partir des pseudo-bases "calcification" et "claviculaire" ainsi que les préfixes et les suffixes découverts par l'alignement ("#" indique le préfixe ou suffixe vide). Seuls sont conservés les préfixes de longueur supérieure ou égale à deux. L'ensemble des préfixes et suffixes obtenus à la fin de cette étape constitue la liste des affixes initiaux.





préfixes	micro-, micro, macro, #	préfixes	axillo-, sus, sus-, sous, sous-, supra, axillo-sus, #
suffixes	s, #	suffixes	s, #

Figure 3 : Alignement des mots contenant les pseudo-bases "calcification" et "claviculaire".

2.3 Découverte des bases et segmentation des mots

Cette dernière phase permet de segmenter tous les mots du corpus en réutilisant les affixes découverts lors de la phase précédente pour obtenir une liste des bases présentes dans le corpus :

1. Pour chaque mot du lexique, recherche des affixes initiaux qu'ils contient. Si l'on retranche d'un mot les diverses combinaisons possibles de ces affixes et de la chaîne vide, il est possible d'obtenir des pseudo-bases. Le mot lui-même constitue une pseudo-base, ce qui permet notamment de conserver dans la liste des pseudo-bases les mots non fléchis. Ces pseudo-bases doivent avoir une longueur minimale de 3. Par exemple, le mot "calcifiées" contient les suffixes initiaux "iées", "ées", "es" et "s". Les pseudo-bases obtenues en retranchant ces suffixes initiaux sont : "calcifiée", "calcifié", "calcifi" et"calcif", auxquelles s'ajoute "calcifiées" (mot complet).

2. Pour chaque pseudo-base, recherche des mots contenant la pseudo-base dans le lexique. Par exemple, les mots contenant la pseudo-base "calcifié" sont : "calcifié", "calcifiée", "calcifiés", "calcifiées".

3. Construction du graphe d'alignement des mots contenant la pseudo-base. Nous appliquons deux contraintes sur le graphe d'alignement : il doit contenir un nombre suffisant d'affixes initiaux et le nombre de mots contenant la pseudo-base ne doit pas excéder un certain seuil. Nous avons fixé expérimentalement la proportion d'affixes initiaux à 75 % et le nombre maximal de mots à 50. Afin d'éviter l'absence d'analyse lorsque ces critères ne sont pas vérifiés, deux tentatives supplémentaires sont effectuées : nous supprimons dans un premier temps les préfixes inconnus du graphe d'alignement, puis, si l'alignement n'est toujours pas considéré comme valide, nous supprimons les suffixes inconnus. La meilleure pseudo-base est celle dont le graphe d'alignement contient le plus grand nombre d'affixes initiaux. Les mots contenant la meilleure pseudo-base sont alors segmentés en fonction de leur alignement. Le tableau 3 représente les pseudo-bases obtenues à partir du mot "calcifiées" ainsi que les segmentations correspondantes. La meilleure pseudo-base est "calcifi".

calcifiées	calcifi	calcifiée	calcif	calcifié
calcifiées	calcifi**+er**	calcifiée	calcif+ier	calcifié
	calcifi**+é**	calcifiée+**s**	calcif+ié	calcifié**+e**
	calcifi**+é+e**		calcif+ié+**e**	calcifié**+s**
	calcifi**+é +s**		calcif+ié+**s**	calcifié**+e+ s**
	calcifi**+é+e+s**		calcif+ié+**e+s**	
	calcifi**+cation**		calcif+ication	
	calcifi**+cation+s**		calcif+ication+**s**	
	micro+calcifi**+cation**		**micro+**calcif+ication	
	micro+calcifi**+cation+s**		**micro+**calcif+ication+**s**	
	micro-+calcifi**+cation+s**		**micro-+**calcif+ication+**s**	
	macro+calcifi**+cation+s**		**macro+**calcif+ication+**s**	

Tableau 3 : Pseudo-bases et segmentations obtenues à partir du mot "calcifiées". Les affixes initiaux sont marqués en gras. Les segmentations validées se trouvent dans la colonne grisée.

3 Évaluation

Nous avons évalué la méthode sur un corpus composé de 80 documents traitant du cancer du sein. Ce corpus se compose de 33 articles scientifiques et de 47 pages web. Il comprend environ 280 000 mots pour 12 587 formes différentes. Nous évaluons la validité des bases associées à chaque mot et non pas la position des points de segmentation. L'ensemble des mots qui contiennent la même base forme une famille morphologique. Nous vérifions si deux mots associés à la même base par l'algorithme (comme c'est le cas pour "microcalcification" et "calcifier") sont effectivement morphologiquement liés. Nous avons automatisé l'évaluation des variantes flexionnelles en utilisant le fichier DLF (formes simples) produit par INTEX (Silberztein, 1993) par l'application du dictionnaire DELAF. Dans ce fichier, chaque forme est associée à un ou plusieurs lemmes. Si deux formes sont associées au même lemme dans le fichier DLF et à la même base par l'algorithme, la relation entre les deux formes est considérée comme valide (relation flexionnelle). L'analyse du lexique par INTEX produit

12 116 relations binaires entre mots du corpus. En ce qui concerne les mots reliés par dérivation ou composition ainsi que les mots ne figurant pas dans le fichier DLF, l'évaluation a été faite manuellement.

Le tableau 4 présente les résultats de l'évaluation en fonction de la taille du lexique d'apprentissage et le tableau 5 donne quelques exemples de familles morphologiques obtenues pour un lexique d'apprentissage de 100 mots. La précision décroît avec l'augmentation du nombre de mots du lexique d'apprentissage tandis que le rappel des relations flexionnelles augmente pour culminer à environ 75 %. Pour 600 mots dans le lexique d'apprentissage, le rappel n'est pas meilleur, ce qui semble indiquer un seuil limite dans l'apprentissage.

Nombre de mots du lexique d'apprentissage	100	200	300	400	500
Préfixes initiaux	35	69	95	113	147
Suffixes initiaux	18	77	112	138	177
Nombre de bases	9 389	6 989	6 740	6 224	5 336
Nombre total de relations	5 284	15 794	17 610	21 889	33 837
Relations flexionnelles (rappel)	3 436 (28,4%)	7 319 (60,4%)	7 291 (60,2%)	8 116 (67,0%)	9 261 (76,4%)
Relations dérivationnelles et compositionnelles valides	1 740	7 145	8 343	10 524	14 807
Relations non valides	108	1 330	1 976	3 249	9 769
Précision	98,0%	91,6%	88,8%	85,2%	71,1%

Tableau 4 : Résultats de l'évaluation.

Bases	Variantes
pathologi	anatomo-pathologie, anatomo-pathologique, anatomo-pathologiques, anatomo-pathologiste, anatomopathologie, anatomopathologique, anatomopathologiques, anatomopathologiste, anatomopathologistes, cytopathologique, cytopathologiste, histopathologie, histopathologique, histopathologiques, pathologie, pathologies, pathologique, pathologiques, pathologiste, pathologistes, physio-pathologique, physiopathologiques, radio-pathologiques, radiologique-anatomopathologique, radio-pathologique, radiopathologiques.
thérapi	chimiothérapie, chimiothérapies, chimiothérapique, chimiothérapiques, hormono-chimiothérapique, physiothérapie, polychimiothérapie, radio-chimiothérapie, radiothérapie, radiothérapique, thérapie, thérapies.

Tableau 5 : Exemples de familles morphologiques obtenues à partir d'un lexique d'apprentissage de 100 mots.

4 Discussion et conclusion

L'algorithme que nous avons mis au point permet de traiter à la fois les procédés de flexion ("pathologique", "pathologiques"), dérivation ("pathologie", "pathologique") et composition ("anatomopathologique"). De plus, il est possible de repérer les préfixes et les suffixes au cours de la même procédure, grâce à l'alignement des segments de mots à partir d'une base qui peut se situer à n'importe quelle position dans le mot.

La taille du corpus a été volontairement limitée afin de permettre une validation manuelle des mots reliés par dérivation ou composition. Les méthodes d'apprentissage sur corpus sont généralement dépendantes de la taille du corpus d'apprentissage : plus le corpus est important, meilleurs sont les résultats. Nous obtenons néanmoins un très bon pourcentage de relations valides (entre 98 et 71 %) avec uniquement 12 587 formes différentes (le français compte plusieurs centaines de milliers de formes différentes).

Nous n'avons pas directement mesuré le rappel. Nous en avons une indication indirecte à partir du nombre de relations découvertes et du rappel des relations flexionnelles du fichier DLF. Ce rappel est assez bon (de l'ordre des 60 à 70 %) et ce d'autant plus qu'un certain nombre de relations flexionnelles ne peuvent être retrouvées par un algorithme de segmentation morphologique (c'est le cas notamment des verbes irréguliers dont des formes très différentes peuvent être ramenées au même lemme; par exemple les formes "veut", "veux", "voudrez", "voudront", "voulant", "vouloir", "voulu" correspondent toutes au même lemme "vouloir"). D'une manière générale, les variantes des bases et les cas de doublement des consonnes constituent des pierres d'achoppement qui peuvent diminuer le rappel de l'algorithme. Ainsi, les mots "estrogènes" et "estrogénique" ont deux bases différentes : "estrogèn" et "estrogéniqu". De même "fractions" et "fractionnement" correspondent aux bases "fraction" et "fractionne". La méthode devrait être améliorée pour prendre ces cas en compte.

Il faut également noter que cette méthode n'est pas a priori limitée à l'analyse du français car elle ne repose sur aucune ressource externe au corpus. Nous avons d'ailleurs obtenu quelques exemples de segmentations de mots anglais, dans la mesure où ces mots contiennent des segments communs au français et à l'anglais (par exemple "radio + therapy"). Il reste à tester la méthode sur des corpus étrangers, notamment des corpus de langues agglutinantes telles que l'allemand, afin d'éprouver son efficacité. Des expériences supplémentaires devraient également être conduites sur des corpus français généralistes et de taille plus importante.

La méthode de segmentation finale des mots décrite utilise une structure de mot simplifiée : (préfixe*) + base + (suffixe*). Nous obtenons par exemple les segmentations suivantes : "lymphangi + osarcome", "lymphangi + te" et "ostéo + sarcom + e". Il est intéressant de noter que l'interfixe "o" reliant "lymphangi" et "sarcom" a bien été repéré. Cependant, seule une base, "lymphangi", a été identifiée, "osarcome" étant considéré comme un suffixe. Il serait souhaitable d'affiner l'algorithme sur ce point afin de repérer plusieurs bases par mot, comme c'est souvent le cas pour les termes spécialisés. Le repérage de plusieurs bases par mot permettrait de mettre le mot "lymphangiosarcome" en relation à la fois avec "lymphangite", "lymphe" ou "ostéosarcome" par le biais des bases "lymph" et "sarcom".

De plus, à l'heure actuelle, l'algorithme utilise uniquement des critères formels pour procéder à la segmentation morphologique des mots. Or les relations morphologiques impliquent à la fois forme et sens. Il paraît donc nécessaire d'intégrer des informations d'ordre sémantique à une méthode de segmentation morphologique. La prise en compte d'informations sémantiques

peut se faire a priori en effectuant l'apprentissage à partir de mots sémantiquement reliés telles que les listes de termes synonymes issus d'une terminologie (Zweigenbaum, Grabar, 2000) ou a posteriori en utilisant des mesures de distance sémantique basées sur la distribution des mots dans le corpus pour valider les segmentations (Schone, Jurafsky, 2001; Zweigenbaum et al., 2003). Nous envisageons de compléter l'algorithme avec des informations de ce type afin d'en augmenter la précision.

Références

Creutz M., Lagus K. (2002), Unsupervised Discovery of Morphemes, *Proceedings of the 6th Meeting of the ACL Special Interest Group of Computaional Phonology (SIGPHON)*, 21-30.

Déjean H. (1998), Morphemes as Necessay Concept for Structures Discovery from Untagged Corpora, *NeMLaP3/CoNLL98 Workshop on Paradigms and Grounding in Language Learning*, 295-298.

Gaussier E. (1999), Unsupervised Learning of Derivational Morphology from Inflectional Lexicons, *Proceedings of the ACL Workshop on Unsupervised Learning in Natural Language Processing*, 24-30.

Goldsmith J. (2001), Unsupervised Learning of the Morphology of a Natural Language, *Computational Linguistics*, Vol. 27(2), pp. 153-198.

Hafer M.A, Weiss S.F. (1974), Word Segmentation by Letter Successor Varieties, *Information Storage and Retrieval*, Vol. 10, pp. 371-385.

Hahn U., Honeck M., Shulz S. (2003), Subword-Based Text Retrieval, *Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03)*, 108.1.

Harris Z. (1955), From Phoneme to Morpheme, *Language*, Vol. 31, pp. 190-222.

Lovis C., Michel PA., Baud, R., Scherrer JR.. (1995), Word Segmentation Processing: A Way to Exponentially Extend Medical Dictionaries, *Proceedings of the 8th World Congress on Medical Informatics*, 28-32.

Schone P., Jurafsky D. (2001), Knowledge-Free Induction of Inflectional Morphologies, *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL-2001)*, 183-191.

Silberztein M. (1993), *Dictionnaires électroniques et analyse automatique de textes : le système INTEX*, Paris, Masson.

Vergne J. (2003), Un outil d'extraction terminologique endogène et multilingue, Actes de *TALN 2003*, 139-148.

Zweigenbaum P., Grabar N. (2000), Liens morphologiques et structuration de terminologie, Actes de *IC 2000 : Ingénierie des Connaissances*, 325-334.

Zweigenbaum P., Hadouche F., Grabar N. (2003), Apprentissage de relations morphologiques en corpus, Actes de *TALN 2003*, 285-294.

Traduction de règles de construction des mots pour résoudre les problèmes d'incomplétude lexicale en traduction automatique Étude de cas

Bruno Cartoni

TIM/ISSCO – ETI – Université de Genève
40 bd du Pont-d'Arve, CH-1205 Genève
bruno.cartoni@eti.unige.ch

Mots-clefs – Keywords

Traduction automatique, morphologie constructionnelle, incomplétude lexicale

Machine Translation, constructional morphology, lexical incompleteness

Résumé – Abstract

Cet article propose d'exploiter les similitudes constructionnelles de deux langues morphologiquement proches (le français et l'italien), pour créer des règles de construction des mots capables de déconstruire un néologisme construit de la langue source et générer de manière similaire un néologisme construit dans la langue cible. Nous commençons par présenter diverses motivations à cette méthode, puis détaillons une expérience pour laquelle plusieurs règles de transfert ont été créées et appliquées à un ensemble de néologismes construits.

This paper presents a method which aims at exploiting constructional similarities between two morphologically-related languages (French an Italian), in order to create word-construction rules that can disassemble a constructed neologism and create in a similar way a constructed neologism into the target language. We present the main motivation for this method and describe an experiment for which transfer rules has been developed and applied to a group of constructed neologism.

1 Introduction

Le présent article s'inscrit dans un travail de recherche qui vise à exploiter les propriétés constructionnelles des mots néologiques construits pour résoudre l'incomplétude lexicale en traduction automatique (ci-après TA). Nous entendons exploiter ces propriétés à la fois pour l'analyse des mots inconnus et pour la génération d'une traduction possible de ces mots.

Cependant, une telle exploitation doit se faire au travers de moyens simples pour garantir la « portabilité » et une certaine efficacité indispensables à un tel système.

Deux hypothèses guident ce travail : premièrement, nous émettons l'idée que les constructions des néologismes sont suffisamment transparentes et peu ambiguës pour être analysées par des moyens simples nécessitant peu de ressources. Deuxièmement, nous pensons que les processus morphologiques sont suffisamment proches d'une langue à l'autre (du moins dans des langues de même famille) pour permettre d'envisager la traduction d'un processus de construction par un autre. Pour ce faire, nous proposons l'utilisation de règles bilingues de construction des mots. Leur élaboration nécessite l'étude approfondie des processus morphologiques à traiter. Dans cette optique, nous décrivons ci-après les différentes motivations qui sous-tendent cette approche (section 2), puis nous proposons un modèle de traitement des mots construits inconnus, que nous appliquons ensuite à une expérience (section 3) dans laquelle nous mettons au point un ensemble de règles bilingues de construction des mots exploitables en TA.

2 Motivation

2.1 Les lexiques informatisés et la néologie formelle

De nombreuses applications de TAL reposent sur un lexique qui contient les mots de la langue traitée (ou un sous-ensemble de ces mots), ainsi que certaines informations associées à ces mots (Sproat, 1992). La couverture et la qualité du lexique dépendent de l'application pour laquelle le lexique est élaboré (Arnold, *et al.*, 1994). Or, l'absence d'un mot dans le lexique cause un certain nombre de problèmes, notamment en TA (Gdaniec, *et al.*, 2001). Les concepteurs de système de TA recourent alors à différentes stratégies pour faire face à l'incomplétude lexicale (comme la simple « transposition » du mot inconnu, sans le traduire). D'autres systèmes, comme Systran (Whitelock *et al.*, 1995), tentent au moins de deviner la catégorie grammaticale du mot inconnu en se basant sur sa terminaison. Ceci permet d'obtenir une analyse syntaxique plus correcte, même si la traduction n'en est pas plus réussie.

Les mots inconnus des lexiques informatisés posent donc un vrai problème. Parmi ces mots inconnus, on trouve un nombre important de noms propres. Toutefois, la majeure partie des mots inconnus a pour origine la créativité lexicale des langues. Celle-ci relève de plusieurs procédés, comme celui de la *néologie formelle* qui consiste en la création de nouveaux mots à partir de matériaux lexicaux préexistant (Gaudin *et al.*, 2000). D'un point de vue quantitatif, la néologie formelle est souvent décrite comme la plus productive et la plus utilisée. Ainsi, par exemple, Cabré (2002) constate que dans la presse catalane, ce procédé est à l'origine de 75 % des néologismes. Elle note également que au sein de la néologie formelle, la préfixation représente la ressource néologique la plus importante (32,3 %), suivie de près par la suffixation (24,7 %).

2.2 Le traitement de la morphologie constructionnelle

La préfixation et la suffixation, source principale de la néologie formelle, font partie des sujets d'étude de la morphologie constructionnelle, qui étudie la formation des mots construits, c'est-à-dire les mots dont le sens est prédictible et entièrement compositionnel par rapport à leur structure interne (Corbin, 1987). La formalisation d'un processus constructionnel passe par l'écriture de règles de construction des mots (Corbin, *ibid.*) (ci-après RCM) soumises à des contraintes permettant de décrire le processus concerné.

La morphologie constructionnelle reste souvent décrite comme irrégulière, et donc difficilement généralisable, et partant non exploitable en TAL, même si un récent article (Dal 2002) remet en cause cette réputation d'irrégularité. Comme le soulignent Dal, *et al.*, (sous presse), le TAL s'intéresse peu à la morphologie constructionnelle, sans doute « parce qu'il voit [dans les données constructionnelles] des phénomènes imprévisibles qui échappent en grande partie au calcul ». Dans de nombreux systèmes de traitement des langues, les informations constructionnelles sont utilisées à la seule fin d'étiquetage morphosyntaxique (Dal, *et al., ibid.*). En TA, l'utilisation des propriétés de la morphologie constructionnelle est rare, sans doute parce que ces applications nécessitent également une opération de génération, qui semble plus difficile pour les mots construits inconnus (Gdaniec, *et al.*, 2001).

Parce que la morphologie constructionnelle est souvent à l'origine de la formation de nouveaux mots, certaines recherches se sont déjà penchées sur le traitement automatique de la néologie formelle, principalement dans des buts de génération (voir par exemple Namer *et al*, 2000). Pour le présent travail, nous partons de l'hypothèse que les néologismes formels doivent être sémantiquement transparents pour être compris des locuteurs. Ainsi, la construction de néologismes ne devrait pas être aussi imprévisible et irrégulière que la morphologie constructionnelle en général, ce qui devrait faciliter le traitement informatique.

Avant de décrire une expérience pratique utilisant certaines propriétés de la morphologie constructionnelle, nous étudions les possibilités de transfert de ces RCM d'une langue à une autre, de façon à pouvoir exploiter ces propriétés dans le traitement des néologismes en TA.

2.3 La traduction des règles de construction des mots

La deuxième hypothèse qui sous-tend notre approche concerne l'exploitation des similitudes morphosémantiques entre deux langues proches pour inférer (ou deviner) l'équivalent d'un néologisme dans l'autre langue. Nous partons du principe que la proximité des lexiques de deux langues proches d'un point de vue morphologique pourrait être exploitée en TA. Ainsi, deux langues morphologiquement proches possèdent des similitudes au niveau des procédés de construction morphologique. Certaines études ont déjà montré la relative proximité entre le français et l'italien (Namer, 2001). Cependant, même si ces deux langues possèdent le même « fonds lexical commun », des divergences se sont développées avec le temps, donnant notamment lieu au phénomène des faux amis. La production néologique semble néanmoins s'effectuer d'une manière plus régulière et avec une relative similitude dans les différentes langues, particulièrement dans les domaines scientifiques et techniques où la mondialisation et les échanges internationaux sont importants, influençant par là même leur vocabulaire. Dans ces domaines, les emprunts ou calques de constructions morphologiques s'effectuent en

mettant la nouvelle unité lexicale à la « sauce » morphologique de la langue empruntante (Gaudin *et al.*, 2000).

Cette hypothèse soulève quelques questions théoriques auxquelles il est difficile d'apporter une réponse franche, au moins à ce stade de nos recherches. Par exemple, même si tout locuteur de deux langues morphologiquement proches sent très bien que le préfixe de réitération italien *ri–* peut être traduit par *re–* en français (comme l'attestent les paires *ricominciare*$_{it}$/*recommencer*$_{fr}$, *rifare*$_{it}$/*refaire*$_{fr}$, etc.), peut-on pour autant considérer que l'utilisation de ces deux procédés est régulière, et donc exploitable en TA ? Et, s'il existe des exceptions, comment les définir pour éviter des analyses ou des générations incorrectes ? Ces questions restent ouvertes, mais l'expérience ci-dessous propose quelques pistes de recherche pour découvrir les régularités morphosémantiques entre deux langues et ainsi extraire des règles bilingues de construction des mots.

3 Expérience

L'exploitation des régularités constructionnelles pourrait donc apporter une solution à certains problèmes d'incomplétude lexicale en TA. Ainsi, dans un cadre plus large, nous proposons un système de transfert d'information morphosémantiques d'une langue à l'autre de façon à générer un néologisme construit en langue cible à partir des informations reconnues dans le néologisme de la langue source.

En présence d'un mot inconnu du lexique d'un système de TA, notre modèle propose donc les étapes suivantes : **(1)** analyse des néologismes selon des RCM ; **(2)** traduction de la base des mots construits analysés en (1) ; **(3)** transfert des informations morphosémantiques ; **(4)** construction d'un néologisme construit grâce à la base traduite en (2) et aux informations provenant de l'étape (3). Ce « transfert » des informations (réalisé en 3) nécessite la mise en correspondance des RCM du français et de l'italien, un peu à l'image des règles de transfert syntaxique de certains systèmes de TA. Dans notre approche, cette mise en correspondance est formalisée par un ensemble de RCM bilingues permettant de décrire différents procédés de construction.

Cependant, même si nos recherches visent, à terme, une exploitation à large échelle dans un système de TA, nous concentrons pour l'instant nos efforts sur des unités lexicales hors contexte. Dans ce même esprit, nous limitons les ressources d'analyse à celles contenues dans un système de TA. Il serait, en effet, peu approprié de proposer des méthodes d'analyse des mots inconnus qui utilisent des procédés gourmands en place et en connaissances, et qui péjorerait ainsi les performances du système de TA en termes de vitesse et de taille. Nous présentons ci-après les différentes étapes de l'élaboration de telles règles, et leurs exploitations *in vivo*.

3.1 Le système dérivationnel choisi

Pour cet article, nous nous sommes limité à l'étude d'un système dérivationnel présent dans nos deux langues de travail, l'italien et le français.

En italien, il existe un préfixe verbal *r(i)–* (parfois sous la forme de *re–,*) qui correspond à la forme syntaxique de « de nouveau » (Dardano, 1978). Ce préfixe est un des plus productif, étant donné qu'il peut potentiellement être associé à tous les verbes (Dardano, *ibid.*). Cette productivité semble particulièrement intéressante pour ce travail. En effet, dans la mesure où ce préfixe peut potentiellement être associé à tous les verbes, il pose un réel problème d'exhaustivité pour les lexiques informatisés. La RCM bilingue qui le traitera peut donc s'avérer très rentable en termes de gain en qualité de traduction.

Dardano (*ibid.*) associe ce procédé de préfixation à celui du préfixe *de–*, en émettant l'hypothèse qu'une action qui doit être **re-**faite, a dû être auparavant **dé-**faite. Il propose alors une série paradigmatique comme *stabilizzare* → ***de****stabilizzare* → ***ri****stabilizzare*. Il note également que cette série possède la capacité de produire une série de noms déverbaux (comme dans *stabilizzazione* → *destabilizzazione* → *ristabilizzazione*).

Pour le français, un système dérivationnel équivalent semble exister. Le préfixe *re–* (et ses allomorphes, *r–* , et *ré–*) signifie également la répétition de l'action décrite par la base verbale (Rey-Debove, 2004). Le lien souligné pour l'italien par Dardano (1978) entre les préfixes *r(i)–* et *de–* est également présent dans le système morphologique du français (Huot, 2001), avec le préfixe *dé–* (qui prend aussi les formes dés– et des–), comme l'attestent les paires ***dé****faire/****re****faire*, ou ***dé****coudre/****re****coudre*. Enfin, les séries de noms déverbaux cités par Dardano (*ibid*) (*stabilizzazione* → *destabilizzazione* → *ristabilizzazione*) se retrouvent en français : *stabilisation* → *déstabilisation* → *restabilisation*.

3.2 Élaboration des règles de construction des mots bilingues

La linguistique informatique s'inspire largement de la linguistique descriptive ou de la linguistique théorique, mais elle doit bien souvent faire des compromis liés aux problèmes d'implantation (Tzoukermann *et al.*, 1997). Dans notre cas, ce sont les contraintes liées à la portabilité et à la rapidité de traitement des systèmes de TA qui guident notre approche. Ainsi, nous implémentons des RCM bilingues basées uniquement sur les chaînes de caractères et à l'aide d'expressions régulières.

La RCM bilingue présentée dans la Figure 1 porte sur le traitement de verbes préfixés en *ri–* en italien et leur traduction en français. Les contraintes portent sur les bases, qui doivent appartenir au lexique de référence (L_{it} et L_{fr}) et qui doivent être la traduction l'une de l'autre. Une règle identique a été créée pour le préfixe *de–*. De plus, étant donné que nous ne travaillons que sur la forme orthographique des mots, les règles sont répétées pour chaque allomorphe (*r–*, *re–*, …). Notons également que faute de place, nous ne mentionnons pas dans les règles les changements morphographémiques résultant de la concaténation de la base et du préfixe qui ont été définies pour la partie française.

$$\text{IT} \begin{pmatrix} \text{X/VERBE} \Rightarrow \textbf{ri}\text{/PREF [Y/VERBE]} \\ \text{Y/VERBE} \in L_{it} \end{pmatrix} = \text{FR} \begin{pmatrix} \text{X'/VERBE} \Rightarrow \textbf{re}\text{/PREF [Y'/VERBE]} \\ \text{Y/VERBE} \in L_{fr} \end{pmatrix}$$

où : Y/VERBE = Y'/VERBE (équivalent de traduction)

Figure 1: RCM bilingue pour le préfixe italien *ri–*

La figure 2 présente la RCM bilingue des noms déverbaux préfixés par *ri–*, procédé étudié par Dardano (1978). Pour contraindre la règle, nous choisissons les suffixes les plus fréquents des noms déverbaux (–zione, –mento, –aggio). Une règle similaire est également construite pour le préfixe *de–* suivi d'un nom.

$$\mathbf{IT}\begin{pmatrix} \text{X/NOM => } \mathbf{ri}\text{/PREF [Y/NOM]} \\ \\ \text{Y/NOM = [a-z]*zione/i |} \\ \text{[a-z]*mento/i | [a-z]*aggio/i} \\ \text{Y/NOM} \in L_{it} \end{pmatrix} = \mathbf{FR}\begin{pmatrix} \text{X'/NOM => } \mathbf{re}\text{/PREF [Y'/NOM]} \\ \\ \text{Y/NOM = [a-z]*tion/s |} \\ \text{[a-z]*ment/s | [a-z]*age/s} \\ \text{Y'/NOM} \in L_{fr} \end{pmatrix}$$

où : Y/NOM = Y'/NOM (équivalent de traduction)

Figure 2 : RCM bilingue pour le préfixe italien *ri–* sur une base nominale

Si l'on ne peut, à ce stade, affirmer avec certitude que la préfixation français en *re-* est la traduction de la construction italienne en *ri–*, il n'en reste pas moins que notre expérience donne des résultats encourageants, que nous présentons dans la suite.

3.3 Le corpus de mots inconnus

La presse écrite étant un terrain particulièrement fertile pour la production néologique (Pruvost *et al.*, 2003), nous avons utilisé comme corpus textuel un recueil de textes italiens publié par ELRA, (corpus MLCC, 1997), contenant les éditions du mois de février 1992 du quotidien italien *Il Sole 24 ore*[1]. Ce corpus contient 1,88 millions d'occurrences.

Dans l'étude empirique des phénomènes de néologie, la découverte de nouveaux mots ne peut se faire qu'à partir d'un lexique de référence, tant la notion de nouveauté est difficile à définir. Pour obtenir une liste de néologismes, nous avons donc confronté le corpus à un lexique de référence qui a joué le rôle de corpus d'exclusion (Gaudin *et al.*, 2000). Ce lexique est celui d'un analyseur morphosyntaxique, qui entre dans un processus complet d'étiquetage morphosyntaxique (*Tatoo*[2]). De cette première confrontation, nous obtenons une liste de 225 075 unités lexicales inconnues de notre lexique de référence, ce qui correspond à environ 12 % du nombre total d'occurrences. Evidemment, ces mots inconnus ne sont pas tous des néologismes. Nous affinons notre liste en excluant les noms propres, qui représentent une part importante de l'incomplétude lexicale en TA (cf. plus haut). Pour ce faire, nous appliquons une simple routine basée sur les majuscules (à l'instar de Maurel, 2004), et nous obtenons un nombre total de mots inconnus potentiellement néologiques de 90 260 occurrences (environ 4,8 % du corpus).

3.4 Résultats et évaluations

Nous avons donc appliqué à notre corpus de mots inconnus les RCM bilingues proposées plus haut, à la fois pour analyser les mots inconnus construits selon le système dérivationnel

[1] http ://www.ilsole24ore.com/
[2] The ISSCO Tagger Tool : http://issco-www.unige.ch/staff/robert/tatoo/tatoo.html

choisi, et pour générer la traduction en français de ces néologismes construits. Nous présentons ci-dessous les résultats des étapes d'analyse et de génération. Notons que la traduction des bases a été effectuée par le système de traduction automatique Systransoft©[3].

3.4.1 Analyse des mots préfixés

Cette première étape consiste analyser les mots construits en utilisant la partie italienne de la RCM bilingue, en recensant semi-automatiquement les dérivés dans notre corpus de mots inconnus. La partie automatique consiste à rechercher les mots qui correspondent aux règles établies plus haut, avec la contrainte que la base soit présente dans le lexique de référence. De cette extraction automatique, nous obtenons en sortie une paire formée du dérivé néologique et de sa base, elle-même accompagnée de son analyse dans le lexique (`"riorganizzare" = "organizzare" Verb […])`. Etant donné que cette automatisation peut générer une certaine quantité de bruit, dû au fait que l'analyse s'effectue uniquement sur la forme orthographique du mot, nous vérifions ensuite manuellement les couples dérivé/base extraits automatiquement. Le tableau 1 présente les résultats de l'analyse des préfixes verbaux (et de leurs allomorphes), accompagnés du nombre de formes « lemmatisées ». Par « lemme », il faut ici entendre la forme lemmatisée de la base du dérivé.

Préfixes verbaux	Occurrences	Lemmes	Erreur
ri–	508	63	5
r–	37	4	2
re–	96	9	0
de-	36	10	6

Tableau 1: analyse des verbes préfixés

Le nombre d'erreurs est calculé sur les formes lemmatisées. Ces erreurs correspondent à des cas où le sens du mot construit n'est pas compositionnel, comme dans **debuttare* = *de* + *buttare*[4]. Elles peuvent paraître fréquentes (13 cas sur 86) mais elles concernent uniquement des formes qui ne sont pas néologiques et qui constituent donc d'avantage des lacunes du lexique de référence utilisé pour cette expérience. D'ailleurs, tous les mots décomposés incorrectement sont connus des systèmes de TA classiques, qui reposent sur des lexiques plus exhaustifs.

Concernant les noms dérivés déverbaux, nous procédons de la même manière avec la RCM bilingue correspondante. Le tableau 2 propose un résumé des résultats de cette extraction et son évaluation. (Pour des raisons de place, nous ne mentionnons pas les formes qui n'ont donné aucun résultat – les noms préfixés avec l'allomorphe *r–* et le suffixe déverbal *–aggio* pour les trois préfixes).

---

[3] SYSTRAN S.A. http://www.systranet.com/systran/net (21 janvier 2005)

[4] ce qui correspondrait en français à **débuter* = *dé* + *buter*

Préfixe verbal	Suffixe déverbal	Occurrences	Lemmes	Erreurs
ri–	*–zione*	119	17	1
	–mento	201	10	0
re–	*–zione*	19	4	0
	–mento	16	1	0
de–	*–zione*	47	11	0
	–mento	0	0	0

Tableau 2: analyse des noms déverbaux préfixés

La vérification manuelle permet de trouver un taux de bruit presque nul. Une seule erreur où le sens du mot construit n'est pas compositionnel a été trouvée, et elle provient d'une erreur de typographie dans le corpus de départ.

De ces deux petites règles, pour les noms et les verbes, nous voyons qu'un nombre important d'unités lexicales peut être analysé d'une manière correcte d'un point de vue constructionnel (1065). Un tel résultat provient sans doute de plusieurs facteurs. Premièrement, les règles ont été soumises à des contraintes (simples mais importantes), ce qui évite un nombre important de mauvaises analyses (qui pourraient être dues par exemple à l'homographie des affixes). Mais le facteur le plus important est sans doute que les mots analysés sont uniquement néologiques (les erreurs ne se retrouvent que dans les mots non néologiques) et que l'analyse s'effectue en fonction du lexique de référence. Tout ceci tend à montrer que les mots construits néologiques sont rarement ambigus, ce qui vérifie, pour ce corpus en tout cas, notre première hypothèse.

Précisons cependant que, même si le bruit est quasiment inexistant, le silence serait également à évaluer. Toutefois, comme la contrainte la plus importante mise sur la règle concerne la présence de la base dans le lexique de référence, les mots non trouvés (le silence) sont forcément des mots dont la base est absente du lexique, et donc « intraitables » par le reste du processus de traduction.

3.4.2 Traduction des mots analysés

Après avoir analysé les mots construits et traduit les bases, nous avons « reconstruit » les équivalents de traduction en appliquant la partie française des RCM bilingues. Nous obtenons alors une liste de mots construits italiens et leur équivalent de traduction, comme *deresponsabilizzazione* = *déresponsabilisation*, *ridistribuzione* = *redistribution*, *riclassificare* = *reclassifier*. Evaluer la correction de la traduction ne nous semblait pas très pertinent, car une bonne traduction dépend également du contexte que nous n'avions pas. Nous avons donc évalué uniquement la correction du néologisme français. Cette évaluation doit se faire en fonction de la correction du néologisme par rapport à un sentiment linguistique propre à la langue française. Mais le sentiment linguistique reste un concept très flou. Les mots ont donc

été évalués sur une échelle de trois valeurs : correct, incorrect, incertain. Cependant, aucun mot n'a été jugé « incorrect ». Les résultats de l'évaluation sont résumés dans le tableau 3.

Procédés constructionnels	Lemmes	néologismes corrects	néologismes incertains
ri– + verbe	58	56	2
r– + verbe	2	2	0
re– + verbe	9	9	0
de– + verbe	4	4	0
ri– + nom	26	22	4
re– + nom	5	5	0
de– + nom	11	11	0

Tableau 3 : évaluation des néologismes en fonction du préfixe

Concernant les néologismes jugés incertains, nous avons utilisé le corpus du Web pour affiner encore l'évaluation. Ainsi, bien que des mots comme *recrocheter* ou *réassociation* aient paru étranges aux évaluateurs, le nombre d'occurrences et la pertinence des sources découvertes sur l'Internet permettent de valider ces créations dans tous les cas.

4 Conclusion et travaux futurs

Le petit nombre de mots étudiés dans cette expérience ne peut nous permettre de tirer des conclusions définitives et une expérience similaire sur d'autres corpus permettrait de confirmer les premières tendances constatées. Il n'en reste pas moins que cette expérience semble montrer que des règles de construction simples et précises donnent des résultats extrêmement fiables. De plus, le faible bruit provenant de l'analyse compositionnelle des néologismes est très prometteur.

D'un point de vue quantitatif, l'ensemble des règles proposées dans cet article a permis de traduire 1065 occurrences. Dans l'absolu, ce chiffre peut paraître faible en comparaison des 90 260 mots inconnus dans le corpus. Cependant, si la pertinence d'une telle règle est validée à plus large échelle et si l'on y ajoute d'autres règles similaires pour d'autres processus de construction réputés productifs, nous profiterions alors d'un mécanisme simple et moins chronophage en terme de gestion des ressources linguistiques pour la TA.

La validité de ces règles reste néanmoins à prouver, notamment pour d'autres processus constructionnels moins transparents et beaucoup moins décrit par la linguistique théorique. Et tout comme pour les règles syntaxiques dans le domaine de la TA, le traitement des divergences entre les deux langues (on dit *clonage* en français, mais *clonazione* en italien) est aussi un large objet d'étude.

Référence

Arnold D., Balkan L., Humphreys R., Meijer S., Sadler , (1994). *Machine translation: An introductory Guide,* Manchester, Blackwell.

Cabré T., Freixa, J., Solé E., (2002), A la limite des mots construits possible, Actes du *Forum de morphologie*, pp. 65-78.

Corbin D., (1987), *Morphologie dérivationnelle et structuration du lexique*, Tuebingen, Niemeyer.

Dal G. (2002). A propos d'une idée reçue, ou de la prétendue irrégularité de la dérivation, *Bulag,* Vol. 27, pp. 57-73.

Dal G., Hathout, N., Namer F., (*sous presse*), Morphologie Constructionnelle et Traitement Automatique des Langues : le projet MorTAL*, Lexique* Vol.16.

Dardano M., (1978), *La formazione delle parole nell'italiano di oggi*, Rome, Bulzoni.

Gaudin F., Guespin L., (2000), *Initiation à la lexicologie française*, Bruxelles, Duculot.

Gdaniec C., Manandise, E., McCord, M., (2001), Derivational Morphology to the Rescue: How It Can Help Resolve Unfound Words in MT. Actes de *MT Summit VIII*.

Huot, H., (2001). *Morphologie, Forme et sens des mots français*. Paris, Armand Colin.

Maurel, D. (2004). Les mots inconnus sont-ils des noms propres? Actes de *JADT 2004*, Louvain-la-Neuve.

Namer, F. (2001), Génération automatique de néologismes bilingues morphologiquement construits en français et en italien. Actes de *TALN 2001*. pp. 281-296.

Namer, F. (2000), GéDériF : Automatic Generation and Analysis of Morphologically Constructed Lexical Resources, Actes de *LREC 2000*, pp. 1447-1454

Rey-Debove J., Ed. (2004). *Brio*, Paris, Dictionnaire Le Robert.

Sproat R. (1992), *Morphology and Computation*. Cambridge, The MIT Press.

Tzoukermann E., Jacquemin, C. (1997), Analyse automatique de la morphologie dérivationnelle et filtrage des mots possibles, Actes de *Forum de morphologie*, pp. 251-260.

Whitelock P., Kilby K, (1995) *Linguistic and computational techniques in machine translation system design,* London, UCL Press.

A la découverte de la polysémie des spécificités du français technique

Ann Bertels

ILT – K.U.Leuven
Dekenstraat 6 – B-3000 LEUVEN (Belgique)
ann.bertels@ilt.kuleuven.ac.be

Mots-clefs – Keywords

sémantique lexicale, langue spécialisée, spécificités, polysémie, cooccurrences

lexical semantics, language for specific purposes (LSP), keywords, polysemy, co-occurrences

Résumé – Abstract

Cet article décrit l'analyse sémantique des spécificités dans le domaine technique des machines-outils pour l'usinage des métaux. Le but de cette étude est de vérifier si et dans quelle mesure les spécificités dans ce domaine sont monosémiques ou polysémiques. Les spécificités (situées dans un continuum de spécificité) seront identifiées avec la *KeyWords Method* en comparant le corpus d'analyse à un corpus de référence. Elles feront ensuite l'objet d'une analyse sémantique automatisée à partir du recouvrement des cooccurrences des cooccurrences, afin d'établir le continuum de monosémie. Les travaux de recherche étant en cours, nous présenterons des résultats préliminaires de cette double analyse.

This article discusses a semantic analysis of pivotal terms (keywords) in the domain of machining terminology in French. Building on corpus data, the investigation attempts to find out whether, and to what extent, the keywords are polysemous. In order to identify the most typical words of the typicality continuum, the KeyWords Method will be used to compare the technical corpus with a reference corpus. The monosemy continuum will be implemented in terms of degree of overlap between the co-occurrents of the co-occurrents of the keywords. We present some preliminary results of work in progress.

1 Introduction et question de recherche

Cet article s'inscrit dans le cadre d'une thèse de doctorat sur la sémantique du vocabulaire spécifique d'un corpus de français technique. Comme le corpus d'analyse relève du domaine

technique des machines-outils pour l'usinage des métaux, l'analyse sémantique porte sur les spécificités[1] d'une langue spécialisée.

Dans la langue spécialisée, les besoins communicatifs requièrent plus de précision, ce que la terminologie traditionnelle définit comme l'univocité, la monoréférentialité et la monosémie des unités terminologiques de la langue spécialisée. La terminologie traditionnelle prescriptive et normative adopte une approche onomasiologique par domaine. Récemment, la monosémie et l'univocité de la langue spécialisée ont été remises en question par la Théorie Communicative de la Terminologie (Cabré, 1998, 2000), par la socioterminologie (Gaudin, 1993) et par la terminologie socio-cognitive (Temmerman, 1997). Les termes font partie intégrante de la langue naturelle, mais véhiculent des connaissances spécialisées (Lerat, 1995). Les partisans de la terminologie descriptive rejettent la dichotomie entre la langue générale et la langue spécialisée et adoptent une approche sémasiologique et linguistique, basée sur l'étude de corpus de textes spécialisés (Condamines & Rebeyrolles, 1997).

Pour quantifier la thèse monosémiste de la terminologie traditionnelle, nous nous proposons de la reformuler en une question de recherche opérationnelle et mesurable : « Y a-t-il une corrélation entre, d'une part, le continuum de spécificité et, d'autre part, le continuum de monosémie (continuum de sens) ? » L'hypothèse de recherche avancée pose que, contrairement à la thèse traditionnelle, les mots (les plus) spécifiques du corpus technique ne sont pas nécessairement (les plus) monosémiques. L'analyse se propose donc de vérifier la polysémie des mots du corpus technique d'analyse, p.ex. le mot *broche* (1) « partie tournante d'une machine-outil qui porte un outil ou une pièce à usiner » et (2) « outil servant à usiner des pièces métalliques ». A cet effet, ces mots sont ordonnés en fonction de leur spécificité et situés sur une échelle de spécificité allant des mots les plus spécifiques aux mots les moins spécifiques, mais comprenant toujours des spécificités statistiquement significatives du corpus technique. Un deuxième classement situe les mêmes mots sur une échelle de monosémie, à partir d'une analyse des cooccurrences de deuxième ordre, c'est-à-dire les cooccurrences des cooccurrences. La question de recherche principale (corrélation entre le degré de spécificité et le degré de monosémie) sera complétée par des questions de recherche secondaires faisant intervenir les facteurs influant sur le degré de monosémie, notamment la fréquence et la classe lexicale. Une analyse de régression multiple permettra de vérifier l'impact des variables indépendantes (spécificité, fréquence, classe lexicale, etc.) sur le degré de monosémie.

2 Corpus technique d'analyse et corpus de référence

Le corpus technique d'analyse est constitué de textes techniques du domaine des machines-outils pour l'usinage des métaux et comprend environ 1.760.000 mots. Le corpus a été étiqueté et lemmatisé par Cordial 7 Analyseur et consiste en 4 sous-corpus, datant de 1998 à 2001 : revues techniques électroniques (800.000) et fiches techniques (300.000) trouvées sur Internet, normes ISO et directives (300.000) et quatre manuels numérisés (360.000). Les textes se situent à différents niveaux de normalisation et de vulgarisation, s'adressant tant à des professionnels (revues et fiches) qu'à des étudiants (manuels). Afin de pouvoir déterminer

[1] Nous adoptons le terme « spécificités » pour désigner les mots les plus spécifiques et caractéristiques du corpus d'analyse, indépendamment de la méthode utilisée (calcul des spécificités vs. KeyWords Method).

les spécificités du corpus technique, il est complété par un corpus de référence de langue générale. Celui-ci est constitué d'articles journalistiques du journal Le Monde (janvier – septembre 1998). Il a également été lemmatisé et comprend environ 15.300.000 mots.

Les fichiers générés par Cordial se composent de trois colonnes, avec un mot par ligne: (1) la forme fléchie ou forme graphique, (2) le lemme ou forme canonique et (3) le code Cordial, comparable à un POS-tag (Part-Of-Speech) indiquant la classe lexicale. Dans les fichiers texte, nous avons corrigé quelques fautes de frappe. Les fichiers lemmatisés ont également fait l'objet d'un nettoyage, à savoir quelques regroupements (p.ex. lemmes avec et sans point *Fig./Fig* et lemmes avec et sans trait d'union) et la correction des erreurs de lemmatisation (p.ex. *machines-outils* sous le lemme *machine-outil*). Ces opérations de nettoyage ont été effectuées, tant pour le corpus d'analyse technique que pour le corpus de référence.

3 Approche méthodologique : spécificités et polysémie

Comme la recherche porte sur la question de savoir s'il y a une corrélation entre le continuum de spécificité et le continuum de monosémie, la réponse et l'analyse linguistique qui en découle requièrent une approche méthodologique double. Il faut d'une part le calcul des spécificités et d'autre part une mesure pour déterminer le degré de monosémie. Ces spécificités se situent, non seulement au niveau des unités simples, p.ex. *fraisage, commande*, mais également au niveau des unités polylexicales, p.ex. *commande numérique*.

3.1 Spécificités

La recherche en langue spécialisée prend généralement comme point de départ l'identification des spécificités, c'est-à-dire des mots spécifiques qui caractérisent le corpus spécialisé et qui le différencient d'un corpus de langue générale. Les spécificités ne sont pas les mots les plus fréquents de ce corpus, mais les mots les plus représentatifs. Du point de vue relatif, ces mots figurent de façon significative plus fréquemment dans le corpus de langue spécialisée que dans un corpus de langue générale. Afin de déterminer les spécificités, les fréquences dans le corpus spécialisé sont comparées aux fréquences dans un corpus de référence, compte tenu de la taille des deux corpus, ce qui revient à comparer la fréquence observée (corpus d'analyse) à la fréquence attendue (corpus de référence). S'il y a une différence entre la fréquence observée et la fréquence attendue, il faut vérifier si elle est statistiquement significative. A cet effet, deux méthodologies sont désormais disponibles : le calcul des spécificités (Lafon, 1984) implémenté dans le logiciel Lexico3[2], outils de statistique textuelle, et la *KeyWords Method* des logiciels WordSmith Tools[3] et Abundantia Verborum[4] (Speelman, 1997). Les deux méthodologies aboutissent grosso modo à des résultats similaires, à savoir une liste de mots spécifiques pourvus d'une mesure statistique indiquant le degré de spécificité. Les différences les plus importantes résident dans la méthodologie et la statistique sous-jacentes.

---

[2] Lexico3 : SYLED – CLA2T, Paris3 : http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/

[3] WordSmith Tools version 3 : http://www.lexically.net/wordsmith/ et http://www.oup.com

[4] Abundantia Verborum : http://wwwling.arts.kuleuven.ac.be/genling/abundant/obtain/

Premièrement, le calcul des spécificités (Lafon, 1984 et Labbé & Labbé, 1991) compare une section d'un corpus au corpus entier afin d'identifier le vocabulaire spécifique de cette section. La comparaison *partie-tout* permet de décider si la fréquence relative d'un mot dans la section est supérieure à ce qui serait attendu, en fonction de la fréquence relative dans le corpus entier. L'analyse statistique sous-jacente au calcul des spécificités utilise le test de Fisher Exact, basé sur les probabilités exactes de la distribution hypergéométrique. Fisher Exact est généralement utilisé pour un ensemble de données de taille modeste ($n \leq 20$). En outre, les factorielles de la formule pour le calcul de la probabilité de la fréquence observée (Lafon, 1984) mèneraient à des nombres astronomiques, si la formule était appliquée à un corpus de quelques milliers, voire des millions de mots. Pour remédier à ce problème, Lafon propose d'utiliser des logarithmes. Dès lors, le résultat du calcul x est à interpréter comme l'exposant de la base 10, d'où résulte la probabilité 10^x. Dans Lexico3, ce sont les exposants (résultats de la formule du calcul hypergéométrique) qui figurent dans la colonne du coefficient de spécificité. Les spécificités positives indiquent un suremploi dans la section analysée, tandis que les spécificités négatives signalent un sous-emploi. Le calcul des spécificités est surtout utilisé par la communauté francophone (Cf. Zimina, 2004 et Drouin, 2004).

La deuxième méthodologie permettant d'identifier les mots les plus spécifiques est surtout utilisée par des utilisateurs du logiciel WordSmith–KeyWords (Berber-Sardinha, 1999). Elle est couramment appelée *KeyWords Method* ou méthode des mots-clefs. Les fréquences dans le corpus spécialisé sont comparées aux fréquences dans un corpus de référence de langue générale, compte tenu de la taille des deux corpus, ce qui permet d'identifier les mots significativement plus fréquents dans le corpus spécialisé. Il s'agit donc de la comparaison de deux corpus différents, et non d'une comparaison *partie-tout*. La deuxième différence entre les deux méthodologies réside dans la statistique sous-jacente. La *KeyWords Method* se sert du ratio du log de vraisemblance (*log likelihood ratio*) (Dunning,1993). Cette statistique de test n'est pas basée sur des probabilités exactes et par conséquent, elle s'applique facilement à des corpus plutôt volumineux. Le ratio du log de vraisemblance sera d'autant plus élevé que le mot est plus fréquent dans le corpus spécialisé par rapport au corpus de référence, indiquant dès lors son degré de spécificité. La valeur p correspondante permet de supprimer les spécificités statistiquement non significatives ($p \leq 0.05$). En plus, le tri des spécificités en fonction de la statistique de test (ratio du log de vraisemblance) permet de les classer par ordre de spécificité décroissante et par conséquent, de les situer dans un continuum de spécificité.

3.2 Polysémie

Les spécificités relevées dans le corpus technique d'analyse, font ensuite l'objet d'une analyse sémantique. Pour chaque spécificité, on déterminera le degré de monosémie, dans le but de vérifier si les mots les plus spécifiques sont en effet (les plus) monosémiques et si les mots moins spécifiques ont plus tendance à être polysémiques. Il s'agira donc d'objectiver et de quantifier l'analyse sémantique, en ayant recours aux cooccurrences. Selon Schütze (1998), Véronis (2003) et d'autres, les cooccurrences permettent de distinguer les différents usages et sens des mots. Audibert (2003) recourt même aux cooccurrences comme critères de désambiguïsation sémantique automatique. Dans Véronis (2003), les cooccurrences d'un mot, à partir d'un grand corpus, sont regroupées suivant leur similarité ou dissimilarité (en fonction de leur co-fréquence) pour identifier les différents sens du mot.

Afin de déterminer le degré de monosémie des spécificités, nous proposons d'aller plus loin et d'étudier les cooccurrences de deuxième ordre. Les cooccurrences des cooccurrences permettent de trouver des synonymes d'un mot, selon Martinez (2000). Pour le mot *mesures*, il trouve les cooccurrents *nouvelles*, *prises*, etc. qui cooccurrent à leur tour avec *décisions*. D'après Denhière & Lemaire (2003), les cooccurrences de deuxième ordre et même d'ordre supérieur déterminent le degré d'association de deux mots M1 et M2, même si ces deux mots ne figurent jamais ensemble. Si les cooccurrences M1-M3 et M2-M3 sont suffisamment fortes, on considère que M1 et M2 sont associés et des cooccurrents d'ordre 2. Il est également possible d'extraire automatiquement les sens des mots à partir d'un réseau de cooccurrences lexicales de deuxième ordre, comme l'explique Ferret (2004). La connectivité des cooccurrents formant un sens est plus importante que leur connectivité avec les autres cooccurrents définissant les autres sens de ce mot, la mesure de cohésion étant l'information mutuelle normalisée (Ferret, 2004).

Les cooccurrents de deuxième ordre étant des critères désambiguïsateurs puissants, ils seront très précieux lors de l'analyse sémantique des spécificités. En effet, le degré de recouvrement des cooccurrences de deuxième ordre sera un indice important du degré de monosémie du mot de base. Pour étudier le caractère monosémique ou polysémique d'une unité linguistique, on vérifie si les contextes peuvent être considérés comme sémantiquement homogènes ou non (Condamines & Rebeyrolles, 1997). L'accès à la sémantique des cooccurrences pourra se faire (automatiquement) par le biais des cooccurrences de deuxième ordre. Le degré de recouvrement de ces cooccurrences de deuxième ordre indiquera à quel point les cooccurrences de premier ordre (contextes du mot de base) sont sémantiquement homogènes.

- Si les cooccurrents des cooccurrents (« cc » ou cooccurrents de deuxième ordre) sont formellement très différents et se recouvrent très peu, les différents cooccurrents (« c » ou cooccurrents de premier ordre) seront sémantiquement plus diversifiés, une structure formelle de cooccurrence différente indiquant un sens différent. Les cooccurrents sémantiquement diversifiés appartenant à plusieurs champs sémantiques, la spécificité aura moins de chances d'être monosémique.

- Et inversement, plus les cooccurrences des cooccurrences se recouvrent, plus les cooccurrents sont sémantiquement homogènes. Le degré de ressemblance ou similarité lexicale des cooccurrents d'un mot étant proportionnel au degré de monosémie de ce mot, un fort recouvrement des cooccurrents de deuxième ordre signale un degré de monosémie plus important.

4 Premiers résultats de la recherche : spécificités et polysémie

4.1 Spécificités

La liste des spécificités du corpus technique d'analyse (1,7 million de mots) est générée avec les logiciels Abundantia Verborum et AV Frequency List Tool. Une expérimentation sur un échantillon restreint du corpus technique sur les trois logiciels disponibles pour le calcul des spécificités montre des résultats comparables en termes de spécificités relevées. Pour garantir la comparaison des résultats, le corpus technique a été incorporé dans le corpus de référence dans le logiciel Lexico3, procédant par comparaison *partie-tout*. Force est de constater que la

même procédure d'incorporation pour le corpus entier (corpus technique de 1,7 million et corpus de référence de 15,3 millions) n'aboutit pas aux résultats escomptés. Même si la liste de fréquence du grand corpus entier s'affiche, l'étape suivante du calcul des spécificités échoue en raison de la taille trop importante du corpus.

Appliquée au corpus technique lemmatisé, la *KeyWords Method* produit une liste d'environ 13.000 spécificités pour le corpus technique ($p \leq 0.05$). A l'aide des codes Cordial, une liste de mots grammaticaux (450) et une liste de noms propres (7200) sont générées, permettant de les supprimer. Les opérations de filtrage et de nettoyage génèrent une liste de spécificités techniques définitive de 7240 mots (lemmes), Cf. Figure 1, pour un aperçu des 25 mots les plus spécifiques. Ces 7240 mots feront l'objet de l'analyse sémantique automatisée pour établir le degré de monosémie. La première colonne (Cf. Figure 1) contient les lemmes spécifiques, les deuxième et troisième colonnes donnent la fréquence absolue dans le corpus technique (FREQ_ABS1) et dans le corpus de référence (FREQ_ABS 2). Dans la colonne 4, on voit la statistique de test LLR indiquant le degré de spécificité et dans la colonne 5, le complément de la valeur p correspondante (1-p). Les colonnes 6 et 7 affichent les fréquences relatives (multipliées par 10000) pour les deux corpus et la dernière colonne informe sur le type de spécificité (1 pour une spécificité positive et –1 pour une spécificité négative). Il est à remarquer que cette liste de spécificités contient aussi des mots de la langue générale (p.ex. *type, permettre*), spécifiques de ce corpus technique. Ce ne sont pas des termes, mais ils sont maintenus car nous nous proposons de comparer leur degré de monosémie à celui des termes (dans le corpus technique) et à leur degré de monosémie dans un corpus de langue générale.

LEMME	FREQ_ABS1	FREQ_ABS2	LLR	1-P	FREQ_REL1	FREQ_REL2	SPEC_POS
machine	12671	1052	50521,91	1	74,51	0,71	1
outil	8306	918	32037,72	1	48,84	0,62	1
usinage	6720	8	30468,41	1	39,52	0,01	1
pièce	7556	2219	24407,46	1	44,43	1,50	1
mm	5490	191	23357,57	1	32,28	0,13	1
vitesse	5283	900	19108,78	1	31,07	0,61	1
coupe	6730	4153	17063,37	1	39,58	2,80	1
broche	2893	12	13010,42	1	17,01	0,01	1
Fig	2680	0	12194,00	1	15,76	0,00	1
axe	3206	420	12079,16	1	18,85	0,28	1
copeau	2557	0	11634,18	1	15,04	0,00	1
plaquette	2407	35	10592,46	1	14,15	0,02	1
diamètre	2415	95	10200,09	1	14,20	0,06	1
commande	2751	850	8765,71	1	16,18	0,57	1
acier	2252	277	8558,49	1	13,24	0,19	1
fraisage	1873	0	8521,34	1	11,01	0,00	1
arête	1870	29	8213,91	1	11,00	0,02	1
précision	2263	541	7663,01	1	13,31	0,36	1
usiner	1577	11	7045,52	1	9,27	0,01	1
surface	2258	758	7037,02	1	13,28	0,51	1
type	2820	1830	6994,07	1	16,58	1,23	1
système	4052	5165	6915,85	1	23,83	3,48	1
fraise	1571	45	6745,88	1	9,24	0,03	1
gamme	1860	545	6006,35	1	10,94	0,37	1
permettre	4883	9504	5848,03	1	28,71	6,41	1

Figure 1 : Les 25 mots les plus spécifiques du corpus technique d'analyse

4.2 Polysémie

Le degré de monosémie (ou inversement de polysémie) dépend du degré de recouvrement des cooccurrents des cooccurrents. Afin d'établir les listes des cooccurrences pertinentes à deux niveaux et de générer une base de données qui sera interrogée pour le calcul automatique du degré de recouvrement, nous avons recours à un algorithme de scripts Python[5].

Pour déterminer le degré de monosémie, nous proposons une formule (Cf. Figure 2), basée sur le recouvrement des cooccurrents des cooccurrents (cc), en tenant compte (1) de la fréquence d'un cc dans la liste des cc (= nombre de cooccurrents (c) apparaissant avec ce cc), (2) du nombre total de c et (3) du nombre total de cc. La mesure d'association utilisée pour déterminer les cooccurrences pertinentes est la statistique LLR (log de vraisemblance). Nous ne prenons en considération que les cooccurrences statistiquement significatives ($p \leq 0.05$).

$$\sum_{cc} \frac{\text{fq}\, cc}{\#\,\text{total}\, c \cdot \#\,\text{total}\, cc}$$

Figure 2 : Formule de recouvrement des cooccurrents des cooccurrents

Dans un premier temps, nous dressons une liste des cooccurrences pertinentes à partir des fichiers lemmatisés du corpus technique, contenant le collocatif, la base[6] et leur co-fréquence. Deux autres fichiers sont dérivés de ces informations et contiennent respectivement les bases et les collocatifs et leurs fréquences. Toutes ces informations permettront de générer une base de données avec des informations statistiques, à savoir la statistique de test LLR et la valeur p. En fait, deux listes de cooccurrences avec leur base de données correspondante sont ainsi dressées : une première liste avec les spécificités comme base et leurs cooccurrents comme collocatif et une deuxième liste de cooccurrences avec les cooccurrents de la première liste comme base et leurs cooccurrents (d'ordre 2) comme collocatif.

Les paramètres modifiables sont le type de cooccurrent à relever (lemme ou forme fléchie) et la fenêtre d'observation. Nous optons pour une fenêtre de [-5,+5], 5 mots à gauche et 5 mots à droite, parce qu'elle apporte assez d'information sémantique, sans qu'il y ait trop de bruit et qu'elle permet un traitement informatique efficace. Au premier niveau d'analyse de la spécificité comme base, la base de la cooccurrence est nécessairement relevée sous forme lemmatisée, afin de pouvoir rattacher les informations sémantiques (degré de monosémie) aux informations de spécificité (liste de spécificités). Pour le collocatif, la forme fléchie s'impose, en raison des informations sémantiques plus riches qu'elle véhicule (Cf. différence entre *pièce à usiner* et *pièce usinée*). Comme ce collocatif est la base du deuxième niveau d'analyse, la forme fléchie s'impose à ce deuxième niveau tant pour la base que pour le collocatif.

Ces deux bases de données sont fusionnées en une grande base de données, interrogée pour l'analyse du recouvrement des cooccurrents de deuxième ordre. A cet effet, la fonction Python de l'algorithme prévoit les paramètres suivants : la base (spécificité à analyser), le

---

[5] http://www.python.org/

[6] la base (anglais : *node*) étant le mot étudié et le collocatif (anglais : *collocate*) étant un de ses cooccurrents

seuil de signification pour les cooccurrents de premier niveau (p.ex. 0.95 pour p≤0.05), le seuil pour les cooccurrents de deuxième niveau et la base de données. Il y a plus de recouvrement, si plus de cooccurrents (c) partagent le même cc, ce qui signifie un poids plus lourd pour ce cc (score près de 1). Un cc moins/pas partagé indique donc peu/pas de recouvrement (score près de 0).

La figure 3 ci-dessous montre les premiers résultats pour le corpus technique (1.7 million) et pour un seuil de signification des c et cc de p≤0.0001. Parmi les 25 mots les plus spécifiques du corpus technique entier, les 2 mots les plus spécifiques se caractérisent par le degré de monosémie le moins élevé (rangs 25 et 24), indiquant peu de recouvrement des cooccurrents d'ordre 2. Les mots en gras sont les moins monosémiques de cet échantillon, ce qui semble contredire la thèse de la corrélation[7] entre le degré de spécificité et le degré de monosémie.

LEMME	FREQ_ABS1	FREQ_ABS2	LLR	DEGRE DE MONOSEMIE	RANG DE MONOSEMIE
machine	12671	1052	50521,91	**0,0231**	25
outil	8306	918	32037,72	**0,0240**	24
usinage	6720	8	30468,41	0,0349	12
pièce	7556	2219	24407,46	0,0310	18
mm	5490	191	23357,57	0,0534	1
vitesse	5283	900	19108,78	0,0402	6
coupe	6730	4153	17063,37	0,0370	10
broche	2893	12	13010,42	0,0394	7
Fig	2680	0	12194,00	0,0483	3
axe	3206	420	12079,16	0,0340	13
copeau	2557	0	11634,18	**0,0299**	20
plaquette	2407	35	10592,46	**0,0282**	22
diamètre	2415	95	10200,09	0,0444	4
commande	2751	850	8765,71	0,0317	17
acier	2252	277	8558,49	**0,0282**	21
fraisage	1873	0	8521,34	0,0350	11
arête	1870	29	8213,91	0,0386	8
précision	2263	541	7663,01	0,0491	2
usiner	1577	11	7045,52	0,0406	5
surface	2258	758	7037,02	0,0321	15
type	2820	1830	6994,07	0,0372	9
système	4052	5165	6915,85	**0,0280**	23
fraise	1571	45	6745,88	0,0319	16
gamme	1860	545	6006,35	0,0324	14
permettre	4883	9504	5848,03	0,0303	19

Figure 3 : Degré et rang de monosémie des 25 mots les plus spécifiques (p ≤ 0.0001)

Pour les 100 mots les plus spécifiques, une première analyse de régression simple fait intervenir le degré de monosémie comme variable dépendante et le degré de spécificité comme variable indépendante ou prédictive. Elle montre qu'il y a une très faible corrélation négative (p=0.03) et que le degré de spécificité n'explique que 3.5% de la variation du degré

[7] Nous ne recourons pas à une simple mesure de corrélation, en raison de l'effet d'interférence attendu des autres facteurs (fréquence, classe lexicale, etc.).

de monosémie. Une analyse de régression multiple, mesurant l'impact de plusieurs variables indépendantes (spécificité, fréquence, classe lexicale, nombre de classes lexicales et longueur), indique comme facteurs significatifs le nombre de classes lexicales (p=0.008), la classe lexicale (p=0.03) et la fréquence (p=0.03) pour une variation totale expliquée de 12% (p=0.004). Parmi les 100 mots les plus spécifiques du corpus technique, les mots les plus polysémiques, affichant le degré de monosémie le plus bas, se caractérisent par une fréquence absolue élevée et par leur appartenance à deux ou plusieurs classes lexicales, principalement les classes 'nom' et 'adjectif'.

5 Conclusion et perspectives

Pour étudier la sémantique des spécificités dans le domaine technique des machines-outils pour l'usinage des métaux, nous avons eu recours à une double analyse. D'une part, la *KeyWords Method* a permis de dresser la liste des spécificités du corpus d'analyse, ordonnées par ordre de spécificité décroissante. D'autre part, la formule pour le recouvrement des cooccurrences des cooccurrences a permis d'accéder à la sémantique des spécificités, en évaluant le degré de monosémie. L'analyse détaillée des résultats de recherche nous apprendra pour un nombre important de mots techniques, s'il y a une corrélation entre leur degré de spécificité et leur degré de monosémie, en fonction d'une série de facteurs, notamment la classe lexicale et la fréquence.

La formule pour le recouvrement des cooccurrents de deuxième ordre et pour le degré de monosémie sera soumise à plusieurs expérimentations en fonction de plusieurs paramètres, afin de mettre au point le calcul du degré de monosémie. Une fois recueillies pour le corpus technique entier, les données sur le degré de monosémie permettront de situer les spécificités dans un continuum de monosémie (continuum de sens). Des analyses statistiques dans R[8] permettront ensuite de vérifier la corrélation entre le continuum de spécificité et le continuum de monosémie et de procéder à une analyse linguistique détaillée en fonction des variables (linguistiques) indépendantes.

Nous envisageons cette double analyse du degré de spécificité et du degré de monosémie pour les collocations spécifiques également, étant donné que les termes se situent souvent au niveau des unités polylexicales. Nous nous proposons aussi de procéder à une validation manuelle de la formule déterminant le degré de monosémie à l'aide de l'analyse des collocations et cooccurrences relevées.

Références

Audibert L. (2003), Etude des critères de désambiguïsation sémantique automatique : résultats sur les cooccurrences, Actes de *TALN 2003*, 35-44.

Berber Sardinha A. (1999), Word sets, keywords and text contents : An investigation of text topic on the computer, *DELTA*, 15-1, 141-149.

[8] http://www.r-project.org/

Cabré M.T. (1998), *La terminologie. Théorie, méthode et applications*, Ottawa, Les Presses de l'Université.

Cabré M.T. (2000), Terminologie et linguistique : la théorie des portes, *Terminologies nouvelles*, 21, 10-15.

Condamines A., Rebeyrolle J. (1997), Point de vue en langue spécialisée, *Meta,* XLII-1, 174-184.

Drouin P. (2004), Spécificités lexicales et acquisition de la terminologie, Actes de *JADT 2004*, 345-352.

Denhière G., Lemaire B. (2003), Modélisation des effets contextuels par l'analyse de la sémantique latente. Actes de *EPIQUE 2003,* http://www.upmf-grenoble.fr/sciedu/blemaire/epique03.pdf

Dunning T. (1993), Accurate methods for the statistics of surprise and coincidence, *Computational Linguistics*, 19-1, 61-74.

Ferret O. (2004), Découvrir des sens de mots à partir d'un réseau de cooccurrences lexicales, Actes de *TALN 2004*, http://www.lpl.univ-aix.fr/jep-taln04/proceed/actes/taln2004-Fez/Ferret.pdf

Gaudin F. (1993), *Pour une socioterminologie. Des problèmes sémantiques aux pratiques institutionnelles*, Rouen, Publications de l'Université de Rouen.

Labbé C., Labbé D. (2001), Que mesure la spécificité du vocabulaire?, *Lexicometrica*, 3, http://www.cavi.univ-paris3.fr/lexicometrica/article/numero3/specificite2001.PDF

Lafon P. (1984), *Dépouillements et statistiques en lexicométrie*, Genève-Paris, Slatkine-Champion.

Lerat P. (1995), *Les langues spécialisées*, Paris, PUF.

Martinez W. (2000), Mise en évidence de rapports synonymiques par la méthode des cooccurrences, Actes de *JADT 2000*, 78-84.

Schütze H. (1998), Automatic Word Sense Discrimination, *Computational Linguistics*, 24-1, 97-123.

Speelman D. (1997), *Abundantia verborum : a computer tool for carrying out corpus-based linguistic case studies*, PhD Thesis, K.U.Leuven.

Temmerman R. (1997), Questioning the univocity ideal. The difference between socio-cognitive Terminology and traditional Terminology, *Hermes*, 18, 51-90.

Véronis J. (2003), Cartographie lexicale pour la recherche d'informations, Actes de *TALN 2003*, 265-274.

Système AlALeR

Alignement au niveau phrastique des textes parallèles français-japonais

Yayoi NAKAMURA-DELLOYE
Université Paris 7 (École Doctorale de Sciences du Langage) - Lattice
30 Rue du Château des Rentiers 75013 Paris, http://www.linguist.jussieu.fr
1 rue Maurice Arnoux 92120 Montrouge, http://www.lattice.cnrs.fr
yayoi@free.fr
Date de soutenance prévue : 2006

Mots-clefs – Keywords

Alignement, corpus parallèles, analyse morphologique japonaise partielle, mémoire de traduction
Alignment, parallel corpora, partial japanese morphological analysis, translation memory

Résumé - Abstract

Le présent article décrit le Système AlALeR (Système d'Alignement Autonome, Léger et Robuste). Capable d'aligner au niveau phrastique un texte en français et un texte en japonais, le Système AlALeR ne recourt cependant à aucun moyen extérieur tel qu'un analyseur morphologique ou des dictionnaires, au contraire des méthodes existantes. Il est caractérisé par son analyse morphologique partielle mettant à profit des particularités du système d'écriture japonais et par la transcription des mots emprunts, à l'aide d'un transducteur.

The present paper describes the AlALeR System, an Autonomous, Robust and Light Alignment System. Capable of aligning at the sentence level a French text and a Japanese one, the AlALeR System doesn't use any external tool, such as morphological parsers or dictionaries, contrary to existing methods. This system is characterized by a partial morphological analysis taking advantage of some peculiarities of japanese writing system, and by the transcription of loan words with a transducer.

1 Introduction

La plupart des méthodes d'alignement au niveau phrastique sont caractérisées par leur simplicité de réalisation et de calcul, obtenue grâce à l'utilisation exclusive d'informations « internes », telles que la distribution de chaînes de caractères ou la longueur de chaîne.

Mais, les systèmes d'alignement du japonais intègrent tous aussi bien un analyseur morphologique que des dictionnaires bilingues pour traiter cette langue très différente des langues telles que l'anglais, l'allemand ou le français.

Cependant, l'alignement étant une opération élémentaire constituant souvent une étape préparatoire pour un autre traitement, un système léger est favorable pour le traitement du japonais également. Ainsi, nous avons conçu le système `AlALeR` adapté à l'alignement des textes japonais, qui ne recourt à aucun moyen extérieur, ni dictionnaire ni analyseur morphologique, en mettant pleinement à profit certaines particularités du système d'écriture du japonais.

Nous présentons dans cet article ce nouveau système d'alignement du japonais : après un bref parcours des techniques antérieures, nous décrirons d'abord le principe de fonctionnement de ce système avec le détail des procédures de certaines opérations, et présenterons ensuite les résultats obtenus.

2 Techniques antérieures et Système `AlALeR`

Les recherches sur la technique d'alignement ont débuté dans le cadre de travaux sur la traduction automatique. Si bien que les précurseurs ont cherché avant tout la simplicité de réalisation et de calcul, donnant ainsi naissance à des méthodes caractérisées par l'utilisation exclusive d'informations internes telles que la distribution lexicale (KAY & RÖSCHEISEN, 1993) ou la longueur des phrases (BROWN *et al.*, 1991), (GALE & CHURCH, 1993).

Les chercheurs occidentaux ont choisi pour améliorer la technique, la poursuite de la voie initiée par ces précurseurs en introduisant de nouvelles notions telles que les cognats ((SIMARD *et al.*, 1992), (LANGLAIS, 1997) et (KRAIF, 2001)), qui ne font pas appel aux informations extérieures.

Néanmoins, du fait que le système d'écriture du japonais ne dispose pas de séparateur graphique indiquant les frontières entre les mots, les chercheurs japonais ont intégré très tôt des analyseurs morphologiques dans leurs systèmes d'alignement (MURAO, 1991). De plus, le japonais est fortement différent des langues principalement traitées dans le TAL – telles que l'anglais, le français ou l'allemand – aussi bien sur le plan syntaxique que sur le plan lexical, ce qui n'a pas permis une simple application des méthodes utilisées pour ces langues au traitement des textes japonais. Aussi, les Japonais ont-ils également dû recourir à des dictionnaires bilingues et rechercher la performance plutôt que la simplicité (HARUNO & YAMAZAKI, 1996).

Notre système a résolu le problème de segmentation par une analyse morphologique partielle basée sur une méthode traditionnelle[1] qui profite d'une particularité du système d'écriture du japonais, possédant plusieurs types de caractères différents. Par ailleurs, la transcription des mots emprunts, à l'aide d'un transducteur, a permis un meilleur alignement des mots sans recourir à un dictionnaire bilingue.

[1] Voir aussi (NAKAMURA-DELLOYE, 2003).

3 Principe de fonctionnement

3.1 Schéma général du Système

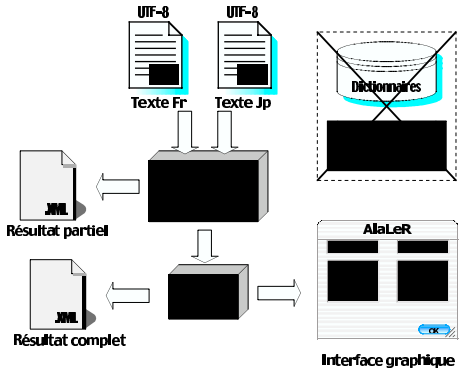


FIG. 1 – Schéma général du Système AlALer

Le système reçoit comme données une paire de textes parallèles rédigés en français et en anglais, ou plus particulièrement d'un texte en français (ou en anglais) et d'un en japonais. Afin de s'affranchir des problèmes d'encodage, fréquents lorsqu'il s'agit de traitements multilingues, AlALeR présuppose comme entrées des textes bruts au format texte, encodés en UTF-8.

Le système peut fournir comme résultat soit un alignement partiel très fiable des textes entrés, soit un alignement complet avec l'option « complet ». Lorsque cette option est choisie, le module de « post-alignement » réalise un appariement des phrases qui n'ont pas été alignées pendant le processus principal. L'appariement de ce module est réalisé selon la probabilité d'alignement de paires de phrases, calculée à partir de la corrélation de leur longueur.

Les résultats sont fournis, soit sous forme de fichier XML, soit par transfert vers l'interface graphique. Celle-ci permet non seulement de visualiser le résultat sous un format plus agréable à lire, mais aussi de faciliter la vérification et éventuellement la modification manuelle des résultats fournis par le système[2].

3.2 Procédure générale

La procédure générale est constituée de deux grandes étapes :

1. **Étape de construction de l'index du lexique**, au cours de laquelle les mots graphiques sont triés pour constituer quatre listes selon leur nature : transfuges, cognats, *katakana* et mots lexicaux ;
2. **Procédure d'alignement** :
   - **étape de préalignement**, au cours de laquelle un premier ancrage est réalisé pour limiter le nombre de possibilités d'alignement, à l'aide notamment des transfuges et des cognats ;
   - **procédure principale**, au cours de laquelle les phrases sont alignées par un calcul de similarité de la distribution des mots qu'elles contiennent. Cette procédure est itérative.

[2]Le système est implémenté en langage C++ et l'interface graphique avec les API Apple Carbon.

Nous allons maintenant présenter chaque étape. Étant donné que le système fonctionne un peu différemment selon les langues traitées, nous ne nous préoccupons ici que du cas d'un alignement de textes français et japonais, afin de mieux montrer la particularité de notre système.

3.3 Construction de l'index du lexique

Cette étape est composée elle-même de quatre étapes :
– Construction de la liste des phrases (`LPH`).
– Construction de la liste des mots graphiques (`LMOT`).
– Création de quatre listes à la suite du tri des mots graphiques : liste des transfuges (`LTRNS`), liste des cognats (`LCOG`), liste des mots en *katakana* (`LKTKN`) et liste des mots lexicaux (`LEX`).
– Création de l'index des mots lexicaux après leur lemmatisation (`ILX`).

3.3.1 Construction de la liste des phrases

Comme il a déjà été mentionné dans (SIMARD & PLAMONDON, 1998), la reconnaissance des phrases représente à elle-seule, malgré l'impression de trivialité que l'on a généralement, une question à part entière. La segmentation en phrases de textes français ou anglais n'est pas évidente à cause du caractère polysémique du séparateur graphique principal de phrase, le point final. Il est donc nécessaire de définir des règles assez détaillées permettant de segmenter correctement les séquences contenant des abréviations ou des sigles (« U.S.A »), des séquences symboliques (« abc@cdf.fr ») ou encore des nombres décimaux (1.5 en anglais).

Le point final japonais est beaucoup moins polysémique, facilitant ainsi la tâche de découpage.

3.3.2 Extraction des mots graphiques

Lors de la deuxième étape, consacrée à la construction de la liste des mots graphiques, la liste pour le texte français est construite par extraction des séquences entourées de séparateurs graphiques des mots – préalablement définis.

Pour le texte japonais – dans lequel il n'existe pas de séparateur graphique indiquant les frontières entre les mots –, une étape lourde de segmentation, réalisée généralement par analyse morphologique, est nécessaire. Cependant, il est possible de reconnaître la plupart des mots lexicaux sans analyse morphologique complète, en profitant d'une particularité du système d'écriture du japonais qui utilise trois types de caractères différents selon la nature des mots : *hiragana*, *katakana* et *kanji*.

– ***hiragana*** : premier syllabaire japonais souvent utilisé pour représenter la partie variable des mots variants et les mots grammaticaux ;
– ***kanji*** : idéogrammes utilisés pour représenter les mots pleins et les radicaux ayant un sens ;
– ***katakana*** : second syllabaire japonais employé pour la transcription des mots emprunts des langues étrangères (sauf le chinois).

Quoiqu'il soit impossible de segmenter totalement de manière correcte une phrase en mots uniquement avec cette méthode, il est possible de reconnaître la plupart des mots lexicaux en extrayant les séquences de *kanji* ou de *katakana*.

La liste ainsi obtenue ne contient néanmoins pas de mots grammaticaux. Nous supprimons donc les mots grammaticaux de la liste LMOT du texte français à l'aide d'une liste de mots grammaticaux préalablement définie.

3.3.3 Tri des mots

Le tri est ensuite réalisé aussi bien pour la liste LMOT du texte français que pour celle obtenue à partir du texte japonais afin de construire quatre nouvelles listes : la liste des transfuges (LTRNS), la liste des cognats (LCOG), la liste des mots en *katakana* (LKTKN) et la liste des mots lexicaux (LEX).

Pour les mots des trois premières listes, leur équivalence traductionnelle est calculable simplement par leur forme. Qui plus est, le résultat de ce calcul est beaucoup plus sûr que le résultat obtenu par la similarité des distributions.

Les « **cognats** », mots apparentés, sont des chaînes de caractères identiques ou proches graphiquement se trouvant dans les lexiques de langues ayant une relation historique plus ou moins étroite, tels que les paires anglais-français *generation/génération* et *error/erreur*. La notion de cognats améliore de manière simple et économique les méthodes statistiques qui n'utilisent aucune information lexicale, encore que son efficacité soit limitée aux langues appartenant à une même famille. Cependant, le japonais intégrant également dans son système d'écriture l'alphabet latin (*rôma-ji*), la possibilité d'obtention d'un résultat a été signalée très tôt dans (CHURCH *et al.*, 1993).

Le système AlALeR ne considère comme cognats que les chaînes alphabétiques totalement identiques apparaissant dans les deux textes entrés. Le système constitue d'abord la liste LCOG du texte japonais en extrayant les mots écrits en alphabet latin. Ensuite, en se référant à la liste japonaise, il construit une liste française en recherchant les séquences identiques aux éléments de la liste japonaise.

Les « **transfuges** » sont des chaînes invariantes à la traduction telles que les chiffres ou les symboles, inclus au début dans les cognats par les définitions du domaine de l'alignement, et regroupés plus tard par (LANGÉ & GAUSSIER, 1995) pour constituer une nouvelle catégorie. Les listes de transfuges LTRANS sont constituées séparément dans les deux langues par simple extraction des séquences de symboles ou de chiffres.

La troisième liste contient les mots du texte japonais écrits en *katakana*. La figure 2 représente la procédure d'appariement d'un mot en *katakana*. Ces transcriptions des mots emprunts sont retranscrites par le système à l'aide d'un transducteur – que nous avons développé spécifiquement – en une ou éventuellement en plusieurs formes en alphabet latin. Au cours du tri des mots français, si la similarité entre le mot français considéré et une séquence de retranscription d'un mot en *katakana* atteint un seuil prédéfini, le mot français est stocké dans la liste LKTKN. Les mots japonais en *katakana* qui n'ont pas trouvé d'équivalent une fois le tri des mots français terminé, sont stockés dans la liste des mots lexicaux pour leur laisser à nouveau une chance d'être finalement alignés par la similarité de distribution.

Le calcul de similarité entre une séquence retranscrite et un mot français est proche de la méthode de la sous-chaîne maximale parallèle utilisée dans (KRAIF, 2001) pour la reconnaissance des cognats. Notre formule, adaptée aux besoins particuliers de la retranscription des *katakana*, diffère de celle de ce dernier par le fait qu'elle tient compte non seulement de la sous-chaîne

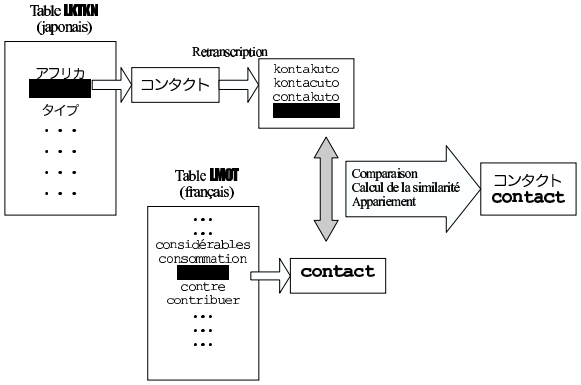


FIG. 2 – Appariement des mots en *katakana*

maximale parallèle mais aussi des consonnes communes. Le nombre de consonnes communes est pris en compte pour favoriser les deux chaînes ayant le plus de caractères consonantiques communs plutôt que celles dont les caractères vocaliques coïncident le plus.

Les paires constituées de deux mots appartenant à l'une de ces trois listes constituent ensuite des listes de paires de mots alignés, appelées table des « Transfuges alignés » (TRAL), table des « Cognats alignés » (COGAL) et table des « Katakana alignés » (KTKNAL).

3.3.4 Lemmatisation

Lemmatisation des mots français Nous avons eu recours à la méthode utilisée à l'étape morphologique dans (KAY & RÖSCHEISEN, 1993). Elle consiste à trouver les sous-chaînes préfixales ou suffixales communes à plusieurs formes effectives des mots graphiques et à trouver ensuite leurs radicaux, porteurs de sens. Ce traitement est implémenté efficacement grâce à l'utilisation d'une structure de données appelée *trie*.

Lemmatisation des mots japonais Après la segmentation par type de caractère, il subsiste encore un cas de segmentation à réaliser : les séquences de mots composés constitués de plusieurs substantifs juxtaposés les uns derrières les autres. Dans ce type de séquence, généralement entièrement en *kanji*, la frontière entre les deux mots composants n'est pas marquée par un changement de type de caractère.

Il s'agit donc également de la recherche des sous-chaînes communes à plusieurs formes effectives. Si bien que nous avons adopté pour le japonais une lemmatisation reposant sur la structure de données *trie*.

La différence dans le cas du japonais est que les parties restantes ne sont pas des suffixes mais un ou même plusieurs autres mots portant eux-mêmes un sens propre. On obtient donc à partir d'un mot graphique ab, non pas un lemme a, mais deux lemmes a et b. La figure 3 représente un exemple d'arbres vérifiant des chaînes préfixales et suffixales, créés à partir de sept entrées. Elle montre comment segmenter les mots japonais à l'aide de ces arbres. En réalisant de manière itérative cette opération, on peut obtenir plus de deux lemmes si la séquence en contient.

Nous obtenons ainsi l'ensemble des données nécessaires à la mise en correspondance des mots permettant d'associer ensuite les phrases à aligner.

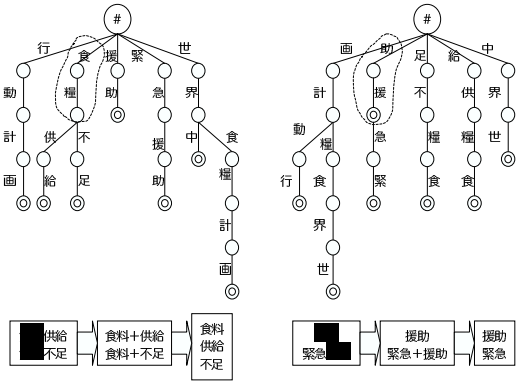


FIG. 3 – Arbres vérifiant des chaînes préfixales (à gauche) et suffixales (à droite)

3.4 Procédure d'alignement

Notre système utilise une technique basée sur les informations des distributions lexicales, présentée par (KAY & RÖSCHEISEN, 1993). Cette méthode, reposant sur l'hypothèse que les phrases correspondantes comprennent des éléments correspondants, est constituée d'un appariement grossier des mots, qui permet ensuite l'alignement des phrases contenant les mots appariés.

Les deux textes sont représentés par une matrice dont les lignes correspondent à chacune des phrases du texte français et les colonnes à celles du texte japonais. L'étape d'alignement est précédée par le préalignement et composée de trois opérations correspondant chacune à la construction d'une structure de données particulière : la table « Candidats des paires de phrases à aligner » (CPR), la table « Mots alignés » (MAL) et la table « Résultat d'alignement » (RAL).

3.4.1 Préalignement

Le préalignement consiste à trouver des ancrages sûrs permettant de réduire la zone de recherche. Le préalignement de notre système, inspiré de la méthode proposée dans (KRAIF, 2001), est réalisé à l'aide des tables TRAL, COGAL et KTKNAL. Il se fait via deux parcours de ces tables.

3.4.2 Table « Candidats des paires de phrases à aligner »

La table CPR est une matrice indiquant les paires de phrases susceptibles d'être alignées. Basée sur l'hypothèse de diagonalité de l'alignement, la zone constituée des cases correspondant aux paires candidates forme une ellipse avec pour axe principal la diagonale de la matrice.

3.4.3 Table « Mots alignés »

La table MAL contient l'ensemble des paires de mots supposés être traductions l'un de l'autre.

L'appariement des mots est réalisé selon la similarité de la distribution de chaque mot. Tous les mots appartenant à un même candidat paire de phrases sont comparés, et leur est attribuée

une similarité basée sur leur distribution. De nombreuses formules ont été proposées jusqu'aujourd'hui pour le calcul de cette similarité de distribution lexicale. Notre méthode est inspirée de l'amélioration par (KITAMURA & MATSUMOTO, 1997) du coefficient de Dice : en plus de la différence de fréquences, elle tient également compte de la fréquence elle-même, donnée contrôlée séparément dans les algorithmes antérieurs. La nouveauté apportée par notre formule est la prise en compte du nombre de phrases où les mots considérés apparaissent. Cette modification améliore les résultats lorsque deux paires ont une similarité identique, situation entraînant des conflits avec les méthodes précédentes.

3.4.4 Table « Résultat d'alignement » et module de post-alignement

La table `RAL` contient l'ensemble des paires de phrases supposées être traductions l'une de l'autre.

L'appariement des phrases utilise la table `MAL` obtenue précédemment, en plus des tables `TRAL`, `COGAL` et `KTKNAL`, pour calculer combien de couples de mots de ces tables contient chaque paire de phrases appartenant à la `CPR`. Si une paire de phrases comporte plus de paires de mots alignés que le seuil défini – en fonction de la taille du texte –, ces phrases sont considérées comme correspondantes traductionnelles. Ces nouvelles paires servant de nouvelles ancres, on crée une nouvelle `CPR` pour réaliser de manière itérative ces opérations d'alignement.

Ce premier résultat partiel peut être complété par une procédure de post-alignement. Le module de post-alignement extrait les sous-matrices constituées des phrases non alignées par le noyau `AlALeR` et calcule la probabilité d'alignement de toutes les paires possibles de phrases. Il réalise ensuite l'appariement de ces phrases avec une méthode de programmation dynamique, de manière à mettre en relation toutes les phrases avec au moins une phrase de l'autre texte.

4 Résultat

Nous avons testé les performances de notre système (sur PowerMac G5) avec cinq textes parallèles français-japonais et deux anglais-japonais : deux articles de *Label France*[3] (désignés ci-après « Bio » et « FIV »), un texte du sommet G8 (« G8 »), *How to Unicode* (« Unicode »[4]), un texte de l'EU (« EUJP ») et deux œuvres littéraires (un extrait de *Zadig* de Voltaire (« Zadig ») et *Balthasar* (« Balth ») d'Anatole France).

Pour chaque texte, nous avons analysé deux résultats : le résultat partiel sans opération de post-alignement (noyau `AlALeR`) et le résultat complet obtenu grâce au post-alignement.

	Bio		FIV		G8		Unicode		Zadig		EUJP		Balth	
Lang	Fr	Jp	Fr	Jp	Fr	Jp	Fr	Jp	Fr	Jp	Ang	Jp	Ang	Jp
Phr	69	75	54	52	53	47	274	268	1206	1376	252	238	321	423
M/C	1418	3615	1176	2597	1398	3077	4224	14155	17912	43426	3881	14308	4835	11491

TAB. 1 – Caractéristiques des textes

[3]Magazine du Ministère des affaires étrangères

[4]Sites internet : (VF) `http://www.freenix.fr/unix/linux/HOWTO/Unicode-HOWTO.html`; (VJ) `http://www.linux.or.jp/JF/JFdocs/Unicode-HOWTO.html`

	Modèles de traduction													partiel		complet
	0-1	1-0	1-1	1-2	1-3	1-4+	2-1	2-2	2-3	2-4+	3-1	3-2	4+ -1	rap	pré	pré
Bio	0	0	55	7	1	0	3	0	0	0	0	0	0	0.81	1	0,98
FIV	0	0	43	3	0	0	2	0	0	0	0	0	1	0,66	1	0,92
G8	0	0	38	1	0	0	7	0	0	0	0	0	0	0,95	1	0,98
Unicode	1	0	195	22	1	0	19	2	0	0	1	1	1	0,90	0,99	0,96
Zadig	7	5	773	188	29	3	69	9	6	1	10	0	2	0,34	0,97	0,78
EUJP	0	4	208	5	1	0	17	0	0	0	0	0	0	0,87	0,98	0,92
Balth	1	2	185	68	16	4	9	13	1	0	0	0	0	0,49	0,97	0,86

TAB. 2 – Modèles de traduction et résultats : rappel et précision

Le tableau 1 montre les nombres de phrases et de mots (textes français) ou de caractères (textes japonais) de chaque texte.

Le tableau 2 présente la répartition par modèle de traduction de chaque paire de textes. La colonne 1-1 montre le nombre de paires en relation traductionnelle, constituées d'une phrase du premier texte (français ou anglais) et d'une du second texte (japonais), la colonne 1-2 le nombre de paires constituées d'une phrase du texte 1 et de deux phrases du texte 2, et ainsi de suite.

Le tableau 2 présente également le résultat d'alignement avec les taux de précision et de rappel, dans le cas du résultat partiel[5].

On peut déduire de l'analyse de ce tableau que le système supporte bien les modèles complexes – i.e. les modèles constitués de plus d'une phrase, tels que 1-3 –, qui perturbaient les systèmes d'alignement basés sur des méthodes probabilistes uniquement, au point de fausser tous les alignements effectués après l'analyse d'un modèle complexe. Cette robustesse est due au résultat partiel extrêmement précis, qui sert d'ancrage fiable pour le post-alignement plus robuste. Le point faible de cette méthode par rapport aux méthodes probabilistes est, comme déjà critiqué par plusieurs auteurs, son utilisation importante de mémoire.

Par ailleurs, le taux de rappel très bas de certains textes est dû non seulement à la présence faible voire l'absence de cognats ou de transfuges, mais aussi à la présence importante des mots de fréquence faible, notamment 1. C'est justement un autre point faible des méthodes basées sur la similarité de distribution. Afin de compenser cet inconvénient, notre système adopte un appariement final basé sur la corrélation des longueurs.

5 Conclusion et perspectives des travaux futurs

Les résultats d'alignement fournis par notre système AlALeR ont montré la possibilité de conception d'un aligneur traitant les textes japonais qui ne recourt à aucun dictionnaire ni analyseur morphologique. Ce résultat est d'abord dû à la stratégie d'appariement des mots japonais en *katakana*. Ceux-ci étant très nombreux dans les textes traduits, la retranscription de mots japonais en *katakana* pour trouver leur mot d'origine a été d'autant plus efficace qu'ils sont souvent absents des dictionnaires. En effet, ce sont très souvent des néologismes ou des noms propres. Cette stratégie s'est montrée extrêmement robuste, ce que nous n'aurions pas pu constater si

[5]Le taux de rappel représente la proportion de phrases appariées avec au moins une phrase de l'autre texte et le taux de précision, la proportion parmi ces phrases appariées de celles l'étant correctement, avec au moins une phrase de l'autre texte.

nous avions dépendu d'un dictionnaire.

Nous avons également testé le système avec quelques traductions de brevets techniques et nous avons obtenu de très bons résultats grâce à la présence très importante de transfuges. Néanmoins, les phrases de ce type de document sont si longues que l'alignement au niveau phrastique ressemble plutôt à un alignement de paragraphes. Comme Simard le fait remarquer dans (SIMARD, 2003), l'alignement à un niveau sous-phrastique est plus bénéfique que celui réalisé au niveau phrastique, notamment en vue de la constitution de mémoires de traduction. Nous allons désormais nous attacher à réaliser un alignement de propositions, qui permettra très certainement de fournir une base de données plus intéressante, aussi bien pour la conception des mémoires de traduction que pour l'étude des linguistiques contrastives.

Références

BROWN P. F., LAI J. C. & MERCER R. L. (1991). Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, p. 169 – 176.

CHURCH K., DAGAN I., GALE W., FUNG P. & J. HELFMAN B. S. (1993). Aligning parallel texts : do methodes developed for english-french generalize to asian languages ? In *Proceedings of the Pascific Asia Conference on Formal and Computational Linguistics*.

GALE W. A. & CHURCH K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, **19**(3), 75–102.

HARUNO M. & YAMAZAKI T. (1996). Bilingual text alignment using statistical and dictionary information. *IPSJ SIG Notes*, **NL 112**(4), 23–30. en japonais.

KAY M. & RÖSCHEISEN M. (1993). Text-translation alignment. *Computational Linguistics*, **19**(1), 121–142.

KITAMURA M. & MATSUMOTO Y. (1997). Automatic extraction of translation patterns in parallel corpora. *IPSJ Journal*, **38**(4), 727– 736. en japonais.

KRAIF O. (2001). Exploitation des cognats dans les systèmes d'alignement bi-textuel : architecture et évaluation. *TAL*, **42**(3).

LANGÉ J.-M. & GAUSSIER E. (1995). Alignement de corpus multilingues au niveau des phrases. *TAL*, **36**(1–2).

LANGLAIS P. (1997). Alignement de corpus bilingues : intérêt, algorithmes et évaluation. In *Bulletin de Linguistique Appliquée et Générale, numéro Hors Série*, p. 245–254. Université de Franche-Comté.

MURAO H. (1991). Studies on bilingual text alignment. Bachelor thesis, Kyoto University. en japonais.

NAKAMURA-DELLOYE Y. (2003). Analyse syntaxique du japonais. Mémoire de D.E.A., Institut National des Langues et Civilisations Orientales.

SIMARD M. (2003). *Mémoire de traduction sous-phrastique*. Thèse de doctorat en informatique, Université de Montréal.

SIMARD M., FOSTER G. & ISABALLE P. (1992). Using cognates to align sentences in bilingual corpora. In *Proceedings of the Fourth International Conference on Thoretical and Methodological Issues in Machine Translation (TMI)*, p. 67 –81.

SIMARD M. & PLAMONDON P. (1998). Bilingual sentence alignment : Balancing robustness and accuracy. *Machine Translation*, **13**(1), 59–80.

Acquisition semi-automatique de relations lexicales bilingues (français-anglais) à partir du Web

Stéphanie Léon & Chrystel Millon

Equipe DELIC – Université de Provence
29, Av. Robert Schuman – 13621 Aix-en-Provence Cedex 1
fanny.leon@orange.fr, Chrystel.Millon@up.univ-mrs.fr

Mots-clefs – Keywords

Traduction, corpus, relations lexicales bilingues, acquisition semi-automatique, World Wide Web.

Translation, corpus, bilingual lexical relations, semi-automatic acquisition, World Wide Web.

Résumé – Abstract

Cet article présente une méthode d'acquisition semi-automatique de relations lexicales bilingues (français-anglais) faisant appel à un processus de validation sur le Web. Notre approche consiste d'abord à extraire automatiquement des relations lexicales françaises. Nous générons ensuite leurs traductions potentielles grâce à un dictionnaire électronique. Ces traductions sont enfin automatiquement filtrées à partir de requêtes lancées sur le moteur de recherche Google. Notre évaluation sur 10 mots français très polysémiques montre que le Web permet de constituer ou compléter des bases de données lexicales multilingues, encore trop rares, mais dont l'utilité est pourtant primordiale pour de nombreuses applications, dont la traduction automatique.

This paper presents a method of semi-automatic acquisition of bilingual (French-English) lexical relations using a validation process via the Web. Our approach consists firstly of automatically extracting French lexical relations. We then generate their potential translations by means of an electronic dictionary. These translations are finally automatically filtered using queries on the Google search engine. Our evaluation on 10 very polysemous French words shows that the Web is a useful resource for building or improving multilingual lexical databases, which are urgently needed in a wide range of applications, such as machine translation.

1 Introduction

Bien qu'elle ait été la première application non-numérique de l'informatique, la traduction automatique a connu des débuts décevants, qui ont jeté un discrédit sur cette technologie pendant plusieurs décennies. Toutefois, des progrès ont été accomplis au cours des dernières années, en raison de l'avènement du Web dans un contexte fortement multilingue, de

l'accroissement très important de la couverture des dictionnaires présents dans les systèmes de traduction, et de la prise en compte d'un nombre croissant d'expressions composées. Par exemple, le système Systran[1] traduit désormais correctement du français vers l'anglais des expressions telles que :

vol à main armée → *armed robbery*
vol à la roulotte → *stealing from parked vehicles*

En revanche, dès que l'on sort de ces listes d'expressions figées, on retombe rapidement dans des erreurs de traduction qui gênent considérablement la compréhension, et lui donnent même parfois un caractère surréaliste. Ainsi, Systran utilise la traduction la plus fréquente du mot *vol,* c'est-à-dire *flight* (VOL AERIEN), dans toutes les autres situations. Si *réserver un vol* est correctement traduit (*to reserve a flight*), *commettre un vol* est traduit par *to make a flight,* ce qui est totalement incompréhensible dans ce contexte pour un anglophone.

Pourtant, la relation lexicale[2] *commettre-vol* est un indice désambiguïsateur très fort, qui, si elle était correctement traduite dans une base de données (*to commit-theft*), pourrait servir à générer des traductions correctes. La combinatoire est toutefois beaucoup plus ouverte qu'avec les expressions composées mentionnées plus haut et la constitution manuelle d'une base de données de combinaisons lexicales à grande échelle est une tâche à peu près impossible. Les dictionnaires bilingues se contentent d'ailleurs de rares indications ponctuelles, se fiant au jugement du lecteur et à sa connaissance du monde, que l'on ne peut guère espérer d'une machine.

Afin de constituer de telles bases de données bilingues, de nombreux travaux se sont appuyés sur des corpus parallèles ou alignés[3] (Véronis, 2000). Toutefois, ces méthodes présentent diverses limites : de telles ressources sont peu nombreuses et concernent des domaines restreints. Le paysage est un peu plus souriant avec des corpus comparables[4] (Morin, Dufour-Kowalski et Daille, 2004), mais les contraintes restent fortes.

Le Web, qui génère des besoins considérables de traduction, offre en même temps un réservoir gigantesque de données qui peuvent être exploitées par des moyens automatiques, en particulier grâce à des moteurs de recherche tels que Google ou Altavista. En se basant sur les travaux de (Kilgarriff & Grefenstette, 2003), on peut estimer à environ 100 milliards le nombre de mots indexés par Google pour la seule langue anglaise. Même si les données du Web sont moins contrôlées, et donc plus « bruitées », elles permettent d'envisager un changement radical d'échelle pour les méthodes basées sur les données, à condition de développer des méthodes et des techniques adaptées. Depuis quelques années, divers domaines du TAL utilisent le Web en tant que ressource de données linguistique. (Cao & Li, 2002) proposent une méthode automatique de génération et de sélection de traductions pour les syntagmes nominaux de l'anglais vers le chinois à partir du Web. Les résultats montrent que le Web peut être utile tant pour l'acquisition de données que pour une aide à la validation.

---

1 http://www.systransoft.com/

2 Concernant la combinatoire lexicale, la littérature présente une terminologie disparate et souvent floue. Certains parlent de « préférences lexicales » (Wilks, 1975), de « restrictions de sélection » (Katz & Fodor, 1964) ou encore de « collocations » (Benson, 1990 ; Smadja, 1993 ; Cruse, 1986). Afin de désigner ce phénomène, nous employons ici le terme « relation lexicale », plus neutre, défini comme une cooccurrence lexicale entre deux lexèmes liés syntaxiquement.

3 Ensemble de textes alignés avec leur traduction au niveau du paragraphe, de la phrase, des expressions ou des mots.

4 Corpus de langues différentes traitant du même domaine mais non parallèles.

L'objectif de notre article est de proposer une méthodologie de validation semi-automatique de traductions anglaises via le Web, à partir de relations lexicales françaises du type *NOM ADJECTIF*, *NOM1 DE NOM2* et *VERBE NOM*(*objet*) extraites d'un corpus français de pages Web en langue générale. Pour revenir à nos exemples *commettre un vol* et *réserver un vol* (voir Figure 1), Google nous permet de valider les traductions correctes, grâce à leur nombre d'occurrences. Par exemple, la requête [*"commit a flight"* OR *"commit the flight"*] retourne seulement 13 résultats. La requête [*"commit a theft"* OR *"commit the theft"*] retourne quant à elle 5110 résultats. Parmi ces deux traductions candidates, les résultats sélectionnent de façon écrasante la relation lexicale satisfaisante (*to commit-theft*).

	Effectifs absolus		***Effectifs par million***	
	flight	**theft**	**flight**	**theft**
commit	13	5510	0	306
reserve	33 500	3	592	0

Figure 1 : Exemples de résultats sur Google (janvier 2005)

2 Méthodologie et traitement des données

Notre étude comporte trois étapes. Nous procédons d'abord à une acquisition de relations lexicales françaises via un corpus de pages Web. Nous générons ensuite de façon automatique leurs traductions potentielles à partir d'un dictionnaire électronique. Nous interrogeons enfin le moteur de recherche Google afin de valider ces dernières de façon semi-automatique.

2.1 Extraction automatique des relations lexicales françaises

Notre méthode a été testée sur la combinatoire lexicale de 10 noms français très polysémiques (*barrage*, *détention*, *formation*, *lancement*, *organe*, *passage*, *restauration*, *solution*, *station* et *vol*). Ces mots ont été jugés comme les plus polysémiques parmi 200 noms de fréquence équivalente, lors du projet Senseval (Véronis, 1998) et constituent donc un banc de test difficile, qui a été utilisé par la suite dans divers travaux. Nous exploitons la version lemmatisée et étiquetée morpho-syntaxiquement du corpus de pages Web francophones élaboré par Jean Véronis (Véronis, 2003) autour de ces 10 noms.

L'acquisition automatique des relations lexicales françaises de type *NOM ADJECTIF*, *NOM1 DE NOM2*[1] et *VERBE NOM*(*objet*) utilise une version améliorée du programme employé dans (Millon, 2004), dans lequel des filtres linguistiques d'extraction sont utilisés (représentation de patrons syntaxiques, filtres de candidats indésirables). Les relations lexicales sont ensuite soumises à un seuil limite de fréquence fixé à au moins 10 occurrences, afin d'obtenir des relations lexicales représentatives de besoins en lexicographie. En effet, un nombre important de relations moins fréquentes sont « accidentelles » (comme par exemple *barrage violet*), et non pertinentes pour notre étude. L'objectif est d'extraire une majorité de relations lexicales « propres » du côté français. A l'inverse, des relations lexicales caractéristiques sont perdues (ainsi *barrage hydraulique* a une fréquence de 4).

Une catégorisation sémantique des relations lexicales n'est pas réalisée, car, outre qu'elle serait extrêmement difficile à obtenir avec fiabilité, elle se perdrait lors de l'interrogation des traductions potentielles dans Google. Pour une relation lexicale française telle que *vol de nuit*, l'objectif est de donner un ensemble d'équivalences en anglais qui reflètent soit des variations

[1] Le nom source peut se trouver en position *NOM1* ou *NOM2*.

lexicales pour un même usage comme dans les traductions *night flight* et *night flying* (usage VOL AERIEN), soit des usages différents comme l'exemple de *night robbery* (usage DELIT). La Figure 2 donne la quantité de données obtenues après filtrage des relations lexicales françaises (seuil de fréquence).

	RLs françaises totales	**RLs Françaises ≥ 10**	
		Occurrences	**% restant**
N ADJ	1332	113	8,48%
N DE N	1940	173	8,92%
V N	1278	57	4,46%
TOTAL	**4550**	**286**	**6,29%**

Figure 2 : Résultats de l'extraction des relations lexicales sources

2.2 Génération automatique des traductions potentielles

Les dictionnaires courants contiennent un nombre très restreint de ces relations lexicales, généralement les plus figées. Ainsi, le *Collins Pocket French-English Dictionary* disponible dans l'équipe sous forme électronique grâce à un accord avec l'éditeur Collins, ne propose de traduction que pour 6,6 % des relations lexicales françaises que nous conservons après filtrage, telles que *barrage routier* traduit par *roadblock* ou *station balnéaire* traduite par *seaside resort*.

Pour chacune des relations lexicales françaises non présentes dans le dictionnaire, nous générons automatiquement toutes les traductions possibles via le *Collins Pocket*. Reprenons pour exemple *réserver-vol*. Le *Collins Pocket* donne les traductions suivantes pour les unités lexicales sources *vol* et *réserver* (unité lexicale source vers unité lexicale cible) :

vol → *flight, theft, flying*
réserver → *to reserve, to book*

Notre programme génère toute la combinatoire :

réserver un vol → *to reserve a flight, to reserve a theft, to reserve a flying, to book a flight, to book a theft, to book a flying*

Afin d'avoir un ensemble de traductions le plus exhaustif possible, nous recensons également les « traductions inversées » des unités lexicales françaises, en recherchant ces dernières lorsqu'elles apparaissent en tant que traduction dans la version *English-French*, ce qui rajoute parfois des traductions, comme pour *vol* :

larceny, robbery, snatch → *vol*

Lorsque le dictionnaire propose une traduction (par exemple *vol libre* → *hand-gliding*), nous n'avons pas généré de traduction supplémentaire. Dans certains cas, comme pour *rampe de lancement*, traduite par *launch pad* et *launching pad* dans le dictionnaire, nous manquerons quelques traductions correctes comme *launching ramp* et *launch ramp*.

La Figure 3 donne la quantité de traductions potentielles générées et la proportion de celles-ci par relation lexicale française.

	Traductions générées	Moyenne par RL française
N ADJ	1215	11
N DE N	5155	30
V N	1012	18
TOTAL	7382	19

Figure 3 : Résultats de l'étape de génération des traductions candidates

2.3 Interrogation automatique du moteur de recherche Google

Le moteur de recherche Google a été interrogé automatiquement à l'aide de l'interface de programmation d'applications API (*Application Programming Interface*)[1] afin de récupérer le nombre d'occurrences[2] de chaque relation lexicale anglaise candidate. Ces fréquences seront utilisées lors de la validation[3].

Pour chaque traduction potentielle, nous générons un ensemble de requêtes (voir Figure 4), en considérant les mots de la requête comme une expression exacte, via l'utilisation des guillemets. La recherche est restreinte aux pages Web de langue anglaise.

Patron syntaxique source	Requête (en langue cible)
NOM ADJECTIF	"the ADJ NOM" OR "a ADJ NOM"
VERBE NOM(objet)	"VERBE the NOM" OR "VERBE a NOM"
NOM1 de NOM2	"NOM_1 of NOM_2"
	"NOM_2 NOM_1"

Figure 4 : Patrons des requêtes des relations lexicales anglaises

La combinaison booléenne pour les patrons syntaxiques de type *NOM ADJECTIF* et *VERBE NOM*(*objet*) ramène un ensemble de résultats qui prend en compte les variations dues aux changements d'article, comme dans l'exemple *"commit a theft" OR "commit the theft"*.

L'utilisation d'articles dans les requêtes du patron syntaxique *NOM ADJECTIF* permet également de réduire le problème de l'ambiguïté catégorielle. Par exemple, *complete* peut être un adjectif (*entier, complet, intégral, total*) ou un verbe (*parfaire, compléter*). La relation lexicale *complete restoration* est ambiguë. L'ajout de l'article permet d'éliminer les cas où *complete* est un verbe.

Le patron syntaxique *NOM1 DE NOM2* pouvant être traduit par différentes structures en anglais selon la relation sémantique considérée entre les deux objets, nous traitons séparément deux types de structures dans les requêtes (Chuquet & Paillard, 1987) : d'une part, le patron *N2 N1* marque une relation étroite entre les deux noms et d'autre part, la structure *N1 of N2* accorde la priorité à l'élément repéré (*N2*)[4].

[1] http://www.google.com/apis/

[2] Des différences ont été remarquées entre le nombre de résultats renvoyés par l'API et par l'interface web. Ce problème a été mentionné dans divers forums, mais aucune explication n'a pu être fournie.

[3] Précisons que si les fréquences de Google sont peu fiables dans le cadre de certaines configurations de requêtes dans « tout le Web » (http://aixtal.blogspot.com/2005/02/web-le-mystre-des-pages-manquantes-de.html), ce problème ne concerne pas l'utilisation que nous faisons de Google puisque nous limitons les requêtes à une langue donnée.

[4] Le cas du génitif (*N1's N2*) n'est pas pris en compte dans le cadre de cette étude.

2.4 Validation semi-automatique des traductions potentielles anglaises

2.4.1 Filtre automatique

Afin de réduire le bruit, un filtre simple a été appliqué aux traductions restantes. Nous ne conservons que celles dont la fréquence sur le Web est au moins égale à un millième des occurrences du mot cible. Nous utilisons à l'heure actuelle une méthode statistique simple qui est concluante pour une première approche. Mais nous pouvons envisager par la suite d'autres techniques plus élaborées. Prenons pour exemple *réserver-vol* et deux de ses traductions candidates *book a/the flight* et *book a/the theft* :

Seuil_theft : 2150000 / 1000 = 2150
Seuil_flight : 5760000 / 1000 = 5760

La relation lexicale *book a/the flight* (avec une fréquence de 244000, donc supérieure au seuil limite pour le nom cible *flight*) est retenue, tandis que *book a/the theft* (avec une fréquence de 61, donc inférieure au seuil limite pour le nom cible *theft*) est rejetée.

Ce filtre provoque évidemment parfois des cas de silence. Ainsi, à partir de la relation lexicale *barrage hydro-électrique*, la traduction *hydroelectric dam* est retenue (fréquence de 3920 et seuil à 1910), tandis que *hydroelectric barrage*, également valide, a une fréquence de 32 (pour un seuil à 139) et est rejetée. Notre approche favorise volontairement la précision, car il s'agit de compléter le plus automatiquement possible des ressources existantes. L'augmentation du bruit obligerait à un filtrage manuel des résultats beaucoup plus long et coûteux. Après le filtre automatique sur les fréquences, 7,5 % des relations anglaises potentielles sont conservées (Figure 5).

2.4.2 Validation manuelle

Une validation manuelle nous permet d'évaluer les relations lexicales restantes après filtre automatique, en faisant appel à divers dictionnaires de langue, ainsi qu'à nos connaissances. Les traductions candidates les plus « délicates » sont vérifiées en contexte sur le Web, par reformulation de la requête à travers Google, ainsi que soumises au jugement de plusieurs locuteurs. Nous détaillons les résultats dans la section suivante.

	Traductions générées	***Filtre automatique***		***Validation manuelle***	
		Filtre seuil fréquence		**Traductions valides**	
N ADJ	1215	136	11,2%	132	97,1%
N DE N	5155	351	6,8%	270	76,9%
V N	1012	63	6,2%	56	88,9%
TOTAL	**7382**	**550**	**7,5%**	**458**	**83,3%**

Figure 5 : Résultats de la validation des traductions

3 Premiers résultats

La précision globale des traductions extraites est de 83,3%. La méthode fonctionne particulièrement bien pour les patrons syntaxiques *NOM ADJECTIF* (97,1%) et *VERBE NOM* (*objet*) (88,9%) (Figure 6). Le patron *NOM1 DE NOM2* pose davantage de difficultés en traduction (76,9%).

MOT SOURCE	N ADJ		N de N		V N		TOTAL	
	Nb total	*% valide*	*Nb total*	*% valide*	*Nb total*	*% valide*	*Nb total*	*% valide*
Barrage	39	94,9	33	81,8	19	94,7	91	90,1
Détention	15	100,0	114	80,7	5	100,0	134	83,6
Formation	16	100,0	76	72,4	4	75,0	96	77,1
Lancement	7	100,0	17	88,2	2	100,0	26	92,3
Organe	27	96,3	21	76,2	4	25,0	52	82,7
Passage	10	90,0	21	57,1	13	92,3	44	75,0
restauration	2	100,0	23	78,3	Pas de RL	Pas de RL	25	80,0
Solution	14	100,0	5	100,0	8	100,0	27	100,0
Station	4	100,0	18	72,2	4	100,0	26	80,8
Vol	2	100,0	23	73,9	4	75,0	29	75,9
TOTAL	**136**	**97,1**	**351**	**76,9**	**63**	**88,9**	**550**	**83,3**

Figure 6 : Précision globale des relations lexicales anglaises évaluées

La Figure 7 présente le nombre moyen de traductions valides par relation lexicale française. En moyenne, on obtient deux traductions correctes pour chaque relation lexicale française.

	RLs françaises	**Traductions validées**	**Moyenne par RL**
N ADJ	63	132	2,1
N de N	124	270	2,2
V N	31	56	1,8
MOYENNE	**218**	**458**	**2,1**

Figure 7 : Nombre moyen de traductions valides par relation lexicale française

La Figure 8 présente un exemple de traductions obtenues pour les relations lexicales *barrage hydro-électrique*, *construction de barrage* et *construire-barrage*.

PATRON	**RL FRANCAISE**	**TRADUCTION**
N ADJ	barrage hydro-électrique	hydroelectric dam
N de N	construction de barrage	barrage building
		barrage construction
		barricade building
		barricade construction
		dam building
		dam construction
		weir building
		weir construction
V N	construire-barrage	to build-barrage
		to build-barricade
		to build-dam
		to build-roadblock
		to construct-dam
		to erect-barricade
		to erect-roadblock

Figure 8 : Traductions des relations lexicales de barrage

Les traductions obtenues permettent de multiplier par 10 la quantité de relations lexicales déjà présentes en entrée dans le dictionnaire pour le patron *NOM ADJECTIF*, par 45 celles pour le patron *NOM1 DE NOM2* et par 56 pour le patron *VERBE NOM*(*objet*). L'apport de la méthode est donc appréciable.

4 Analyse des problèmes et perspectives

4.1 Erreurs et difficultés

Les erreurs et les limites de notre stratégie sont catégorisées selon les types de problème et les perspectives d'amélioration.

4.1.1 Erreurs syntaxiques

Certaines erreurs sont dues à des problèmes d'ambiguïté de rattachement syntaxique concernant les traductions candidates testées sur l'API Google. Le moteur de recherche ne permettant pas un accès direct aux catégories morpho-syntaxiques, une analyse syntaxique des contextes des traductions candidates n'a pu être réalisée. Prenons l'exemple suivant :

The Library of Congress ***set the changeover*** *date.*

La relation lexicale recherchée est *set the changeover*. Or dans ce cas, *changeover* est régi par le nom *date* et non par la forme verbale *set*. Une autre limite concerne les problèmes d'ambiguïté catégorielle. Les formes en *-ing*, par exemple, sont propices à ce type d'erreurs. Une perspective d'amélioration des problèmes d'ordre syntaxique est donc manifestement l'application d'un analyseur syntaxique aux pages Web anglaises.

Enfin, il arrive qu'une relation lexicale ne soit pas traduite par une structure de même longueur. Par exemple, *barrage routier* se traduit par une unité lexicale simple : *roadblock*. La formation des mots composés en anglais, avec ou sans espace ou tiret, est évidemment un cas extrêmement difficile, mais on peut envisager de générer des requêtes du type N-N ou NN (sans espace).

4.1.2 Erreurs sémantiques

Un type d'erreurs d'ordre sémantique concerne l'acquisition de relations lexicales anglaises valides, mais non correspondantes à la relation lexicale source, comme dans l'exemple :

cours de formation --> group rate (59900 occurrences)

Ici, *group rate* signifie *tarif de groupe*.

De plus, la traduction d'une relation lexicale n'est pas toujours obtenue de façon compositionnelle (Melamed, 2001, cité par Morin *et al.*, 2004). Par exemple, en anglais, il n'existe pas une traduction littérale de *forcer un barrage* : la traduction va dépendre du contexte situationnel (*to drive through a roadblock*, *to run through a roadblock*, etc.).

Une autre limite est due à l'absence de certains usages dans le dictionnaire. C'est le cas par exemple de l'acception sportive du nom *barrage* (*playoff*). Ainsi, la relation lexicale *match de barrage* est traduite par *weir game* et *barrage game* au lieu de la traduction correcte *playoff game*.

4.1.3 Erreurs techniques

Google ne prend pas en compte des phénomènes tels que la ponctuation, les majuscules, ou la marque du génitif (*'s*), lors des recherches sur le Web. Certaines relations lexicales anglaises

erronées, comme *to reserve-theft* (fréquence de 3) n'ont pas une fréquence nulle pour cette raison. L'exemple suivant montre que les mots *reserve* et *a theft* appartiennent à deux syntagmes différents :

> *A man will face court next month charged with stealing three date palms from a Swansea* ***reserve, a theft*** *which sparked three months of community outrage.*

Le filtre sur les fréquences permet d'éliminer une partie des relations lexicales erronées. Néanmoins, ces problèmes concernent également des traductions correctes, et « bruitent » quelque peu les fréquences de l'API Google[1].

4.2 Perspectives d'amélioration du protocole

4.2.1 Changement de seuil de filtrage des relations lexicales françaises

Les relations lexicales françaises qui comptent moins de 10 occurrences au sein du corpus sont éliminées. Or, un certain nombre de relations lexicales correctes ont des fréquences inférieures à ce seuil. C'est le cas, par exemple pour l'usage sportif de *barrage* (*match de barrage, barrage aller, barrage retour,* etc.). L'algorithme *HyperLex* (Véronis, 2003, 2004) nous permettrait d'identifier les usages peu fréquents des mots (jusqu'à environ 1% des occurrences). Une amélioration consisterait à ajuster le seuil des relations lexicales françaises selon la fréquence de l'usage du nom concerné.

4.2.2 Description des patrons syntaxiques de l'anglais

Contrairement à notre méthode d'extraction des relations lexicales françaises, ne sont pris en compte que les patrons de « base » de l'anglais sans autres variations que celles de l'article dans le cadre de notre acquisition des traductions. Ces patrons « de base » ont donné un premier éventail de résultats qui ont permis d'évaluer notre méthode. Des améliorations ultérieures visent à augmenter le panel de patrons morpho-syntaxiques anglais des traductions candidates, en procédant à une analyse syntaxique du contenu des pages Web prospectées.

4.2.3 Variations morpho-syntaxiques au sein des requêtes anglaises

Une amélioration vise à considérer des variations morpho-syntaxiques au sein de nos requêtes telles que des variations dues aux formes verbales ou au pluriel, comme dans l'exemple :

> *"commit a theft", "commit the theft", "commit thefts" , "commit the thefts", "committed a theft",* etc.

5 Conclusion

Nous avons décrit une méthode d'acquisition semi-automatique de relations de traductions du français vers l'anglais, en montrant que le Web s'avère être un outil particulièrement efficace d'aide à la validation de traductions candidates. Les résultats sont particulièrement intéressants pour les patrons syntaxiques de type *NOM ADJECTIF* (précision de 97,1 %) et *VERBE NOM*(*objet*) (précision de 88,9 %). La méthode reste imparfaite pour le patron *NOM1 DE NOM2*, mais le taux de précision est honorable (76,9%), surtout étant donné la difficulté

[1] Une autre limite de l'API Google est de ne pouvoir lancer que 1000 requêtes par jour, ce qui impose des contraintes de temps. Depuis, mars 2005, l'API Yahoo a été lancé et offre une possibilité de 5000 requêtes par jour.

volontaire du banc de test choisi (mots très polysémiques). Même en l'état, la méthode permet un accroissement important des relations lexicales contenues dans notre base de données lexicale. Ces résultats constituent un premier échantillon significatif des possibilités qu'offre le Web en matière de validation de relations lexicales. Nos perspectives d'évolution concernent principalement la mise en place d'une analyse syntaxique des relations candidates en contexte, ainsi qu'une désambiguïsation des relations lexicales à traduire.

Références

Benson M. (1990), Collocations and general-purpose dictionaries, *International Journal of Lexicography*, vol. 3(1), pp. 23-35.

Cao Y., Li H. (2002), Base noun phrase translation using web data and the EM algorithm. In *Proceedings of CoLing*.

Chuquet H., Paillard M. (1987), *Approche linguistique des problèmes de traduction: anglais-français*, Gap, Ophrys.

Cruse D. A. (1986), *Lexical Semantics*, Cambridge, Cambridge University Press.

Katz J. J, Fodor J. A. (1964), The structure of a semantic theory, In J. A. Fodor and J. J. Katz, editors, *The Structure of Language,* chapter 19, pp. 479-518.

Kilgarriff A., Grefenstette G. (2003), Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics, 29*(3): 333-348.

Melamed I. D. (2001), *Empirical methods for exploiting parallel texts*, MIT Press.

Millon C. (2004), Acquisition de relations lexicales désambiguïsées à partir du Web, *Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2004),* Fès, (Maroc).

Morin E., Dufour-Kowalski S., Daille B. (2004), Extraction de terminologies bilingues à partir de corpus , *Actes de TALN'2004*, Fès (Maroc).

Smadja F. (1993), Retrieving collocations from text : Xtract, *Computational Linguistics*, Vol. 19, pp. 143-177.

Véronis J. (1998), A study of polysemy judgements and inter-annotator agreement, *Programme and advanced papers of the Senseval workshop,* pp. 2-4, Herstmonceux Castle (England).

Véronis J. (Ed.) (2000). Parallel Text Processing: Alignment and use of translation corpora, Kluwer Academic Publishers.

Véronis J. (2003), Hyperlex : cartographie lexicale pour la recherche d'informations, *Actes de TALN'2003,* pp. 265-274, Batz-sur-mer (France): ATALA.

Véronis J. (2004), *HyperLex : cartographie lexicale pour la recherche d'informations.*Rapport Interne Equipe DELIC, Université de Provence. [En ligne : http://www.up.univmrs.fr/veronis/pdf/2004-hyperlex-rapport.pdf]

Wilks Y. A. (1975), Preference Semantics, In: Keenan, E. (ed), *The Formal Semantics of Natural Language,* Cambridge University Press, pp. 329-348.

Linguistic representation of Finnish in the medical domain spoken language translation system

Marianne Santaholma
University of Geneva, ETI, TIM/ISSCO
40, bvd du Pont-d'Arve
CH-1211 Geneva 4, Switzerland
Marianne.Santaholma@eti.unige.ch

Mots-clefs – Keywords

Grammaire d'unification, traduction automatique multilingue de la parole, interlingue, sous-langage, finnois.

Domain specific unification grammar, multilingual spoken language translation, interlingua, sub-language, Finnish.

Résumé – Abstract

Dans cet article nous décrivons le développement des ressources linguistiques du finnois pour un système de traduction automatique de la parole dans le domaine médical: MedSLT. Le travail inclut la construction des corpus médicaux en finnois, le développement de la grammaire finlandaise pour la génération, le développement du lexique finlandais et la définition des règles de mapping interlingue-finnois pour la traduction multilingue. Nous avons découvert que le finnois peut être introduit dans l'architecture existante de MedSLT sans trop de difficultés. En effet, malgré les différences entre l'anglais et le finnois, la grammaire finlandaise a pu être créée en adaptant manuellement la grammaire anglaise originale. Les premiers résultats de l'évaluation de la traduction anglais-finnois sont encourageants.

This paper describes the development of Finnish linguistic resources for use in MedSLT, an Open Source medical domain speech-to-speech translation system. The paper describes the collection of medical Finnish corpora, the creation of a Finnish grammar by adapting the original English grammar, the composition of a domain specific Finnish lexicon and the definition of interlingua to Finnish mapping rules for multilingual translation. It is shown that Finnish can be effectively introduced into the existing MedSLT framework and that despite the differences between English and Finnish, the Finnish grammar can be created by manual adaptation from the original English grammar. Regarding further development, the initial evaluation results of English-Finnish speech-to-speech translation are encouraging.

1 Introduction

The basic architecture of a speech-to-speech translation system typically includes several components. Any speech-to-speech translation system requires at least a module for the source language speech recognition, a translation module which converts the recognised and parsed source language string into the target language, and a speech synthesis module for the target language output speech generation. These components may be based on different kinds of architectures. For example translations may be obtained using a variety of translation methodologies, like rule-based, statistical or example-based translation engines. In past years statistical methods have been commonly used in speech systems. This even to the point that it may have given the impression that rule-based methods are no longer relevant. The general success of statistical methods over rule-based methods is based principally on the general robustness of the statistical systems and on the overall easiness of system development. However in some special fields, like for example in the medical domain, the reliability of the system is more important than the general robustness of the system. This suggests that in these domains rule-based methods can be better suited (Knight et al., 2001). MedSLT is an Open Source project which is developing a generic platform for building this kind of rule-based system where reliability is a crucial issue (See Rayner, Bouillon, 2002, Rayner et al., 2004). To compare rule-based to statistical methods there exist two versions of the system, on based on grammar-based language modelling (GLM) and one on statistical language modelling (SLM). These versions are trained on the same corpus, and evaluated on a test corpus collected using both versions of the system. The experiments show that in terms of number of sentences translated, the GLM and SLM scored equally well. However, (Rayner et al., 2004) concluded that the GLM was preferable in terms of presenting a more predictable interface.

A rule-based spoken translation system implies several different resources: a description of the source language (SL) and of the target language (TL) and a set of translation rules, for example transfer rules or interlingua mapping rules. Since in general the development of linguistic resources used in translation systems is laborious and time consuming, in order to reduce the development effort needed for multilingual rule-based systems, we focus on developing general unification grammars that can be used for speech recognition, analysis, and generation. The main feature is that the general grammars will be automatically specialised for these different tasks with a corpus and an example-based learning method (Rayner et al, 2000). The grammar specialisation is necessary in order to compile the grammar into CFG form, to reduce the ambiguity of the grammar and to build the generation grammar.

This paper presents the development of linguistic resources for Finnish for the MedSLT system. The development includes the collection of the medical sub-domain corpora, the creation of the Finnish generation grammar and lexicon, and the definition of interlingua to Finnish mapping rules, used by the multilingual translation module. The interest of working on the Finnish language is that despite different natural language processing (NLP) projects including Finnish, it has not yet been used extensively in speech-to-speech translation systems. Another motivation is that as Finnish is not an Indo-European language, it does not necessarily share the same word and sentence structure with English and French. Therefore it allows the study of the grammar adaptation and the entire multilingual MedSLT system architecture including the MedSLT interlingua representation from a new perspective.

The paper is organised as follows. Section 2 describes the Open Source speech translation system MedSLT. Section 3 presents the Finnish module (sub-domain corpora, Finnish generation grammar and lexicon, and interlingua to Finnish mapping rules). Section 4 presents the evaluation of the MedSLT English to Finnish translation performance and Section 5 concludes.

2 The MedSLT system

MedSLT (MedSLT, 2005, Rayner et al., 2003) is a medical domain spoken language translation (SLT) system, which is developed to translate doctor-patient examination dialogue. Translation is one-way; the system translates the diagnosis questions asked by the doctor. The questions are formulated so that the patient can answer them non-verbally by nodding or shaking the head, by pointing at a body part or similar. The system coverage is organised into medical sub-domains by symptom classes. The current system sub-domains include the emergency relevant sub-domains of headaches, chest pains and abdominal pains, each supporting a vocabulary of between 300 and 500 words. The current system prototype translates from English into such structurally different languages as French, Japanese and Finnish. The system includes also initial versions of French-English, Japanese-English, Spanish-English and English-Spanish.

The basic architecture adopted in the MedSLT-system is a compromise between the fixed-phrase translation (e.g Phraselator, 2005) and the rule-based linguistic methods (Wahlster, 2000, Rayner et al., 2000). At runtime the system behaves like a phrasal translator, which translates beforehand defined patterns. In contrast, the compile time architecture is based on general linguistic resources. The grammars used in the MedSLT system are written in unification grammar formalism in a SICStus Prolog based feature-value notation. The unification grammars are compiled into grammar-based language models using the Open Source Regulus toolkit (Regulus, 2005) (figure 1: Regulus compile time component). Language models are in GSL form, suitable for use with the Nuance platform (Nuance, 2005). The translation is based on the interlingua approach of MT.

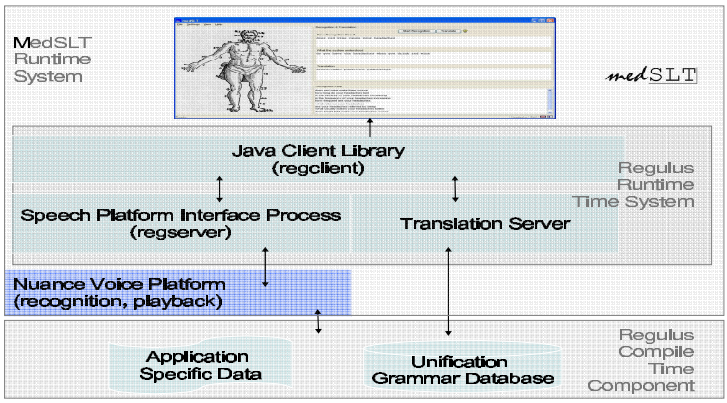


Figure 1 : MedSLT system architecture

The MedSLT runtime system is accessed through a GUI (illustrated in figure 1), which allows the simple utilisation of the system for the diagnosing doctor. The flow of information in the

MedSLT system is as follows. First the input speech is recognised using the recogniser built on the Nuance platform. The output of the recogniser is the semantic representation of the input produced by using the specialised grammar. This semantic representation of the SL is then passed to a discourse processing module, which interprets it in the context of the previous dialogue, in order to resolve possible ellipsis. The resolved SL representation is transformed into an SL independent interlingua representation. In the MedSLT interlingua representation each clause is treated as a flat list of attribute-value pairs (see section 3.4.). The interlingual form is transferred into a TL surface string using a generation grammar, and finally passed to a speech synthesis unit. The mapping of the SL dependent representation into interlingua and the mapping of interlingua into a TL dependent representation is obtained by manually developed interlingua mapping rules.

3 Finnish linguistic resources

3.1 Sub-domain corpora

The first step towards the development of the Finnish module for the MedSLT system was to create the Finnish headache and chest pain sub-domain corpora. These corpora serve as the primary source to decide what kind of structure rules and vocabulary is necessary to introduce to the Finnish module. The corpora were created by translating (and adapting) the original English corpora. The objective was to find the equivalent Finnish questions for the original English diagnosis questions. Since in the current MedSLT system Finnish is used only as output language it was not regarded necessary, at this point, to take into consideration the other possible questions a Finnish doctor might want to include in the system coverage, or the different variations of the same question. Therefore it was justified to translate the original English corpora into Finnish instead of collecting authentic Finnish data. The translated Finnish diagnosis questions were, however, revised by Finnish medical doctors (Santaholma, 2005).

Two essential issues were taken into consideration when translating the diagnosis questions into Finnish: the particular character of spoken language and the special situation in which the utterances were intended to be used. The spoken language style differs markedly from the written style. Generally the spoken language is more informal and commonly contains the use of ill-formed language, such as incomplete sentences, wrong word cases, and unusual word order. This special character of spoken language influenced the content of the Finnish corpora and consequently the structure and lexical rules of the Finnish MedSLT grammar. In whole the comprehensibility, reliability and simplicity of the utterances were regarded to be more important than the actual formulation or style of the sentences. In the context of medical examination it is important that the patient feels comfortable and confident. Even more so if the questions are asked by a doctor speaking a language the patient does not understand and if he/she is listening to the translations of the questions spoken out by a machine. Thus the output of the MT system should sound as natural as possible. For the Finnish output the aim was to preserve the simplicity of the original English questions without letting the translation be influenced too much by the expressions and the structure of the SL.

The current Finnish MedSLT headache corpus consists of 170 utterances and the chest pain corpus of 187 utterances. The concepts of these two corpora overlap considerably, subsequently so does the structure of the diagnosis questions. In most cases the questions of

the sub-domains differ only in the vocabulary. The system input languages -like English- include commonly some variation in the way the questions can be posed, which makes the system more practical to use since the doctor is not obliged to remember the exact formulation of the questions but rather the main concepts of the questions. For the output language this variation is not necessary. The English question variants corresponding to one concept in the corpora are translated into Finnish by the same utterance. Due to this, the Finnish corpora are slightly more restricted in comparison to the SL corpora.

3.2 Finnish MedSLT grammar rules

The MedSLT Finnish generation grammar is so far a domain specific grammar for speech adapted from the general Regulus English grammar used in the MedSLT system (Regulus, 2005). Currently the Finnish grammar contains 57 grammar rules and around 530 lexical entries. The current grammar rules cover the basic constructions, which are necessary for the MedSLT headache and chest pain sub-domains. The grammar includes syntactic rules for declarative, interrogative and elliptical clauses, formation of yes/no questions using subject-predicate inversion, wh-questions, clause lacking the grammatical subject (replaced by the object of the phrase), rules for various kinds of nominal phrases and verbal phrases (like transitive and intransitive phrases), rules for adjectival modifiers, including comparatives, passive sentences, sentences with past-participles, and rules for different verb and sentence modifiers like adverbial modifiers and adverbs. The MedSLT Finnish generation grammar is more limited than the standard Finnish grammar regarding the variety of constructions the grammar includes. However the grammar does not contain particular structure rules that would be considered being merely specific constructions of a medical domain sublanguage. The syntax reduction in the range of constructions does rather reflect the specific text type and discourse of the domain than the domain specific language itself. Furthermore, we believe that a specialised grammar is not solely domain specific but is also constructed after a particular discourse type. (Santaholma, 2005)

```
vp:[sem=concat(Vbar, concat(Advp, Np)), vform=Vform, subcat=A, inv=Inv, agr=Agr,
subj_n_case=Case, np_n_type=nonsubj, subj_sem_n_type=SubjType, gapsin=null, gapsout=null] -->
        vbar:[sem=Vbar, vform=Vform, inv=Inv, subcat=(trans\/personal), subcat=A, agr=Agr,
np_n_type=nonsubj, subj_n_case=Case, subj_sem_n_type=SubjType, obj_sem_n_type=ObjType,
obj_case=B, takes_adv_type=AdvpType],
        ?advp:[sem=Advp, sem_adv_type=AdvpType],
        np:[sem=Np, wh=n, agr=Agr, sem_n_type=ObjType, n_type=nonsubj, case=(ptv\/nom),
case=B, gapsin=GIn, gapsout=GOut].
```

Figure 2 : Finnish transitive verb phrase rule

The natural languages appear to have quite a lot of common structure. Consequently the exhaustive grammars of different languages share structural rules and properties at least to some point. During the Finnish grammar development was discovered that the basic English structures were relatively easy to adapt to corresponding Finnish constructions. This at least when using as a reference a grammar that covers similar kinds of systematic patterns of the same restricted discourse type. When comparing the MedSLT English and Finnish grammars, most of the Finnish rules are very similar to the English counterparts from which they have been adapted. When adapting the English grammar the most significant difference between Finnish and English is that in Finnish more phenomena are resolved at morphology level rather than in the syntax like in English. Finnish is a highly agglutinative language, in which

nouns, adjectives, pronouns and numerals inflect in (around) 15 cases. Therefore an essential feature in the Finnish MedSLT grammar rules is the feature 'case'. For example in the Finnish verbal phrase rule used for generating clauses including a transitive verb the allowed inflectional case of the subject and the object of the utterance are defined (figure 2). This is necessary in order to prevent the over-generation. Furthermore, in Finnish the different grammatical functions as well as time, place, ownership, manner etc. for which English normally uses a preposition are expressed by suffixes. The correspondence of the Finnish cases with the English prepositions is, however, not exactly straightforward. As a whole, Finnish is a very complex and productive language regarding morphology whereas the syntax is rather straightforward and free to certain point.

3.3 Lexicon and lexical entries

The Finnish MedSLT lexicon currently includes around 530 distinct Finnish lexical entries covering the MedSLT headache and chest pain sub-domains. However, it is noteworthy that the different inflections of the same Finnish entry are counted as distinct lexical entries. Therefore, the actual total of different Finnish lemmas is slightly smaller than the figure may indicate. The Finnish lexicon includes rules for the common part-of-speech categories – i.e. for verbs, nouns, adjectives, adverbs, specifiers, wh-question words, post-positions and for prepositions. The multiword expressions (~lexicalised NPs) that define the sentence or the verb of sentence are placed under the category of adverbials.

The Finnish lexical entries include a fairly comprehensive amount of different information. The features defined for instance in the verb entries include, - among others - the verb type, the sub-categorisation, semantic type of the possible subject, object, predicative, adverb and adverbial, as well as the allowed inflectional cases of these constituents in the context of the verb in question (figure 3). The Finnish verbs inflect in tense, mode and person.

```
verb:[sem=[[event, lievittää], [tense, present]], vform=q_ko, agr=sg, subcat=trans,
subj_n_case=nom, subj_sem_n_type=(cause\/activity), obj_sem_n_type=perception_body,
obj_case=ptv, takes_adv_type=frequency] --> lievittääkö.
```

Figure 3 : Finnish verb entry. The question form of the verb 'lievittää'; '*to relieve*', in the present, third person singular.

As a consequence of the considerable amount of the different inflectional cases, the amount of different word forms of the same lexical entry may be quite extensive in the Finnish lexicon. An advantage of a limited domain application, like MedSLT system, is that the amount of distinct word forms necessary in the application is restricted. The lexicon is actually possible to write manually (Morphological tools like Mmorph (Petitpierre/Russell, 1995), or PC-Kimmo (Koskenniemi, 1983) are not integrated in the current MedSLT system). Evidently the enumeration of all the possible inflectional cases for every lexical entry is laborious and contains a lot of repetition. However the encountered repetition may be decreased to a certain point by the systematic use of macros in the lexical rules. The macros are extensively used in the MedSLT English lexicon. The Finnish lexicon currently contains macros mainly in adjective and noun entries.

3.4 Interlingua-Finnish mapping rules

The interlingua mapping rules enable the transformation of the **a)** SL representation through **b)** Interlingua into the **c)** TL representation. For example if we want to translate the English utterance '*Is the headache made worse by red wine*?' in Finnish "*Pahentaako punaviini päänsärkyä*?"; (*make_worse red wine headache?), we first need to write rules to transfer the English source representation:

a) source_representation=[[adj,worse],[cause,red_wine],[event,make_adj],[prep,subj], [secondary_symptom,headache],[spec,the_sing],[tense,present],[utterance_type,ynq], [voice,passive]]

into the corresponding interlingua representation:

b) interlingua=[[sc,when], [clause,[[utterance_type,dcl], [pronoun,you], [tense,present], [voice,active], [action,drink], [cause,red_wine]]], [event,become_worse], [symptom,headache], [tense,present], [utterance_type,ynq], [voice,active]]

After that we still need to develop rules for transferring the Interlingua representation into the Finnish target representation:

c) target_representation=[[cause,punaviini], [event,pahentaa], [symptom,päänsärky], [tense,present], [utterance_type,ynq]]

MedSLT makes use of two types of interlingua rules: **transfer_lexicon** rules and more complex **transfer_rules.** The previous ones, the **transfer_lexicon** entries, are employed when there is one-to-one correspondence between the interlingua expression and the natural language expression. In practice, both, the source part and the target part of the rule, contain only one element. **Tranfer_rule** entries map together several elements.

The MedSLT interlingua representation of an utterance is mostly based on the flat list of semantic features obtained in the analysis. Only some causal and temporal structures are represented as slightly nested structure (like above 'Is the headache made worse by red wine?'). This kind of representation is possible in the restricted domain like the one of MedSLT. Corresponding the character of the application, the MedSLT interlingua is aimed to be easily portable to new medical sub-domains. Furthermore, the mapping rule development is desired to be as straightforward as possible for every interlingua □ natural language pair.

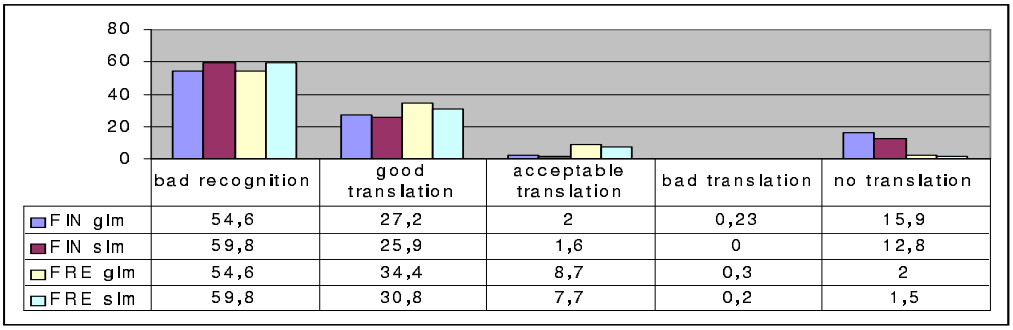
The interlingua-Finnish mapping rules currently enable the translation from other MedSLT system languages into Finnish in the headache sub-domain. The nested structures for causal and temporal expressions are not yet implemented in Finnish but the current generated Finnish semantic representations of utterances are based solely on the flat representations. In whole, the interlingua representation is more atomic than the actual Finnish target representation. The Finnish output representation resembles in fact more the English source representation. Thus interlingua-Finnish mapping rules contain a lot of complex **transfer_rules** in order to map the different interlingua and Finnish target language structures. The advantage of the more complicated transfer rules is that the word context is included in the rule. The disadvantage is that if the context is always required the translation may lose robustness.

4 Evaluation of the translation

The translation performance of the MedSLT English-Finnish language pair was evaluated on unseen data and the obtained results were compared with the corresponding results of the English-French language pair. The (speech) data used for the evaluation was collected during November 2004 in twelve data collection sessions on the headache sub-domain. A total of 870 spoken utterances were collected. For the recognition of English input were used both GLM and SLM based versions of the English recogniser (Recognition results are analysed and described in detail in Rayner et al, 2004, Rayner et al, 2005). The correctly recognised English sentences (judged by English native speakers) were translated into Finnish and the acceptability of these translations were judged by 3 Finnish native speakers with grades of 'good' (semantically and grammatically correct sentence), 'acceptable' (semantically correct translation) and 'bad' (semantically and grammatically incorrect sentence).

The translation performance into Finnish was somewhat weaker than into French but comparable if taking into consideration the non-translated sentences (figure 4). Out of the correctly recognised utterances (395 utterance; 45,4% of a total of 870 utterance) 60% of Finnish translations were judged as 'good', 4,4% of translations were assessed as 'acceptable' and 0,5% as 'bad'. The corresponding figures for French were 'good' 75,8%, 'acceptable' 19,2% and 'bad' 0,7%. Generally the Finnish judges graded the translation as 'bad' if it contained a word in the wrong inflectional case -even if the word itself was correct. The utterances judged as 'acceptable' contained mostly special medical terminology or particular expressions describing the pain that were not familiar for the judges.

The most remarkable difference between the Eng-Fin and Eng-Fre translation performance was thus the amount of utterances left without translation (see figure 4: 'no translation'): of correctly recognised English utterances 36% were not translated into Finnish, whereas only 4,4% of utterances were left without translation into French. When analysing the sentences that were not translated into Finnish it was noticed that in most cases the translation failed because the Finnish lexicon either lacks a lexical entry or a certain form (inflectional case, verb tense/person) of the lexical entry (lexical gaps). Even if the lexicon contained the word in some form, the grammar prevents the generation of sentences using in-correct word forms. Furthermore the un-translated sentences were mainly not in coverage sentences (Proportion of not in coverage 453 (52.1%) and in coverage 417 (47.9%) utterances in corpus of total of 870 sentences).



	bad recognition	good translation	acceptable translation	bad translation	no translation
FIN glm	54,6	27,2	2	0,23	15,9
FIN slm	59,8	25,9	1,6	0	12,8
FRE glm	54,6	34,4	8,7	0,3	2
FRE slm	59,8	30,8	7,7	0,2	1,5

Figure 4 : Comparison of English-to Finnish and English-to French translation performance

The following examples-show lexical gaps: *"Does the pain radiate to the neck?"* (in coverage sentence) and *"Is the pain in the neck?"* (not in coverage sentence). The Finnish lexicon

includes the word "kaula"; *'neck'* in the ablative case, which is used in the system in the context of the verbs *'to radiate'* and *'to spread'*. A translation gap is produced when trying to translate the utterance *"Is the pain in the neck"*, where the verb "olla"*; 'to be'*, requires the adessive case of the word neck. The same problem is encountered, among others, in the sentences *"Does your headache extend to the back?"* and *"Does the pain spread to your eye?"* The Finnish lexicon does not include the words *'back'* and *'eye'* in the inflectional cases required by the verb context and the utterances are left without translation even if the system translates the words correctly in utterances *"Is the pain above the eye"* and *"Is the pain in the back"*. In some cases the translation was also unsuccessful because of the lack of needed grammar rules. Because of a lacking grammar rule sentences like the following were left without translation: *"Do you have nausea when you have headaches?"* (subordinate structure); *"Do your headaches come after anxiety?" / "Do you get the headache after drinking red wine?"/ "Is the pain relieved after sleep?"* (post-positional structure)

As a whole the acceptability of Finnish translations is comparable to the French, and in general the Finnish translations are comprehensible and thus acceptable. Most of the work to be done now is on the coverage of the Finnish grammar and lexicon.

5 Conclusion

This paper has described the development of Finnish linguistic resources for use in MedSLT, an Open Source medical domain speech-to-speech translation system. The development was partly done by adapting the already existing resources, and in particular the Finnish grammar was created by grammar adaptation from the original English grammar. The grammar adaptation was proved to be an efficient way to develop the Finnish MedSLT grammar. The syntax rules were mostly highly similar with the original English grammar rules they were adapted from. Most difficulties were caused by the complex morphology of Finnish. To avoid the generation of non-grammatical sentences the grammar and lexicon rules have to be carefully constrained. The manual enumeration of the lexical entries and the different inflectional cases of the words is laborious but still feasible by the use of macros in the restricted domain application like MedSLT. In more general domains, the use of integrated morphology tools is preferable.

The evaluation of the translation performance of English-Finnish language showed encouraging results and by some changes in the coverage of grammar and lexicon the translation result will be improved and eventually the Finnish module will be more robust. This also confirms that the MedSLT system architecture as a whole is adaptable on restricted domain to translate between multiple different languages.

Acknowledgements

I would like to thank the developers of the original MedSLT system framework Pierrette Bouillon and Manny Rayner for all their help and advice.

References

Knight S., Gorrell G., Rayner M., Milward D., Koeling R., Levin I. (2001), Comparing grammar-based and robust approaches to speech understanding: a case study, In *Proceedings of Eurospeech 2001*, Aalborg, Denmark, 1779–1782.

Koskenniemi K. (1993), *Two-level Morphology: A General Computational Model for Word-Form Recognition and Production*, University of Helsinki, Department of General Linguistics.

MedSLT (2005), https://sourceforge.net/projects/medslt/. As of 31 January 2005.

Nuance (2005), http://www.nuance.com. As of 15 January 2005.

Petitpierre D., Russell G. (1995), *MMORPH-The Multext Morphology Program*, Version 2.3: October 1995.

Phraselator (2005), http://www.phraselator.com. As of 15 January 2005.

Rayner M., Carter D., Bouillon P., Digalakis V., Wirén M. (2000), *The Spoken Language Translator*, Cambridge, Cambridge University Press.

Rayner M., Bouillon P. (2002), A flexible Speech to Speech Phrasebook Translator, In *Proceedings of ACL-02 Workshop on Speech-to-Speech Translation: Algorithms and Systems*, Philadelphia, 69-76.

Rayner M., Bouillon P., Dalsem Van V., Hockey B.A., Isahara H., Kanzaki K. (2003), A limited-domain English to Japanese medical speech translator build using REGULUS 2, In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (demo track), Sapporo, Japan, 137-140.

Rayner M., Bouillon P., Hockey B. A., Chatzichrisafis N., Starlander M. (2004), Comparing Rule-Based and Statistical Approaches to Speech Understanding in a Limited Domain Speech Translation System, In *Proceedings of TMI 2004*, Baltimore, MD USA, 21-29.

Rayner M., Hockey B A., Bouillon P. (2005), Using Regulus, http://cvs.sourceforge.net/viewcvs.py/regulus/Regulus/doc/RegulusDoc.htm. As of 31 January 2005.

Bouillon P., Rayner M., Chatzichrisafis N., Hockey B. A., Santaholma M., Starlander M., Nakao Y., Kanzaki K., Isahara H. (2005) A Generic Multi-Lingual Open Source Platform for Limited-Domain Medical Speech Translation, In *Proceedings of EAMT 2004*, Budapest, Hungary. Forthcoming.

Santaholma M. (2005), *Linguistic representation of Finnish language in speech-to-speech translation system*, Masters thesis. Geneva University, Department of translation and interpretation.

Wahlster W. (Ed.) (2000), *Verbmobil: Foundations of Speech-to-speech Translation*, Berlin, Heidelberg, New York, Springer-Verlag.

Constitution d'un corpus de français tchaté

Achille Falaise

GETA, CLIPS-IMAG – UJF - Université Grenoble I
385, rue de la Bibliothèque, B.P. 53, 38041 Grenoble Cedex 9
achille.falaise@imag.fr
Date prévue de la thèse : janvier 2008

Mots-clefs – Keywords

langue tchatée, ressources linguistiques, collecte de données

chat language, linguistic resources, resource acquisition

Résumé – Abstract

Nous présentons dans cet article un corpus de français tchaté, destiné à l'étude de la langue du tchat. Ce corpus, collecté et encodé automatiquement, est remarquable avant tout par son étendue, puisqu'il couvre un total de 4 millions de messages sur 105 canaux, hétérogènes sur les plans thématique et pragmatique. Son codage simple ne sera toutefois pas satisfaisant pour tous les usages. Il est disponible sur un site Internet, et consultable grâce à une interface web.

We present in this article a french chat corpus, intended for the study of chat language. This corpus, automatically collected and coded, is especially remarkable for its extent, since it covers a total of 4 million messages on 105 channels, heterogeneous from a thematic and pragmatic point of view. Its simple coding will not, however, be sufficient for all purposes. It is available on an Internet site, and viewable using a web interface.

Introduction

Alors que de nouveaux outils de communication écrite synchrone, tels que les salons de discussion, la messagerie instantanée et le texto, connaissent un essor indéniable depuis quelques années, leurs spécificités linguistiques, pourtant reconnues, restent encore peu étudiées dans le détail. Il est vrai que le manque de ressources les concernant, tout au moins pour la langue française, ne fait rien pour en faciliter l'étude. Nous nous intéressons ici plus particulièrement à la langue du tchat, la « langue tchatée », produite à l'aide d'outils de tchat tels que les salons de discussion ou les différents logiciels de messagerie instantanée, laissant ainsi de côté la langue des textos, que nous supposons assez différente. Aujourd'hui, toute étude de la langue du tchat passe par la constitution d'un corpus, mais les contraintes de temps font que, bien souvent, ce dernier est assez réduit, aussi bien du point de vue de la longueur, que du nombre d'utilisateurs impliqués, ou encore des thèmes abordés; ce qui peut amener à s'interroger sur la portée réelle des résultats obtenus. Pourtant, ces nouveaux outils de

communication, moins normatifs que l'écrit traditionnel, offrent une opportunité de jeter un regard nouveau sur la langue écrite.

Au printemps 2004, au cours d'un stage de M2R[1] portant sur la traduction automatique de tchat (Falaise, 2004), un important corpus de français tchaté a été collecté, afin d'évaluer les difficultés posées par la la langue du tchat à son traitement automatique. Nous souhaitons contribuer à l'étude de cette langue en mettant ce corpus à disposition. Après un bref aperçu des principes du tchat et des caractéristiques du « français tchaté »[2], nous présenterons donc ce corpus, depuis sa méthode de collecte originale, jusqu'à son mode de diffusion.

1 Qu'est-ce que le tchat ?

1.1 Les outils de tchat

Nous considérons, à la suite de (Latzko-Toth 2001), que les dispositifs tels que les salons de discussion et les messageries instantanées peuvent être regroupées sous l'appellation « d'outils de tchat ». Les salons de discussion se confondent généralement avec le principal protocole sur lequel ils s'appuient, à savoir le protocole IRC. Il existe de nombreux réseaux IRC, et chaque réseau, se divise en canaux (ou salons) indépendants les uns des autres, et souvent associés à un thème précis. La messagerie instantanée, qui se distingue des salons de discussion, ouverts à tous, par son caractère privé, repose quant à elle sur un grand nombre de protocoles, chaque réseau ayant recours au sien.

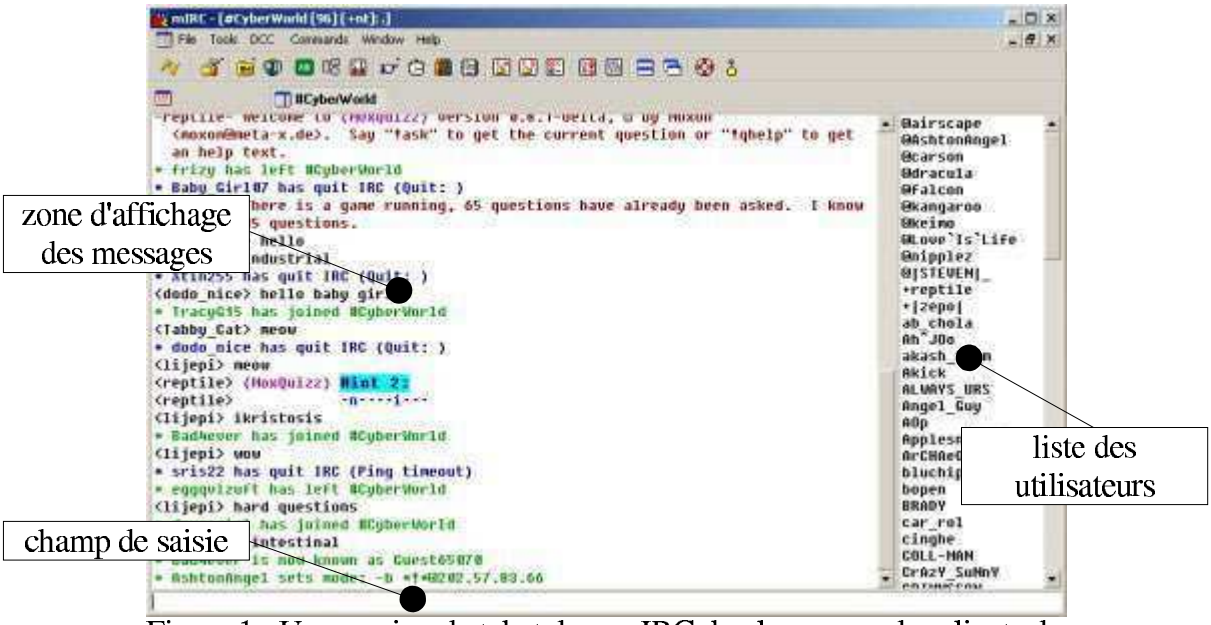


Figure 1 : Une session de tchat dans mIRC, le plus connu des clients de tchat pour le réseau IRC.

Ces dispositifs se distinguent notamment d'autres outils informatiques, tels que le forum ou le courriel, par leur caractère synchrone; en cela, ils se rapprochent plus du texto. De plus, comme ce dernier, le tchat est *volatil*. Le contenu d'une session de tchat, à l'instar d'une conversation verbale, n'a pas vocation à être enregistrée. Ainsi, lorsqu'un utilisateur se

1 Seconde année de master recherche.

2 Pour reprendre l'expression de (Pierozak, 2003).

connecte, il est dans l'ignorance totale de ce qui a été dit avant son arrivée, de même qu'il ne pourra pas savoir ce qui se dira après sa déconnexion. Mais à la différence du texto, les outils de tchat proposent aux utilisateurs des espaces communs, qui rendent aisée la communication entre de nombreuses personnes, là où le texto peut difficilement mettre en relation plus de deux personnes. De plus, il faut garder à l'esprit que les textos sont facturés au message, ce qui incite l'utilisateur à être synthétique et à en envoyer le moins possible, alors qu'au contraire le tchat, moins contraignant de ce point de vue, autorise toutes les digressions.

1.2 La langue du tchat

Les principales caractéristiques de la langue du tchat sont présentées, notamment, dans (Pierozak, 2003) et (Guimier de Neef & Véronis, 2004), et on n'en exposera donc ici que les grands principes.

Le « français tchaté », ou « clavardage » pour reprendre le terme québécois, se caractérise en particulier par une syntaxe proche de la langue parlée, et surtout par sa graphie originale. Loin d'être une limitation, le caractère écrit des conversations de tchat semble en effet en être l'un des principaux attraits (Herring, 1999, et Latzko-Toth, 2001). En fait, dans la langue du tchat, la graphie d'un lexème semble plus relever de la fantaisie de son auteur que de la norme orthographique, selon un processus de création lexicale permanente que (Pierozak, 2003) qualifie de « ludogénèse », et suffisamment souple pour permettre à certains utilisateurs de développer leur propre « voix » graphique.

Comme le souligne (Pierozak, 2003), la syntaxe des énoncés de tchat tient beaucoup de l'oral. Entre autres phénomènes propres à la langue parlée, les topicalisations y sont fréquentes, ainsi que les constructions du type *situation* + *thème* + (*rhème*). En outre, pour éviter les messages trop longs, les usagers découpent souvent leurs messages en propositions, ce qui n'est pas sans rappeler le découpage en groupes prosodiques (Falaise, 2004).

Pour caractériser les spécificités lexicales du tchat par rapport à l'écrit « standard », on pense bien entendu aux émoticons (« :-) », « ^^ », etc.) et aux abréviations, fréquentes en tchat, et parfois spécifiques à ce mode de communication (comme « lol » , « mdr », « tlm », etc.). On relève par ailleurs fréquemment une graphie que l'on pourrait qualifier de phonétique (« salut les zamis »), et qui sert souvent à transcrire des variantes phonolgiques (« kikoo » pour « coucou », « oki » pour « okay », etc.) : il s'agit d'une modification de la graphie d'un mot, destinée à lui donner une prononciation légèrement différente de la norme. Parfois, ces variantes phonologiques correspondent à des allongements vocaliques, comme dans « kikooooooo » par exemple. Comme l'a fait remarquer (Guimier de Neef & Véronis, 2004), on voit réapparaître en tchat des phénomènes de créativité phonético-graphique que l'on pensait réservés aux systèmes d'écriture anciens (hiéroglyphes égyptiens, anciens sinogrammes, glyphes mayas, entre autres...) : l'insertion de signes autonomes, porteurs d'une signification propre, dans des mots, d'après leur valeur phonétique et sans tenir compte de leur valeur sémantique. Par exemple « 2m1 », « 2main » et « dem1 »[3] pour « demain ». On peut aussi relever des graphies résultant d'une créativité sémantico-graphique, telles que « Micro$oft »[4] pour « Microsoft », dans lesquelles on insère un signe possédant des valeurs lexicale et sémantique propres, qui sont cette fois toutes deux conservées dans la graphie ainsi formée. Ces phénomènes de créativité graphique, bien que bien connus et pour certains aussi anciens que l'écriture elle-même, ne sont pas présents dans la langue écrite normée moderne, et sont donc généralement négligés en TALN.

[3] Respectivement 163, 31 et 137 occurrences de ces formes dans notre corpus.

[4] 14 occurrences dans notre corpus.

La figure 2 donne une idée générale de l'importance de ces divergences lexicales, à défaut d'une analyse plus fine. On constate que les deux tiers environ des « mots » (caractères entre deux espaces, y compris émoticons) sont correctement orthographiés, et que les mots involontairement mal orthographiés (« orthographe mal formé ») sont rares. Par contre, un nombre significatif de mots relèvent de la graphie phonétique (« orthographe phonétique ») décrite au paragraphe précédent. Le recours aux abréviations, aux émoticons, aux onomatopées, ainsi que les références aux autres utilisateurs, se révèlent aussi assez courants. Enfin, on relève quelques cas de xénismes (des anglicismes en l'occurrence) et de fusions de mots (« jme demande », « jte dis », « ça mva », etc.).

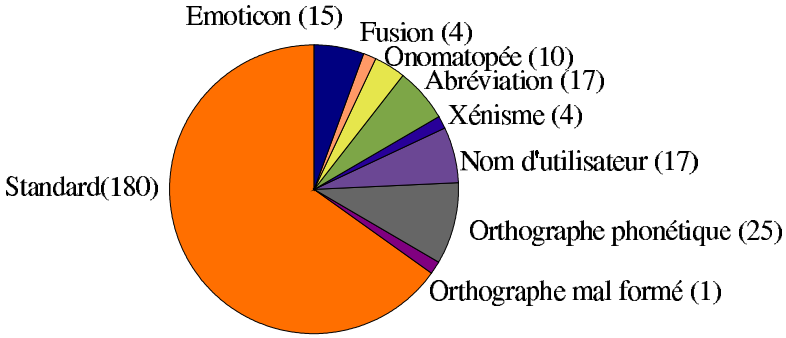


Figure 2 : répartition des principaux phénomènes lexicaux dans un ensemble de 77 messages déterminés aléatoirement au sein du canal #18-25ans; nombre d'occurrences entre parenthèses.

Il faut souligner néanmoins que malgré toutes ces divergences par rapport à l'écrit « standard », la finalité du tchat demeure le dialogue, et les tchateurs sont soucieux, jusqu'à un certain point, de la clarté de leur messages. En témoignent les corrections effectuées *a posteriori* par les tchateurs eux-même, généralement lorsqu'un mot mal orthographié peut être confondu avec l'un de ses homophones hétérographes (« s/pere/paire/ », vu sur le canal #c++).

Ainsi, l'un des intérêts scientifiques de l'étude de la langue du tchat tient au fait qu'elle vient remettre en cause certains *a priori* fréquents en TALN, comme par exemple l'approche scholastique de la notion de grammaticalité ou encore le caractère fermé du lexique.

2 Collecte du corpus

2.1 Ethique

La collecte d'un corpus tchaté ne va pas sans soulever certaines questions éthiques. En effet, même en nous limitant aux salons de discussion, publics, il s'agit tout de même d'enregistrer des conversations en principe éphémères. Et à moins de créer un canal dédié à la collecte d'un corpus, il n'est pas possible de prévenir tous les utilisateurs du fait qu'ils vont être enregistrés.

Nous avons donc choisi de constituer ce corpus à partir d'enregistrements de salons de discussion, librement consultables sur Internet, plutôt que d'enregistrer nous-même des sessions de tchat. Un autre avantage de cette méthode est qu'elle permet de récupérer en une seule fois une grande quantité de données.

2.2 Droit d'auteur

Cette approche ne lève pas toutefois tous les doutes qui se posent au niveau des droits d'auteur. Pour certains canaux, représentant environ 15% des messages du corpus, les conditions d'utilisation[5] stipulent que les enregistrements sont consultables par tous et reproductibles : « tout utilisateur accepte que les propos qu'il tient sur les canaux officiels puissent être visibles et transmis ». Rien n'est précisé pour les autres canaux. Le code de la propriété intellectuelle (CPI) se montre peu clair en ce qui concerne les dialogues publics anonymes. En admettant que les dialogues de tchat soient protégés par le droit d'auteur, ce qui n'est pas évident au vu de l'article L.112-2 du CPI[6], il semble difficile de considérer chaque message comme une œuvre à part entière, à moins de considérer des textes tels que « :-) », « salut » ou « ouaip » comme des œuvres originales. De plus, dans un tel contexte de dialogue, les messages peuvent difficilement être compris sans se référer à leur contexte. Par conséquent, on pourrait plutôt considérer chaque canal de tchat comme une œuvre collective[7] anonyme[8]. Selon l'article L.113-6 du CPI, les œuvres anonymes sont gérées par leur éditeur, c'est à dire en l'occurrence, soit le responsable du serveur de tchat, soit celui du site sur lesquels les enregistrement de tchat sont publiés. Ce problème, et surtout sa solution, nous étant apparus assez tard, nous cherchons actuellement à déterminer laquelle de ces deux personnes est, aux yeux de la loi, dépositaire des droits d'auteur, afin de la contacter pour obtenir une autorisation de publication en bonne et due forme.

2.3 Collecte

Notre corpus de langue tchatée est constitué à partir des enregistrements disponibles sur le site http://www.botstats.com. Ce site publie les résultats du service web Botstats, dédié aux canaux de tchat, et qui permet notamment la tenue de statistiques et l'archivage des discussions pour une durée maximale de trois mois. Ce service est utilisé par quelques centaines de canaux du réseau IRC EpikNet. La publication des enregistrements est un choix de la part du créateur du canal; et ce dernier peut en outre restreindre leur consultation aux utilisateurs enregistrés. Toutefois, quelques créateurs de canaux (une centaine) ont décidé de les rendre accessibles à tous, et ce sont les enregistrements de ces canaux que nous avons regroupés au sein du corpus.

Les enregistements, consultables sous forme de pages HTML sur Internet, sont tout d'abord extraits à l'aide d'un « aspirateur de sites », puis les fichiers HTML obtenus sont convertis automatiquement en XML grâce à des expression régulières, afin de pouvoir être exploités.

2.4 Format des données

L'activité d'un canal de tchat peut être représentée par une succession de messages, produits par différents auteurs. Outre les messages « normaux », rédigés par un auteur humain à destination de lecteurs humains, il faut distinguer quelques cas particuliers :

---

5 http://www.epiknet.org/legal/

6 Cet article, qui décrit ce qui est protégé par le droit d'auteur, n'est toutefois pas restrictif.

7 Au sens de l'article L.113-2 du CPI.

8 Au sens de l'article L.113-6 du CPI.

- les commandes, qui sont destinées au serveur (afficher la liste des utilisateurs par exemple) ou à un robot (sur un canal de tchat, un robot, ou « bot », a le statut d'utilisateur et peut par exemple intervenir pour rappeler le thème du canal, donner l'heure, gérer des jeux, mener des dialogues de type Eliza[9], etc.), et qui appartiennent à un langage formel;

- les messages pré-enregistrés, déclenchés à l'aide de raccourcis clavier ou lors de certains événements (déconnexion de l'utilisateur par exemple), qui ne relèvent pas du même contexte pragmatique;

- les messages envoyés par des robots;

- les événements, notifiés par le serveur (quelqu'un vient se connecter, de changer de surnom, etc...).

Le corpus est codé en XML. L'élément racine <*log*>, peut avoir quatre types d'éléments-fils :

- l'élément <*commentaire*>, dont le contenu est un commentaire sur le canal;

- l'élément <*message*>, comportant un message envoyé par un utilisateur, humain ou non, et destiné à être lu par les autres utilisateurs humains;

- l'élément <*commande*>, comportant une commande destinée au serveur;

- l'élément <*evenement*>, dont le contenu est un événement notifié par le serveur.

Les éléments <*message*>, <*commande*> et <*evenement*> possèdent des attributs *date* et *heure*. Les éléments <*message*> et <*commande*> comportent en outre un sous-élément <*auteur*>, contenant le surnom de l'utilisateur ayant produit le message, ainsi que des attributs indiquant son type, humain ou robot. <*evenement*> comporte quant à lui des sous-éléments précisant le type d'événement et simplifiant leur traitement automatique.

Le corpus, encodé automatiquement, respecte ces spécifications, à deux exceptions près. La valeur des attributs *type* (utilisateur humain ou robot), qui ne peut être déterminée automatiquement, puisque rien ne distingue formellement les messages d'un humain de ceux d'un robot, n'est pour l'instant pas renseignée. De plus, un certain nombre de commandes n'ont pas été reconnues par les expressions régulières chargées de les identifier, et ces dernières devront être affinées. En effet, les commandes peuvent généralement être reconnues par l'expression régulière ^!.+ , correspondant à une ligne débutant par un point d'exclamation, mais certains canaux proposent des commandes supplémentaires, dont la syntaxe est différente, et qui ne seront par conséquent pas détectées comme telles. Inversement, il arrive parfois, bien qu'assez rarement, qu'un message normal débute par un point d'exclamation : ce message sera alors considéré à tort comme une commande. Il convient donc d'élaborer une nouvelle expression de recherche pour chaque canal, en tenant compte de la liste des commandes et de leur syntaxe exacte.

[9] « ELIZA est un célèbre programme informatique écrit par Joseph Weizenbaum, qui simulait un psychothérapeute rogérien en reformulant la plupart des affirmations du "patient" en questions, et en les lui posant. » (Wikipédia, http://fr.wikipedia.org/wiki/ELIZA)

	<log xml:lang="fr">
	<commentaire>Exemple</commentaire>
A> soir tlm	<message date="29/03/2004" heure="15:13">
	<auteur type="humain">A</auteur>
	soir tlm
	</message>
A> kikooooooooB :*)	<message date="29/03/2004" heure="15:20">
	<auteur type="humain">A</auteur>
	kikooooooooB :*)
	</message>
B> kikoooA :)	<message date="29/03/2004" heure="15:20">
	<auteur type="humain">B</auteur>
	kikoooA :)
	</message>
<C vient de se connecter>	<evenement date="29/03/2004" heure="15:25">
	<connexion>
	<utilisateur>C</utilisateur>
	</connexion>
	C vient de se connecter
	</evenement>
	</log>

Exemple 1 : exemple de discussion et code XML correspondant.

3 Première évaluation des résultats

3.1 Quantification du corpus

D'un point de vue quantitatif, la somme de données collectées est assez considérable : 4 192 033 messages, couvrant environ 3 mois de conversations sur 105 canaux de tchat. Si l'on considère un mot comme une suite de caractères délimitée par les signes de ponctuations traditionnels (cette définition n'est pas forcément la plus adaptée à la langue du tchat, mais est acceptable en première approche), alors le corpus comporte 23 011 876 mots, soit une moyenne d'environ 5,5 mots par message. De ce point de vue, ce corpus apparaît sans commune mesure avec l'existant, et ce d'autant plus qu'on peut l'étendre en continu, au fur et à mesure que de nouveaux enregistrements sont générés par le service Botstats. A titre de comparaison, le plus important corpus auquel nous ayons eu accès est le corpus d'italien tchaté constitué par la société Eulogos (Eulogos, 2001), qui comporte 849 510 mots.

3.2 Evaluation qualitative

Les thèmes abordés par les canaux sont variés, et vont du tchat généraliste où l'on discute de tout et de rien, au tchat spécialisé dans les problèmes de programmation, ou encore les débats concernant l'actualité. On relève aussi des différences d'ordre pragmatique. En plus des traditionnels bavardages, certains canaux sont plus ou moins dédiés aux jeux (pendu, quizzes), alors que dans d'autres la conversation est alimentée par des dépêches AFP. D'autres enfin sont consacrés à des discutions techniques, sous forme de question/réponse, par exemple sur un canal consacré aux questions de programmation.

Certains canaux semblent à première vue assez originaux sur le plan linguistique. Ainsi on peut constater en comparant les figures 3 et 4 que le canal #edelweiss comporte peu de formes pour sa taille (42% de moins que #ffparadise, de taille pourtant plus réduite). D'un point de vue plus général, il semble qu'un corpus de tchat comporte nettement plus de formes qu'un

corpus écrit « standard » ou oral équivalent, quand on compare notre corpus avec ceux décrits dans (Gendner V. & Adda-Decker M., 2002).

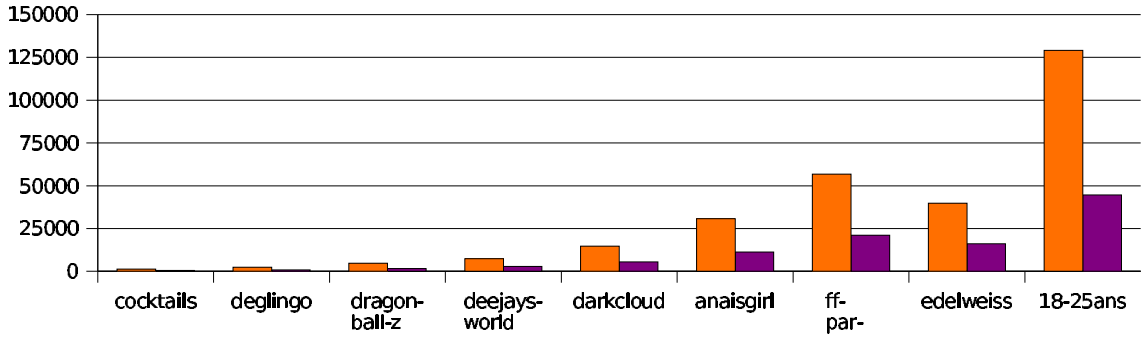


Figure 3 : nombre de formes (barre de gauche), et nombre de formes présentes au moins deux fois (barre de droite), dans quelques canaux du corpus.

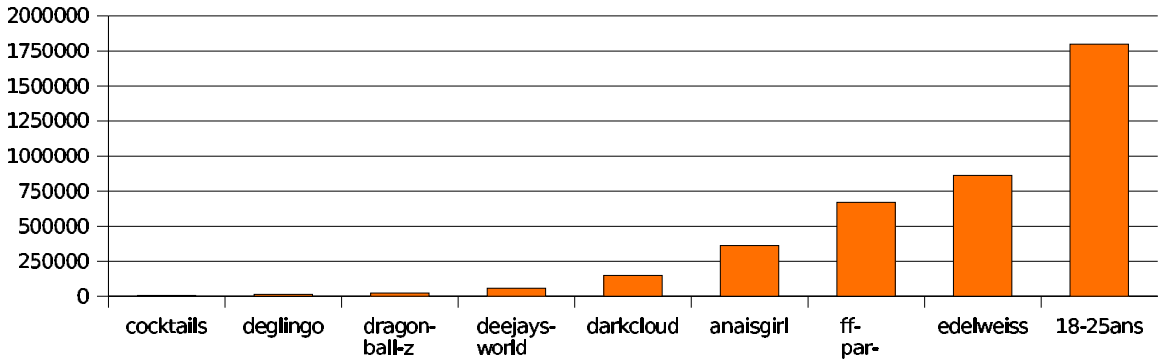


Figure 4 : nombre de mots dans quelques canaux du corpus.

Toujours par comparaison avec les corpus écrit et oral de (Gendner V. & Adda-Decker M., 2002), on peut constater que le nombre de formes présentes plusieurs fois, par rapport au nombre total de formes, est beaucoup plus faible en tchat (35% des formes pour le canal #18-25ans) que pour l'écrit « standard » (62%) et l'oral (70%).

3.3 L'avenir

Dans un premier temps, il est nécessaire d'effectuer un classement des canaux, en fonction de leur thème mais aussi de leurs spécificités pragmatiques (type d'interaction et d'interacteur[10]), afin de pouvoir ensuite sélectionner ceux qui correspondent le mieux à ce que l'on veut étudier. On ne s'attend pas, en effet, à ce qu'un canal de jeu ait les mêmes propriétés qu'un canal de conversation classique. L'encodage doit aussi être amélioré, comme décrit en 2.4, de façon à correspondre aux spécifications; il s'agit d'un travail semi-manuel relativement important, en particulier en ce qui concerne la notation du caractère humain ou non de l'auteur du message (attribut *type*).

Enfin, une annotation lexicale plus fine est envisageable, permettant d'identifier, et par conséquent de quantifier, les différentes particularités graphiques de la langue du tchat, comme les émoticons, déformations graphiques, corrections orthographiques, etc. fréquemment relevés dans l'étude de la langue tchatée, de façon plus fine que ce qui est présenté en première partie. Toutefois ce dernier traitement, manuel, est très lourd à mettre en œuvre, et ne peut pas concerner tout le corpus. Il est aussi possible d'obtenir des résultats

10 Humain ou robot.

intéressants pour caractériser un canal ou un utilisateur, à partir de quelques centaines de mots prélevés aléatoirement, comme nous en avons donné un bref aperçu en 1.2.

4 Mise à disposition du corpus

Notre corpus est consultable en ligne[11], grâce à une interface web. Afin de permettre sa consultation dans des conditions raisonnables, celui-ci a été transféré dans une base de données MySQL; un script PHP regénère à la volée le code XML correspondant à la partie du corpus sélectionnée dans l'interface. Ce code XML est associé à une feuille de style XSL permettant une visualisation simple des données au sein de l'interface, à condition d'utiliser un navigateur supportant ce format[12]. L'intérêt de l'utilisation dynamique d'une feuille XSL est que le code XML reste disponible, par exemple lorsque l'on demande au navigateur d'afficher le code source de la page.

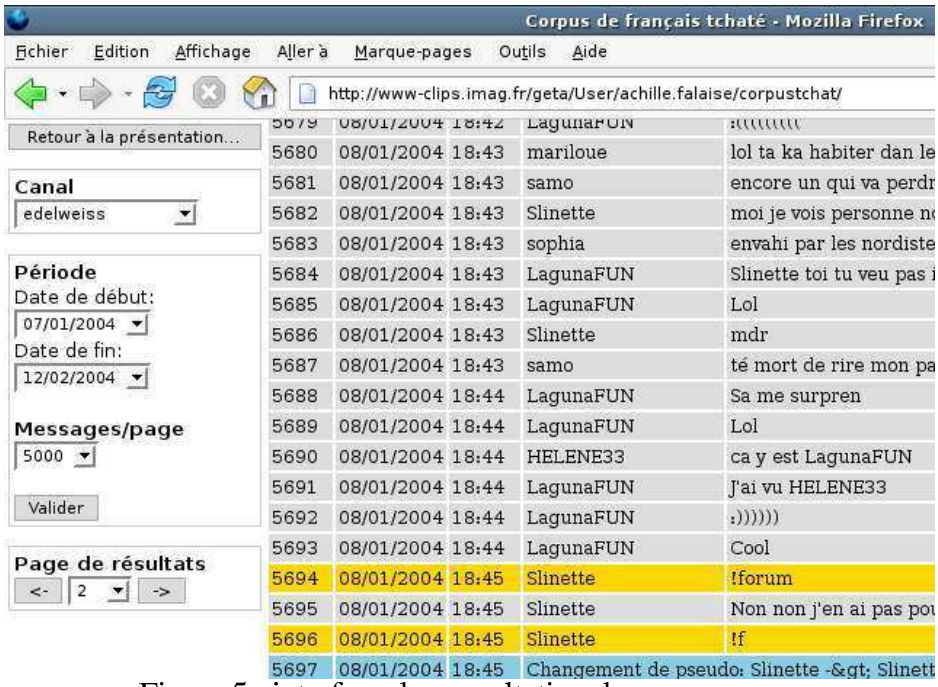


Figure 5 : interface de consultation du corpus.

A terme, un système d'enregistrement et d'authentification des utilisateurs de ce système sera mis en place, afin de restreindre son utilisation au seul monde scientifique.

Conclusion

Ce corpus, de par son étendue, tant du point de vue du nombre de mots (plus de 23 millions), que du nombre de canaux (105), est assez représentatif de la langue tchatée. Il peut ainsi, malgré certaines insuffisances, être utilisé en complément de corpus plus précis mais aussi plus restreints, dans le cadre de l'étude de la langue tchatée, ou encore pour évaluer

---

11 http://www-clips.imag.fr/geta/User/achille.falaise/corpustchat/

12 Ce format est supporté par Firefox 1.0, ainsi que par Internet Explorer 4.5+ sous Windows, après l'installation de la librairie MSXML pour les versions de Windows antérieures à Windows XP.

rapidement un outil de TALN dans ce cadre linguistique, ce pourquoi il était conçu à l'origine. C'est pourquoi nous pensons utile de le mettre à disposition de la communauté du TALN.

Progressivement, les nouveaux outils de communication écrite nous amènent à élargir notre conception de la langue écrite, et à reconsidérer certains principes du TALN qui semblaient acquis (grammaticalité des énoncés, lexique sous forme de listes, etc.). Nous pouvons constater, avec l'exemple de la langue tchatée, à quel point ces conceptions sont liées au caractère contraint de « l'écrit standard », et demandent à être élargies pour vraiment rendre compte des réalités cognitives à l'œuvre dans la langue.

Références

Eulogos (2001), « Corpus di conversazioni da chat-line in lingua italiana, da registrazioni effettuate nel primo trimestre 1998 »
http://www.intratext.com/X/ITA0192.HTM

Falaise (2004) : Premier pas vers une TA interactive pour le tchat, rapport de stage de master, Université Joseph Fourier, Grenoble, 63 pages.

Gedner V. & Adda-Decker M. (2002), Analyse comparative de corpus oraux et écrits français: mots, lemmes et classes morpho-syntaxiques, *Actes des XIVe Journées d'Etude sur la Parole*, Nancy.

Guimier de Neef E. & Véronis J. (2004) : « 1 pw1 sr la kestion ;-). », Journée d'étude de l'ATALA, *Le traitement automatique des nouvelles formes de communication écrite (e-mails, forums, chats, SMS, etc.)*, Paris.

http://www.up.univ-mrs.fr/~veronis/je-nfce/resumes.html#1pw1

Herring S. (1999), « Interactional coherence in CMC », *Journal of computer-Mediated Communication*, Vol. 4, n°4.

Latzko-Toth G. (2001), « Un dispositif construit par ses utilisateurs ? Le rôle structurant des pratiques de communication dans l'évolution technique de l'Internet Relay Chat », *Actes du IIIème colloque international sur les usages et services des télécommunications*, pp. 556-564.
http://grm.uqam.ca/textes/Latzko_ICUST2001.pdf

Pierozak I. (2003), Le "français tchaté" : un objet à géométrie variable ?, *Langage et Société*, n° 104, pp. 123-144.

Pujade L. (2001), L'écrit sur internet, mémoire de maîtrise, Toulouse, 179 pages.

Shortis, T. (2000), *The Language of ICT*, London, Routledge.

Étude de quelques problèmes de phonétisation dans un système de synthèse de la parole à partir de SMS

Rémi Bove[1]

Équipe DELIC – Université de Provence
29, Av. Robert Schuman, 13621 Aix-en-Provence Cedex 1
remi.bove@voila.fr

Mots-clefs – Keywords

SMS, phonétisation, synthèse de la parole.

SMS, phonetisation, speech synthesis.

Résumé – Abstract

Cet article présente une étude dont l'objectif était d'améliorer la phonétisation d'un système de synthèse vocale de SMS en ce qui concerne trois types de problèmes : l'écriture rébus (chiffres et lettres utilisés pour leur valeur phonique), les abréviations sous forme de squelettes consonantiques et les agglutinations (déterminants ou pronoms collés graphiquement au mot qui suit). Notre approche se base sur l'analyse d'un corpus de SMS, à partir duquel nous avons extrait des listes de formes permettant de compléter les lexiques du système, et mis au point de nouvelles règles pour les grammaires internes. Les modifications effectuées apportent une amélioration substantielle du système, bien qu'il reste, évidemment, de nombreuses autres classes de problèmes à traiter.

This article presents a study whose goal is to improve the grapheme-to-phoneme component of an SMS-to-speech system. The three types of problems tackled in the study are: rebus writing (digits and letters used for their phonetic value), consonant skeleton abbreviations and agglutinations (determiner or pronouns merged with the next word). Our approach is based on the analysis of an SMS corpus, from which we extracted lists of forms to enhance the system's lexicons, and developed new grammatical rules for the internal grammars. Our modifications result in a substantial improvement of the system, although, of course, there remain many other categories of problems to address.

---

[1] Cette étude a été réalisée dans le cadre d'un contrat de recherche avec France Télécom Division R&D.

1 Introduction

La synthèse automatique de la parole à partir de texte permet la conversion d'un document écrit en son équivalent parlé. Depuis 1965 et l'apparition sur le marché commercial d'un système de synthèse vocale développé par IBM, les programmes informatiques visant à synthétiser de la parole n'ont pas cessé de s'améliorer et de se multiplier (D'Alessandro, 2001). Les dispositifs actuels permettent de produire un signal acoustique de qualité jugée suffisante pour des applications nombreuses dans des domaines tels que l'aide aux personnes handicapées, le monitoring vocal, la communication homme/machine ou encore les services de télécommunication. Cependant, les progrès potentiels restent vastes tant en termes d'amélioration du naturel de la voix qu'en termes d'amélioration de la restitution du contenu sémantique. Un effort particulier doit être réalisé sur l'analyse linguistique pour parvenir à la vocalisation automatique d'un plus grand éventail de textes et en particulier celles des textes dits « mal formés » — il conviendrait sans doute de dire plutôt « non-standards » — comme ceux issus des SMS.

Cette communication rend compte de notre collaboration avec l'équipe du laboratoire Langues Naturelles (LN) au sein de France Télécom Recherche et Développement (FT R&D, pôle basé à Lannion, Côte d'Armor) sur le projet « *SMS2Voice* ». Celui-ci propose un dispositif de vocalisation de SMS développé par France Télécom R&D sous maîtrise d'ouvrage partagée Orange et FDF (Fixe et Distribution France), basé sur le couplage entre un analyseur syntaxique et un module de synthèse vocale. Le but de notre intervention était de tenter d'améliorer au mieux la couverture actuelle du système, en étudiant le fonctionnement du dispositif et en ciblant les traitements à effectuer. Les résultats de ce travail sont exposés dans ce papier.

2 SMS et synthèse de la parole

Le SMS (acronyme de *Short Message Service*) est un service d'échange de messages écrits (limités à 160 caractères), proposé par tous les opérateurs de téléphonie mobile. La vocalisation automatique de SMS a de nombreuses applications. Elle intéresse tout d'abord les différents opérateurs de téléphonie mobile pour la réception de SMS sur poste fixe. La synthèse vocale à partir de SMS pourrait également rendre ceux-ci accessibles aux aveugles et malvoyants. Enfin, la technologie aurait un intérêt non négligeable pour tous les métiers où les mains sont occupées (par exemple, les chauffeurs routiers) et pour lesquels il serait plus pratique de pouvoir écouter ses messages plutôt que de les lire.

Lorsqu'on s'intéresse à ce type de messages, on se rend rapidement compte que ceux-ci présentent de nombreuses particularités linguistiques problématiques pour la vocalisation (Anis, 2001, 2002 ; Guimier de Neef & Véronis, 2004). Il suffit pour s'en convaincre d'écouter la sortie produite par un synthétiseur sur des messages tels que :

`dsl ma bel! ms c dernié tps javé pa tro le tps`
(*Désolé ma belle ! Mais ces derniers temps j'avais pas trop le temps.*)

`slt jsp ke tt va bil`
(*Salut, j'espère que tout va bien.*)

Si les procédés d'écriture employés dans les SMS sont divers et nombreux (nous renvoyons le lecteur à Anis, 2002, pour une étude analytique du domaine), ils ne sont pas tous problématiques pour un traitement automatique. Il est possible de dresser des listes exhaustives pour les sigles par exemple (SMS, RER, …). En revanche, d'autres phénomènes sont productifs et ne peuvent être réglés par une simple liste. C'est notamment le cas des trois phénomènes spécifiques de l'écriture SMS que nous exposons dans cette communication : l'écriture *rébus* (chiffres et lettres), l'écriture par *squelettes consonantiques*, et le procédé d'*agglutination*. Nous pensions en effet, après un relevé systématique des erreurs de vocalisation produites par SMS2Voice, qu'ils étaient à l'origine de nombreuses erreurs de traitement par l'analyseur linguistique (que nous décrivons plus bas). Ces procédés sont présentés ci-dessous.

2.1 L'écriture rébus

Nous entendons par « rébus » le procédé d'écriture par lequel certaines séquences de lettres sont remplacées par un arrangement de chiffres et/ou de lettres correspondant au même phonème que la séquence en question. Exemples[2] :

`2m1` = *demain* [rébus chiffre]

`Kfé` = *café* [rébus lettre]

2.2 Squelettes consonantiques

Nous considérons comme squelettes consonantiques les mots dont les voyelles ont été supprimées, réduisant ainsi la forme à une succession des consonnes principales du mot. Comme le souligne justement (Anis, 2002), nous savons depuis longtemps grâce à la théorie de l'information, que les consonnes ont une valeur informative plus forte que les voyelles. Le mot français écrit est fortement charpenté autour des consonnes, dont certaines n'ont pas de contrepartie phonique. Exemples :

`slt` = *salut*

`prtt` = *pourtant* (notons que le « *n* » du son « *an* » n'est pas conservé alors que le « *t* » muet final l'est)

`tjs` = *toujours* (de la même manière ici le « *r* » pourtant prononcé n'est pas conservé alors que le « *s* » muet l'est)

2.3 Les agglutinations

Nous appelons enfin « agglutination » la formation d'un mot par la réunion de deux ou plusieurs unités lexicales (`jpouré` = « je pourrai »). Il est à noter que nous nous avons traité uniquement le cas des agglutinations binaires (combinaison de deux unités). Le tableau 1 donne un échantillon d'exemples pour les séquences d'agglutinations avec clitiques.

---

[2] Les exemples sont extraits du corpus sur lequel nous avons travaillé et que nous présentons plus loin.

Forme agglutinée	Exemples de patrons observés	Exemples de formes
« **JE** »	j + pronom j + verbe	`jte (= je te)` `jsui (= je suis)`
« **QUE** » (sous la forme « KE »)	*k + article défini* *k + pronom démonstratif* *k + pronom personnel*	`kle (= que le)` `kce (= que ce)` `ktu (= que tu)`

Tableau 1 : Exemples d'agglutinations

3 Environnement technique

Rappelons que cette étude porte sur l'amélioration du dispositif *SMS2Voice*. La figure 1 représente globalement la structure de cette application. Notre travail a porté sur la brique logicielle T*i*LT (TLiT → TiLT, **T**raitement **LI**nguistique de **T**extes ; cf. Guimier de Neef *et al.* 2002), déjà en utilisation chez France Télécom pour de nombreux projets (indexation et filtrage linguistique, résumé automatique, classement thématique, etc.), et qui constitue une véritable « boîte à outils » pour le Traitement Automatique des Langues. Un module permet de segmenter un texte d'entrée en mots ; ceux-ci sont ensuite soumis à une analyse lexicale, morphologique, et éventuellement des corrections. Ces diverses informations sont ensuite traduites afin d'affecter une catégorie grammaticale à chaque type de mot, et pour pouvoir procéder à une analyse syntaxique et sémantique des données ainsi traitées. Pour qu'il puisse fonctionner il lui faut donc (entre autre) :

- Des **lexiques** (avec informations morpho-flexionnelles pour l'association de chaque mot à ces différentes analyses hors contexte)
- Des **stratégies de corrections** (pour la correction de formes erronées)
- Des **grammaires** (pour la désambiguïsation lexicale par exploration du contexte)

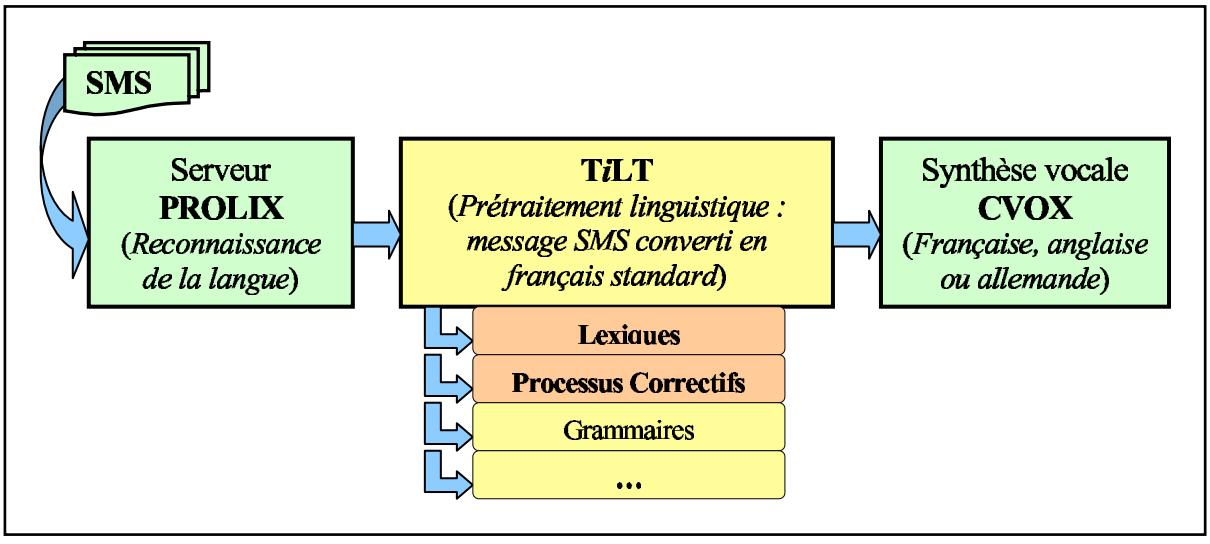


Figure 1 : Architecture générale du système *SMS2Voice*

Ce système est initialement prévu pour analyser des textes au format « standard » et nous avons participé à son adaptation aux traitements de textes non-standards. Le dispositif actuel était opérationnel sur de nombreux points, mais son fonctionnement pouvait être amélioré pour certains phénomènes étant encore partiellement traités, tels que :

`gcrai` (j'essaierai) vocalisé tel quel.

`attd` (attend) vocalisé tel quel.

`Deboulot` (de boulot) traduit « *déboule* ».

A la suite des observations, nous nous sommes rendus compte que les systèmes de correction et les lexiques préalablement en place présentaient certaines limites. C'est donc essentiellement sur ces deux modules que nous avons travaillé. Il semblait donc important de poursuivre cette étude avec un travail sur corpus afin d'approfondir la connaissance des phénomènes SMS choisis, et améliorer au mieux le système.

4 Méthodologie et corpus

Le premier prérequis important à cette étude était de disposer d'un corpus à partir duquel il était possible de faire un certain nombre d'observations et de traitements. Le corpus de SMS utilisé dans cette étude a été réalisé par des étudiants de l'Université de Provence de 2000 à 2004 dans le cadre de travaux pratiques et de mémoires. Celui-ci contient 13 400 messages représentant près de 156 620 mots. Ce volume de données reste évidemment modeste, mais cela est principalement dû à la difficulté de recueil des messages du fait de leur caractère intime et personnel pour de nombreux usagers.

A partir de ce corpus, nous avons commencé par dégager des listes d'occurrences représentant les mots les plus fréquents selon divers critères d'identification, afin d'étudier notamment si les formes les plus courantes étaient les plus problématiques ou non. Pour ce faire, nous avons comparé les listes de mots extraites du corpus avec un lexique de français standard de référence, le lexique MULTEXT (Ide & Véronis, 1994). Le but de notre approche était de voir pour chaque phénomène donné quelles sont les formes que le système est déjà en mesure de traiter, et d'isoler ainsi les occurrences problématiques pour leur appliquer le traitement adéquat en fonction de leurs particularités (longueur, complexité, etc.). Précisons que notre analyse a porté sur les lexiques (et donc sur des listes de mots hors contexte), ainsi que sur les processus correctifs. Nous avons donc travaillé sur la production des corrections possibles et non sur le choix de celles-ci (qui relève plutôt du module grammatical), afin d'assurer au mieux la conversion des forme SMS en français standard.

4.1 Extraction des occurrences

La première étape pour étudier chacun des procédés nécessite d'extraire le plus précisément possible du corpus la liste des occurrences correspondant aux critères d'identification du phénomène, par l'intermédiaire d'expressions régulières. Pour ce faire, nous avons mis en place des traitements qui permettent, à partir de la liste des formes inconnues en français standard, d'extraire le plus précisément possible les formes voulues. L'opération suivante consiste à éliminer manuellement les formes qui ont été extraites par le script sur une base

purement formelle, mais ne correspondent pas au phénomène recherché. En effet, la majorité des phénomènes est difficile à filtrer et à extraire directement, et dans un premier temps nous ne pouvons nous baser que sur des indices pour les identifier.

4.2 Étude quantitative

A la suite de l'étape d'extraction des occurrences, nous avons donc été contraint de procéder à une phase de tri manuel des formes extraites pour ne garder que celles correspondant au patron voulu. Le tableau 2 donne, pour chacun des procédés étudiés, les têtes de listes des occurrences (ainsi que leur fréquence d'apparition) lors de la première extraction, puis la liste des occurrences restantes à traiter après filtrage manuel.

TÊTES DE LISTE POUR LES REBUS CHIFFRES			
1ère extraction de formes		**Formes restantes après filtrage manuel**	
Fréquence	*Occurrence*	*Fréquence*	*Occurrence*
63	2m1	63	2m1
62	bi1	62	bi1
35	2min	35	2min
25	2main	25	2main
17	2pui	17	2pui
13	b1	13	b1
13	1er	12	vi1
12	vi1	10	qq1
12	2m	7	dem1
10	qq1	6	pl1

TÊTES DE LISTE POUR LES REBUS LETTRES			
1ère extraction de formes		**Formes restantes après filtrage manuel**	
Fréquence	*Occurrence*	*Fréquence*	*Occurrence*
29	reCu	9	paC
26	MonNuméro	4	pE
9	paC	3	mR
4	pE	2	vE
4	gpVeuilz	2	trouV
4	faCon	2	jaV
3	mR	2	jV
2	vE	2	danC
2	trouV	2	creV
2	reCus	2	cT

TÊTES DE LISTE POUR LES SQUELETTES CONSONANTIQUES			
1ère extraction de formes		Formes restantes après filtrage manuel	
Fréquence	*Occurrence*	*Fréquence*	*Occurrence*
209	sms	167	slt
167	slt	145	svp
145	svp	120	stp
120	stp	52	qqch
52	qqch	48	lgtps
48	lgtps	39	rdv
39	rdv	37	bcp
37	bcp	35	jrs
35	jrs	33	msg
33	msg	29	tps

TÊTES DE LISTE POUR LES AGGLUTINATIONS			
1ère extraction de formes		Formes restantes après filtrage manuel	
Fréquence	*Occurrence*	*Fréquence*	*Occurrence*
209	sms	55	jsui
167	slt	42	jte
145	svp	23	jt
144	tro	20	jp
132	ds	14	jme
120	stp	12	jm
97	st	11	ktu
64	ms	10	1pe
63	2m1	9	jvai
60	ns	9	jtapel

Tableau 2 : Têtes de liste des occurrences pour les différents procédés

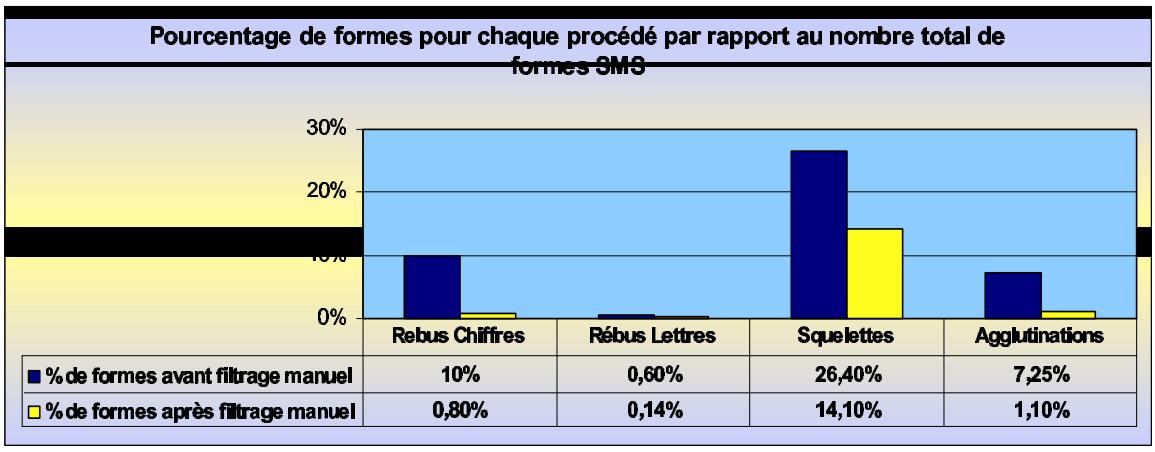


Figure 2 : Filtrage manuel

La figure 2 donne le récapitulatif statistique de nos observations pour chacun des phénomènes étudiés, une fois le tri manuel effectué. Les pourcentages sont faibles mais restent en

proportion suffisante pour repérer un nombre important de dysfonctionnements du système. On note qu'une grande partie des formes correspondant aux patrons d'extraction ne correspond pas aux procédés recherchés, ce qui justifie une analyse linguistique manuelle. Un système automatique basé sur de tels patrons aurait des performances très faibles.

5 Etude qualitative et amélioration du système

Dès que les formes correspondant au procédé sont identifiées, il est nécessaire de vérifier si elles sont connues et correctement traitées par le système, afin d'améliorer les ressources de celui-ci. Pour cela nous avons recherché (à l'aide de scripts à base d'expressions régulières) les listes d'occurrences extraites et non traitées par le système. Lorsque ces formes inconnues sont isolées, une stratégie de traitement doit être adoptée. Deux possibilités sont offertes (la décision dépend généralement de la productivité du phénomène traité) :

- Figer l'encodage des formes SMS dans un lexique ;
- Modifier les règles des processus correctifs

Par exemple pour le cas des squelettes consonantiques, l'observation des listes extraites a permis de constater que bien que les occurrences soient fréquentes, elles présentent cependant une faible diversité. Les formes inconnues du système (`slt, bcp,` etc.) ont donc simplement été encodées dans le fichier des formes figées. En revanche, des règles ont été créées pour les rébus et agglutinations.

Concernant le procédé d'écriture rébus (chiffres et/ou lettres), nous avons ajouté de nouvelles règles phonétiques. Par exemple, une règle de type `1 → jî` donne la possibilité pour le chiffre « `1` » de correspondre à la séquence « `ien` » (exemple : « `rev1` » = reviens ; « `ch1` » = chien ; etc.). Ce type de modification permet donc de générer un nouveau phonétiseur et d'améliorer ainsi la sortie produite lors de l'accès au lexique (Tableau 3).

Avant modification du phonétiseur	Après modification du phonétiseur
> v1 Mot analysé : v1 Nombre de solutions : 10 v1 vain vainc vaincs vains vin vingt vins vint vînt	> v1 Mot analysé : v1 Nombre de solutions : 12 v1 vain vainc vaincs vains **viens** **vient** vin vingt vins vint vînt

Tableau 3 : Exemple d'accès au lexique

Pour le cas des agglutinations, on remarque que ce phénomène ne concerne que certains patrons syntaxiques (clitiques+verbe, déterminant+adj., etc.). Nous avons donc modifié et/ou ajouté des règles de grammaire adaptées.

Exemples :

```
((LETTRE_ELIDEE ADV_NEG_NE)(VERB))
```
*(pour le cas d'agglutination entre « **ne** » et un **verbe**, ex. : `nviendrai`)*

```
((LETTRE_ELIDEE PREP)(VERB INF))
```
*(pour le cas d'agglutination entre **préposition** et **verbe à l'infinitif**, ex. : `dpartir`)*

Nous avons ensuite effectué différents tests de «non-régression » afin de garantir la pertinence des modifications opérées. Ces tests consistent à appliquer T*i*LT sur une série de fichiers spécifiques (ex : messages avec systématiquement le chiffre `2`, la lettre `d` ou encore la forme interrogative « `eske` » [est-ce que]). Les fichiers de sortie permettent notamment de voir la différence entre les résultats de l'ancien test et ceux de celui venant d'être effectué. Cette phase est particulièrement importante car elle permet de s'assurer qu'il n'y a pas eu de problème majeur avant de passer à d'autres modifications sur les données, et elle permet de juger de la pertinence des améliorations apportées.

6 Conclusions / perspectives

Le travail réalisé a permis d'améliorer de façon significative la performance du système *SMS2Voice*. Une évaluation quantitative détaillée pourrait être réalisée, mais sa mise en œuvre aurait demandé un effort important qui dépassait le cadre qui nous était imparti. De plus, une telle évaluation serait plus pertinente à entreprendre lorsque d'autres phénomènes auront pu être traités (tels que ceux listés par Anis, 2003). Par une analyse qualitative systématique, nous avons en effet pu mettre en évidence de nombreux points qui méritent eux aussi amélioration. Au-delà des interventions que nous avons décrites, nous avons pu proposer différentes grilles d'analyse et des typologies (morpho-lexicale, morpho-syntaxique, etc.) qui permettront de faciliter la suite de ce travail.

Quelques points restent à améliorer pour les procédés que nous avons traités. Concernant le phénomène d'agglutination notamment, qui est soumis à d'importantes contraintes linguistiques, diverses règles grammaticales peuvent être encore ajoutées (par exemple, les observations menées sur les agglutinations binaires méritent d'être étendues au cas des agglutinations ternaires (ex : `jtsouhaite` [je te souhaite])). Il reste également un nombre important de formes et de procédés que nous n'avons pas abordés et dont le traitement s'annonce extrêmement délicat. Par exemple, une forme telle que « `1dpdte` » (indépendante) fait appel à trois procédés cumulés :

1. Rébus avec chiffre : « `1` » (in-)
2. Rébus avec lettre : « `d` » (-dé-)
3. Squelette consonantique : « `pdte` » (-pendante)

C'est également le cas pour les squelettes consonantiques. Si la réduction aux consonnes concerne la totalité du mot, le phénomène est plutôt limité (`slt`, `bjr`, `vrmt`, …). Il devient au contraire très productif lorsqu'il n'affecte que partiellement le mot. Par exemple, pour la forme « `sincèrement` » : `s1cèrmnt`, `s1cRmnt`, `sincRmnt`, `s1cRment`, etc.

Ce type de combinaisons pose des problèmes de vocalisation difficilement résolubles dans l'état actuel des connaissances. En effet, pour traiter de telles occurrences le système T*i*LT

devrait avoir simultanément recours à diverses stratégies de correction. La difficulté vient de l'architecture modulaire du système (qui est d'ailleurs commune à la plupart des systèmes de TAL actuels) : chaque module de correction intégré à l'analyseur linguistique intervient de façon indépendante, séparément des autres modules. Il n'y a pas de traitement global pour l'application des corrections, et l'application systématique de toutes les combinaisons de modules produit une explosion combinatoire extrêmement coûteuse et néfaste *in fine* à la précision du système.

Il demeure donc de nombreux aspects intéressants pour des études futures. L'écriture SMS n'est ni normée, ni stable ; une vocalisation de qualité nécessite donc une explicitation de tout le message textuel quels que soient les procédés d'écriture employés. Il est donc certainement nécessaire et important de poursuivre l'approfondissement des travaux de linguistique et de traitement automatique menés sur les données issues de SMS, ainsi que d'autres formes d'écrit non standards qui utilisent des procédés analogues (chats et e-mails, cf. Torzec, 2001).

Remerciements

Je tiens à remercier Emilie Guimier de Neef qui a encadré ce stage au sein de France Télécom, ainsi qu'à toute l'équipe du laboratoire Langues Naturelles qui m'a accueilli. J'adresse également tous mes remerciements à Jean Véronis qui a dirigé ce travail. Je leur suis reconnaissant pour leurs nombreux commentaires sur cet article.

Références

Anis, J. (2001). (Dir.), (2001), « *Parlez-vous texto ?*», Paris : Le Cherche Midi Éditeur.

Anis, J. (2002), Communication électronique scripturale et formes langagières : chats et SMS. *Actes des Quatrièmes Rencontres Réseaux Humains / Réseaux Technologiques*, Université de Poitiers. [http://oav.univ-poitiers.fr/rhrt/2002/actes%202002/jacques%20anis.htm]

D'Alessandro. C. (2001). 33 ans de synthèse de la parole à partir de texte : une promenade sonore (1968-2001), *Traitement automatique de la parole,* Vol. 42(1), pp. 297-321.

Guimier De Neef E., Boualem M., Chardenon C., Filoche P., Vinesse J. (2002), Natural Language processing software tools and linguistic data developped by France telecom R&D, *CDAC Conference*, India.

Guimier de Neef, E. & Véronis , J. (2004). 1 pw1 sr la kestion ;-). *Journée d'Étude de l'ATALA "Le traitement automatique des nouvelles formes de communication écrite (e-mails, forums, chats, SMS, etc.)*", Paris.

Ide, N., & Véronis, J. (1994). MULTEXT (Multilingual Tools and Corpora), *14th International Conference on Computational Linguistics, COLING'94*. Kyoto. 588-592.

Torzec, N., Moundenc, T., Emerard, F. (2001). Prétraitement et analyse linguistique dans le système de synthèse TTS CVOX : Application à la vocalisation automatique d'e-mails, *Traitement automatique de la parole*. Vol. 42(1), pp. 17-46.

RECITAL 2005

9ème Rencontre des Étudiants Chercheurs
en Informatique pour le
Traitement Automatique des Langues Naturelles

POSTER

Synchronisation syntaxesémantique, des grammaires minimalistes catégorielles (GMC) aux Constraint Languages for Lambda Structures (CLLS)

Amblard Maxime (1)
(1) LaBRI -Université de Bordeaux 1
351 cours de la libération
33405 Talence cedex amblard@labri.fr
date de soutenance prévue : novembre 2006

Motselefs – Keywords

logique, grammaires minimalistes catégorielles, λealcul, portée des quantificateurs, Constraint Language for Lambda Structures
logic, minimalist grammars, λealculus, quantifiers scope, Constraint Language for Lambda Structures

Résumé -Abstract

Ces travaux se basent sur l'approche computationelle et logique de Ed Stabler (**?**), qui donne une formalisation sous forme de grammaire du programme minimaliste de Noam Chomsky (**?**). La question que je veux aborder est comment, à partir d'une analyse syntaxique retrouver la forme prédicative de l'énoncé. Pour cela, il faut mettre en place une interface entre syntaxe et sémantique. C'est ce que je propose en utilisant les Grammaires Minimalistes Catégorielles (GMC) extension des GM vers le calcul de Lambeck. Ce nouveau formalisme permet une synchronisation simple avec le λealcul.

Parmi les questions fréquemment rencontrées dans le traitement des langues naturelles, j'interroge la performance de cette interface pour la résolution des problèmes de portée des quantificateurs. Je montre pourquoi et comment il faut utiliser un λealcul plus élaboré pour obtenir les différentes lectures, en utilisant Constraint Languages for Lambda Structures €LLS.

This work is based on the computational and logical approach of Ed Stabler (**?**), which gives a formalization of the minimalist program of Noam Chomsky (**?**). The question I want to solve is, starting from a syntactic analysis, how to find the predicative forms of a sentence.

I propose an interface between syntax and semantic by using Categorical Minimalists Grammars (CMG) extension of the MG towards the Lambeck calculus and Constraint Language for Lambda Structures (CLLS). This interface is powerful for the resolution of quantifier scope ambiguities.

Introduction

Dans un premier temps, le formalisme des GMC, extension des grammaires minimalistes vers le calcul de Lambeck, (**?**) et (**?**), sera présenté. Puis une interface entre syntaxe et sémantique -calcul des formes prédicatives en logique d'ordre supérieur -sera exposée. Mais cette interface ne suffit pas pour conserver tous les calculs possibles avec les grammaires de Lambeck. Dans une seconde partie, j'étudie une solution qui consiste à utiliser des λtermes structurés avec contexte, via CLLS (**?**).

1 Grammaires Minimalistes Catégorielles -GMC

1.1 Syntaxe

Ce formalisme, présenté dans (**?**), est l'extension de la formalisation des grammaires minimalistes de Stabler (**?**). Tout comme ces dernières, les GMC sont lexicalisées et les expressions sont des arbres finis, ordonnées avec projection. La projection est la relation entre les éléments permettant de retrouver la tête du constituant.

Les fonctions génératrices dans les grammaires minimalistes sont de deux types. La première est la fusion qui permet d'agglomérer deux structures. La seconde est motivée par Chomsky comme la nécessité de vérifier certains traits, mettant en relation deux éléments de l'analyse, permettant de faire monter dans la dérivation une feuille en position basse : le déplacement.

Fusion : élimination de / ou \ -structurellement la concaténation droite ou gauche. Ce qui se traduit sous forme de séquents par les formules suivantes :

$$\frac{\Gamma \to x : A/B \quad \Delta \to y : B}{\Gamma, \Delta \to xy : A}\,[e/] \qquad \frac{\Delta \to y : B \quad \Gamma \to x : B\backslash A}{\Gamma, \Delta \to xy : A}\,[e\backslash]$$

Graphiquement, on ajoute une branche binaire à notre arbre entre une nouvelle feuille et la structure actuelle. Le nouveau sommet contiendra la réduction des types.

Déplacement : cette opération se base sur la notion de lien entre deux positions d'une dérivation. Elle se déroule en deux temps. Dans la dérivation principale, on introduit certaines hypothèses. Dans une seconde dérivation, on construit un élément de type $\times$. Si le type des éléments autours du $\times$ est le même que celui des hypothèses dans la dérivation principale, on peut substituer dans cette dernière les éléments de la seconde.

$$\frac{\Gamma \to w : A \times B \quad \Delta, x : A, y : B, \Delta' \to z : C}{\Delta, \Gamma, \Delta' \to let(x, y) = (\pi_1(w), \pi_2(w)) \; in \; z : C}\,[e\times]$$

Où π_1 est la projection de la première composante et π_2 de la deuxième.

Selon l'hypothèse de Chomsky, les formes phonologiques et logiques fonctionnent séparément. Le déplacement met en relation deux positions dans l'analyse et ces formes peuvent se substituer soit en position basse, soit en position haute ce qui motive cette opération.

Graphiquement, on ajoute une branche unaire à l'arbre pour marquer que la substitution a eu lieu. Pour une meilleure visibilité, on marquera les projections reliant les deux dérivations.

La section suivante présente un calcul sémantique parallèle à l'analyse syntaxique, dont un exemple est exposé.

1.2 Interface syntaxe -sémantique

La sémantique utilisée est la forme prédicative des constituants formant l'énoncé. Pour l'obtenir, le λ-calcul est la voie la plus naturelle. Cependant un λ-calcul classique ne permet pas d'aboutir au résultat escompté car les hypothèses utilisées pour le déplacement excluent une construction itérative.

L'analyse syntaxique est supposée normalisée pour rencontrer l'objet puis le sujet, ce qui donne l'ordre des variables dans le λ-terme du verbe.

Les règles de l'interface λ-termes contextués. A chaque règle syntaxique correspond une règle sémantique : la fusion est une application dans la contrepartie sémantique. On aura donc $[\backslash E]$ ou $[/E]$ devenant $[\rightarrow E]$.

L'équivalence du déplacement est un peu plus complexe. Il faut distinguer deux situations. On peut introduire soit une variable de type simple et dans ce cas, il faut faire une application standard, soit une variable de type supérieur, ayant subi un type-raising, donc inverser l'application. Ce qui se traduit par les deux règles suivantes :

$$\frac{\Delta \vdash z : T \rightarrow U \rightarrow V \quad \Gamma \cup [x : T] \vdash y : U}{\Delta \cup \Gamma \vdash z(\lambda x.y) : V}\ [RAISE/]$$

$$\frac{\Delta \vdash z : T \rightarrow U \rightarrow V \quad \Gamma \cup [x : T] \vdash y : U}{\Delta \cup \Gamma \vdash (\lambda x.y)(z) : V}\ [NORAISE/]$$

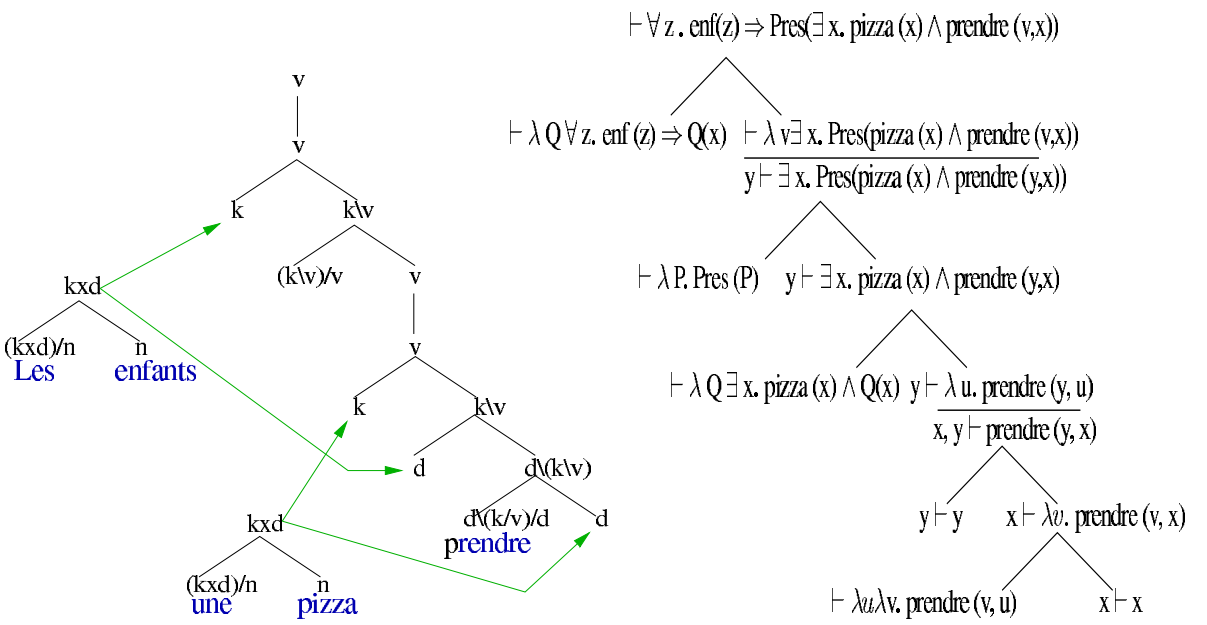
Pour les hypothèses, nous introduisons des variables neutres. Mais, avant qu'un déplacement n'intervienne, une λ-abstraction sur cette variable est nécessaire. Pour éviter de perdre la position relative à cette variable lors d'applications précédentes, on utilise un contexte. Les variables neutres ont une redondance dans le contexte et lors de leur extraction de ce dernier, on abstrait sur la bonne variable. Les contextes se conservent par application et sont marqués par le symbole $\vdash$.

Interface syntaxe-sémantique SYN est le calcul syntaxique ($\times$, / et $\backslash$) et SEM le calcul sémantique ($\rightarrow$, $[RAISE]$ et $[NORAISE]$).

Chaque étape dans un calcul trouve sa contrepartie dans l'autre. Deux preuves, l'une dans SYN et l'autre dans SEM sont dites synchronisées si : chaque feuille dans SEM est en bijection avec une feuille de SYN et chaque étape et sa contrepartie sont réalisées en même temps. Pour synchroniser ces deux calculs, on construit pour la sémantique un arbre binaire en utilisant le λ-terme associé dans la partie syntaxique.

Exemple d'analyse sémantique synchronisée à l'analyse syntaxique : *Les enfants prendront une pizza.*

items lexicaux	λ-terme sémantique	types syntaxique
$prendre$	$\vdash \lambda u \lambda v.\ prendre(v,u)$	$(d\backslash k\backslash v)/d$
les	$\vdash \lambda P \lambda Q \forall x.\ P(x) \rightarrow Q(x)$	$k \times d/n$
une	$\vdash \lambda P \lambda Q \exists x.P(x) \wedge Q(x)$	$k \times d/n$
$pizza$	$\vdash \lambda x.pizza(x)$	n
$enfants$	$\vdash \lambda x.enfant(x)$	n
$infl$	$\vdash \lambda P.Pres(P)$	$(k\backslash v)/v$



La première partie de l'analyse syntaxique sature les positions du verbe avec des hypothèses dans les deux calculs. On obtient dans SEM : $x, y \vdash prendre(y, x)$ où deux variables neutres ont saturé les positions dans le verbe et sont également présente dans le contexte.

Puis le déplacement relatif à l'objet est déclenché dans SYN, marqué par une branche unaire. Dans SEM, une λabstraction s'opère par extraction d'un élément du contexte. Le déplacement permet l'arrivée effective de la forme logique, ici : $\vdash \lambda Q \exists x.pizza(x) \wedge Q(x)$. Une application est donc possible et suit immédiatement cette arrivée dans la dérivation.

La dérivation se poursuit avec l'inflexion qui apporte dans la dérivation la nécessité d'une hypothèse de type cas. Celle-ci correspond à la vérification que le constituant introduit via sa position aura le bon cas, en l'occurrence le nominatif pour le sujet. Elle déclenche le second déplacement, avec projection des formes logiques et phonologiques. La contrepartie dans SEM est une λabstraction puis une application.

L'analyse syntaxique est terminée et acceptante. Dans le même temps la formule sémantique voulue a été construite : $\vdash \forall z.enf(z) \Rightarrow Pres(\exists x.pizza(x) \wedge prendre(v, x))$.

Cependant, à chaque analyse syntaxique ne correspond qu'une forme sémantique. Or, lorsqu'un énoncé contient des quantificateurs, plusieurs lectures sont possibles. La section suivante présente une extension de l'interface pour calculer les différentes formules.

2 Utilisation de λstructures

CLLS (Constraint Language for Lambda Structures) est un formalisme élaboré par Markus Egg et al en Allemagne. L'idée est de représenter les λtermes par des arbres binaires et de relier ces arbres entre eux par deux types de relations pour permettre la sousspécification.

Il existe donc trois types de liens possibles entre les divers éléments de la structure.

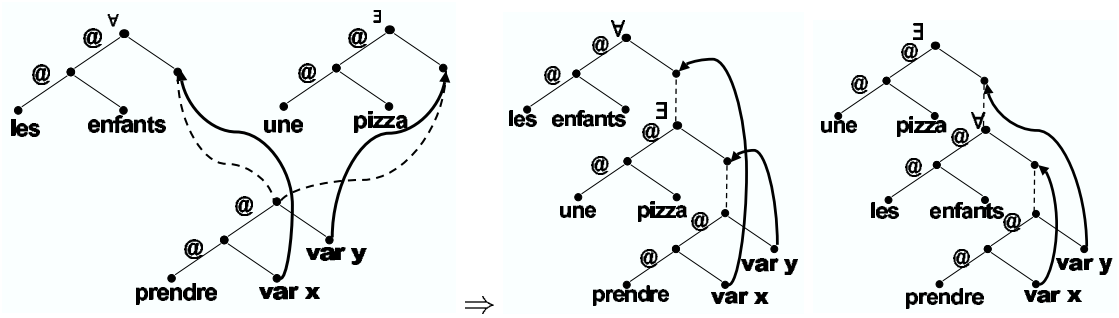
1. une arrête d'un arbre qui est une application d'un terme à un autre, notée @.

2. une relation de substitution qui lie une variable à une autre suite à une application entre deux arbres. Graphiquement ce lien sera une flèche en gras.

3. une relation de dominance entre deux arbres, reliant une feuille d'un arbre au sommet d'un autre, graphiquement celuiei sera en pointillés. Elle intervient lorsque deux arbres entrent en relation par application. La racine d'un arbre peut être sujet de plusieurs relations de dominance.

Grâce à ces définitions, la construction de représentations sousspécifiées entre les différents constituants est permise. En effet, les relations de type 2 permettent de conserver l'ordre dans lequel les variables sont substituées et la relation de dominance (type 3) permet de modifier la portée des quantificateurs.

Exemple de CLLS et de réductions : *les enfants prendront une pizza*



lecture
nombre enfants = nombre pizzas

lecture
une unique pizza

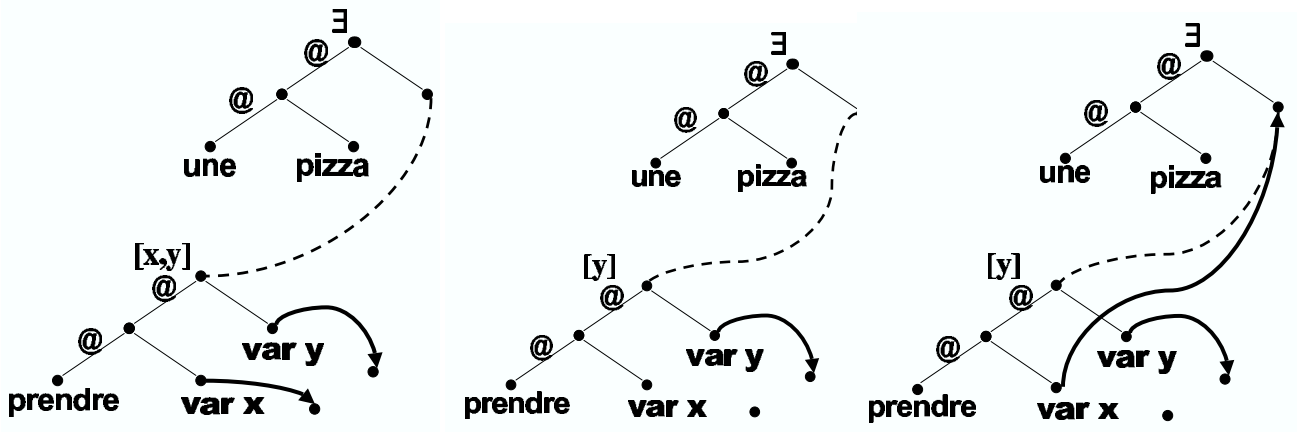
λstructures contextués

Pour synchroniser les GMC et CLLS, il faut conserver la notion de contexte dont on a vu la nécessité dans la partie précédente.

Pour cela, on ajoute sur la racine de la CLLS une liste représentant le contexte, notée entre crochet : [ ]. On admettra la propriété selon laquelle les contextes sont sédentaires. Ce qui signifie que lorsque la structure est liée par une application à une autre structure, on maintient le contexte sur la racine de la sousstructure.

Détails du calcul dans la partie sémantique pour le déplacement :

(1) (2) (3)



(1) verbe chargé de ses deux hypothèses avec redondance dans le contexte

(2) première phase du déplacement, lambda-abstraction, une variable est libérée en fonction du contexte. Ici, celle correspondant à l'objet est à nouveau libre.

(3) application avec l'arbre correspondant à l'objet. Les deux arbres sont maintenant liés et on peut réaliser une réduction de la relation de dominance. La variable esseulée disparaît.

Cette représentation est nécessaire car, par définition, l'apport logique d'un constituant dans les grammaires minimalistes se fait toujours en position haute.

Le défaut de ce modèle est qu'il est finalement trop souple. En effet, pour une phrase avec deux quantificateurs, toutes les lectures ($2! = 2$) sont possibles. Par contre pour une phrase à trois quantificateurs, il n'y a pas 6 ($3! = 6$) lectures mais seulement 5. Une des solutions à l'étude est d'ajouter un ordre sur les relations de dominance. Pour cela il faut différencier ces relations selon leur genre. Une classification sur ces dernières est en cours d'étude.

Conclusion

A partir d'une idée calculatoire simple qui sous-tend la théorie de Stabler et qui encode le programme minimaliste, on obtient un système formel pour l'analyse syntaxique, les GMCs. A partir de cette analyse, j'ai proposé une interface pour le calcul de la sémantique où chaque utilisation de règle dans la partie syntaxique trouve sa contrepartie dans celle sémantique, par l'utilisation de λ-termes structurés avec contexte. L'utilisation du λ-calcul est naturelle pour le calcul de la forme prédicative d'un énoncé, cependant, il ne suffit pas dans sa forme standard. CLLS avec contexte est une solution viable qui permet de reprojeter la structure obtenue vers plusieurs formules et d'obtenir les différentes lectures.

Le concept de synchronisation entre analyse syntaxique et sémantique montre qu'il est difficile de mettre en place un système formel répondant à toutes les caractéristiques de la langue, car il se heurte à de nombreuses questions. Celle de la portée des quantificateurs reste ouverte. Une implémentation à partir de celui des GM devient envisageable.

Références

Amblard M., Lecomte A. et Retoré C. (2004), Synchronization Syntax Semantic for a minimalism theory *Journée Sémantique et Modélisation 2004*,

Chomsky Noam (1995), *The Minimalist Program*, Cambridge, MIT Press.

Egg M., Koller A. et Niehren J. (2001) The Constraint Language for Lambda Structures, *Journal of Logic, Language, and Information*, To appear.

Heim I. et Kratzer A. (1998), *Semantics en Generative Grammar*, Oxford, Blackwell.

Koller A., Burchardt A. et Walter S. (2004), Computational semantics, Actes de *ESSLLI 2004*.

Lecomte A. et Retoré C. (2001), Extending Lambeck Grammars, Actes de *Algebraic Mahods in Lanuage Processing 2003*, 354-361

Stabler Ed. (1997), Derivational Minimalism, Actes de *Logical Aspect of Computational Linguistics 1996*, vol 1328, Springer-Verlag.

Stabler Ed. (1999), Remnant movement and structural complexity, Actes de *Constraints and Resources in Natural Language Syntax and Semantics 1999*, 299-326.

Recherche d'information en langue arabe : influence des paramètres linguistiques et de pondération de LSA

Siham Boulaknadel (1,2), Fadoua Ataa-Allah (2)

(1) LINA FRE CNRS 2729 – Université de Nantes
2 rue la Houssinière BP 92208 44322 Nantes cedex 03, France
siham.boulaknadel@univ-nantes.fr
(2) GSCM – Université Mohammed V
BP 1014 Agdal Rabat-Maroc
fadoua_01@yahoo.fr

Mots-clefs – Keywords

Recherche d'information, Analyse de la sémantique latente, Langue arabe, Racinisation

Information retrieval, Latent semantic analyses, Arabic language, Stemming

Résumé – Abstract

Nous nous intéressons à la recherche d'information en langue arabe en utilisant le modèle de l'analyse sémantique latente (LSA). Nous proposons dans cet article de montrer que le traitement linguistique et la pondération des unités lexicales influent sur la performance de la LSA pour quatre cas d'études : le premier avec un simple prétraitement des corpus; le deuxième en utilisant un anti-dictionnaire; le troisième avec un racineur de l'arabe ; le quatrième où nous avons combiné l'anti-dictionnaire et le racineur. Globalement les résultats de nos expérimentations montrent que les traitements linguistiques ainsi que la pondération des unités lexicales utilisés améliorent la performance de LSA.

We are interested in information retrieval in Arabic language by using latent semantic analysis method (LSA). We propose in this article to show that the linguistic treatment and weighting of lexemes influence the performance of LSA. Four cases are studied: the first with a simple pretreatment of the corpora; the second by using a stopword list; the third with arabic stemmer; the fourth where we combined stopword list and arabic stemmer. Broadly the results of our experiments show that the linguistic treatments as well as weighting of lexemes used improve the performance of LSA.

1 Introduction

En recherche d'information, le problème d'accès au texte est essentiellement dû à l'écart entre les termes utilisés dans les requêtes et les documents. L'appariement entre requête et document se fait donc par l'intermédiaire de leur représentation respective. Le modèle de recherche le plus souvent utilisé est le modèle vectoriel (Salton, 1983). Un des problèmes de ce modèle réside dans l'hypothèse d'indépendance faite sur les termes d'indexation : chaque terme d'indexation constitue une dimension de l'espace vectoriel, sans considération d'éventuelles relations entre termes.

En l'absence d'une connaissance approfondie de la collection de documents, la requête peut être formulé en des termes proches mais non identiques à ceux employés dans un document. Un certain nombre de chercheurs se sont intéressés à ce problème, soit par l'utilisation de réseaux sémantiques qui consiste à recourir à une base de connaissances linguistiques regroupant les mots sémantiquement proches et structuré selon des relations hyperonymiques et/ou synonymiques (Grefenstette, 1994), soit par l'extension de requêtes, opération par laquelle un certain nombre de termes issus de documents de la collection sont ajoutés à une requête.

La troisième possibilité que nous avons choisie, consiste à se servir des relations sémantiques implicites induites par les cooccurrences entre termes dans les documents. Ainsi le modèle de l'analyse sémantique latente (LSA) (Deerwester et al., 1990) consiste à réduire le nombre de dimensions de l'espace vectoriel en s'appuyant sur le fait que les documents traitant des mêmes sujets ont des vocabulaires proches et sont donc proches dans l'espace vectoriel.

Dans notre travail, nous avons sélectionné les schémas de pondération qui améliorent la performance de la méthode LSA pour le calcul de similarité, dans le cas de cinq corpus de petites tailles en langue arabe, tout en évaluant l'importance de différents paramètres linguistiques utilisés.

2 Présentation de LSA

L'analyse sémantique latente (LSA) consiste à réduire le nombre de dimensions de l'espace vectoriel par le biais d'une décomposition en valeurs singulières (SVD), de la matrice A en un produit de trois autres matrices : $A = U S V^T$

Où U est une matrice orthogonale de taille (m x n) de description d'unité lexicale, V est une matrice orthogonale de taille (n x n) de description d'unité textuelle et S une matrice diagonale de taille (n x n).

À partir d'un certain nombre k<n, nous nous apercevons de l'existence de valeurs singulières très faibles et qui peuvent être négligées dans la matrice.

De ce fait, il est démontré qu'il y a une meilleure approximation Ak de A qui est donnée par :

$$A_k = U_k S_k V_k^T$$

Cette réduction va permettre de ne garder que les unités lexicales les plus significatives. À noter que k est déterminé de façon empirique en fonction du corpus utilisé et du degré de performance voulu.

Pour évaluer la performance de la LSA on utilise les deux mesures traditionnelles de précision et de taux de rappel (Salton, 1989).

3 Paramètres de pondération

La pondération des unités lexicales consiste à transformer l'occurrence d'une unité lexicale dans l'unité textuelle par une combinaison de pondérations locales L(i,j), indiquant l'importance de l'unité lexicale i dans l'unité textuelle j et pondérations globales G(i), indiquant l'importance de l'unité lexicale i dans l'ensemble des unités textuelles de la collection.

Avec f_{ij} la fréquence de l'unité lexicale i dans l'unité textuelle j, df_i le nombre d'unités textuelles auxquelles l'unité lexicale i appartient, gf_i le nombre total de fois où l'unité lexicale i apparaît dans la collection, N est le nombre d'unités textuelles, M le nombre des termes dans le corpus et p_{ij} est le rapport de f_{ij} par gf_i.

Pondération globale

Nom du schéma	Formule	Intérêt
Entropie	$1-\sum_{j}^{N}\frac{p_{ij}\log p_{ij}}{\log(N)}$	Elle tient compte de la distribution des unités lexicales dans les unités textuelles et permet d'attribuer un poids minimum aux termes qui sont distribués de la même façon dans toutes les unités textuelles et un poids maximum aux termes qui sont concentrés dans quelques unités textuelles
Normal	$\frac{1}{\sum_{j}^{N} f_{ij}^{2}}$	Elle a pour effet de donner un poids élevé aux termes peu fréquents et elle ne dépend que de la somme des fréquences au carré et pas de la distribution de ces fréquences.
Gfldf	$\frac{gf_i}{df_i}$	Elles pondèrent tous deux les termes par le nombre des unités textuelles différentes dans lesquelles ils apparaissent. La différence entre les deux c'est que GfIdf augmente le poids des mots fréquents.
Idf	$\log_2\left(\frac{N}{df_i}\right)$	

Figure 1 : Paramètres de pondération utilisés

4 Traitements linguistiques

L'arabe est une langue sémitique s'écrivant de droite à gauche elle comporte 28 consonnes et 6 voyelles standard (3 longues :ا و ي et 3 courtes : ُ / ِ / َ). Le traitement automatique de l'arabe est difficile vu ses variations orthographiques et sa structure morphologique complexe.

Deux approches sont utilisées dans l'analyse morphologique de l'arabe, la première que nous avons choisie (Darwish, 2002) est une analyse morphologique assouplie ou racinisation qui consiste à essayer de déceler si des suffixes ou préfixes ont été ajoutés à l'unité lexicale : par exemple pour le duel (ان) dans (معلمان, deux professeurs), le pluriel des noms masculins (ون , ين) dans (معلمون, des professeurs) et féminins (ات) dans (مسلمات, musulmanes) ; la forme possessive (نا, كم, هم) dans (كتابهم, ses livres) et les préfixes dans les articles définis (ال, وال, بال, كال, فال).

La deuxième est une lemmatisation qui consiste à réduire les formes déclinées à une représentation canonique.

5 Expérimentations

Notre objectif est de sélectionner les schémas de pondération qui améliorent la performance de la méthode LSA pour le calcul de similarité, dans le cas des corpus de petite taille, tout en évaluant l'importance de l'utilisation d'un anti-dictionnaire et d'un racineur.

5.1 Données

Afin de bien évaluer nos résultats sur les corpus de petite taille, nous avons choisi sur Internet une version arabe des contes partiellement voyellés de 1800 mots : « Le paysan énergique »[1], de H.Darwish « Fleurs du miel »[2] , «Musique de la nature »[3] et « Sous les branches »[4] de K.Abid et «Les chaussures en bois»[5] de J.alhamad. Nous avons appliqué la transcription de Buckwalter qui consiste à transcrire l'alphabet arabe en alphabet latin (Buckwalter, 2002). Nous avons décidé de construire un anti-dictionnaire général qui contient l'ensemble des unités lexicales grammaticales extraites du dictionnaire arabe[6] ensuite nous avons choisi de formuler les requêtes avec aussi peu de variations que possible par rapport à la formulation d'origine. L'ensemble des requêtes que nous avons établi est de l'ordre de 71 requêtes.

5.2 Calculs des courbes

Nous avons segmenté par la suite chacun de ces contes en paragraphes, ce qui nous a permis de construire cinq corpus dont le nombre d'unités textuelles (paragraphes) varie entre 8 et 24. Après avoir transformer ces contes et requêtes textuelles en mode vectoriel, nous avons calculé la précision moyenne sur l'ensemble des requêtes de chaque corpus, en faisant varier k de 2 à r (le rang de la matrice correspondant à chaque corpus). Les paragraphes retournés étaient ceux dont le vecteur faisait un angle de cosinus supérieur à un seuil de 0.9 avec les vecteurs requêtes.

5.3 Résultats

Nous avons effectué des tests pour vingt trois schémas de pondération, plus un autre test où nous avons utilisé la méthode LSA sans appliquer aucune pondération à la matrice originale. Effectivement, d'après les tests appliqués sur le corpus « Musique de la nature », nous avons remarqué que la performance de la LSA sans pondéré la matrice originale est relativement

---

[1] http://www.awu-dam.org/book/99/child99/5-a-d/book99-ch008.htm
[2] http://www.awu-dam.org/book/99/child99/11-h-a/book99-ch001.htm
[3] http://www.awu-dam.org/book/99/child99/11-h-a/book99-ch003.htm
[4] http://www.awu-dam.org/book/99/child99/11-h-a/book99-ch012.htm
[5] http://www.comp.leeds.ac.uk/latifa/research.htm
[6] http://www.almeshkat.net/books/archive/books/muajm arabia.zip

inférieure de 6% par rapport à celle où nous appliquons le schéma de pondération 'Pondération Local Logarithmique * Entropie de Dumais' ; et les schémas 'Pondération Local Logarithmique*Entropie Globale' et 'Pondération Local Logarithmique*GFIDF'.

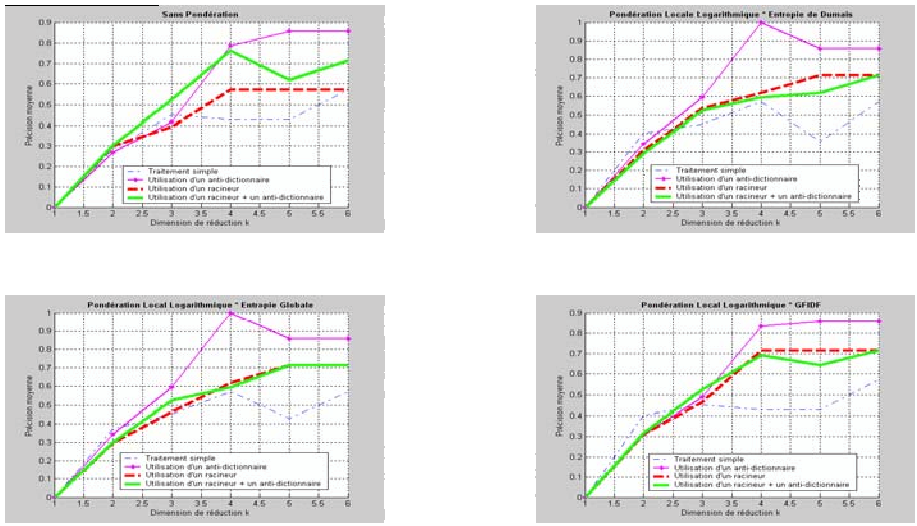
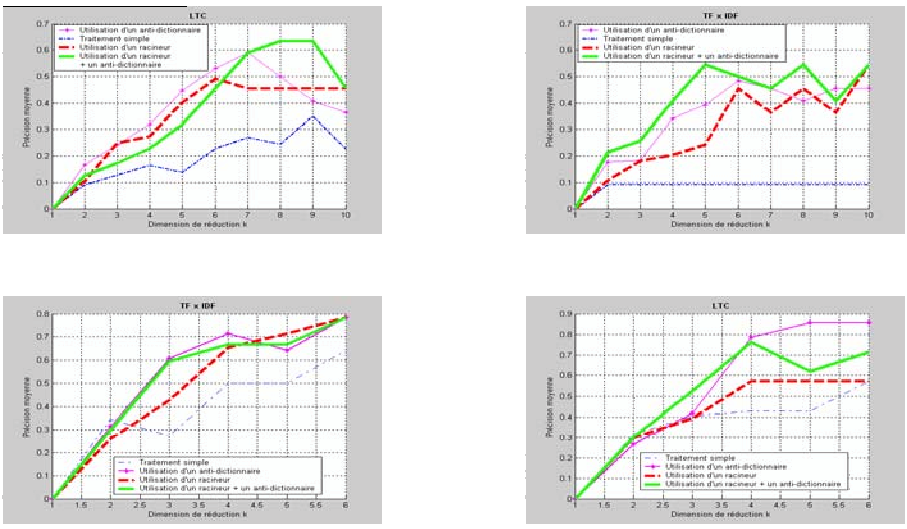


Figure 2 : L'influence des schémas de pondération sur la performance de la LSA

Pour évaluer l'importance de l'utilisation d'un anti-dictionnaire et d'un racineur, nous avons extrait la précision maximale de l'ensemble des précisions moyennes résultantes des tests réalisés. Nous avons présenté l'évolution de la précision moyenne de la méthode LSA, en appliquant les deux schémas de pondération « Tf x IDF » et « LTC », pour le corpus « Sous les branches » sur la figure 3-(a) et sur la figure 3-(b) pour le corpus « Musique de la nature ».



(a)

(b)

Figure 3 : Evolution de la précision en fonction du nombre de dimensions de l'espace pour un seuil de 0.9

Globalement les courbes calculées montrent que la performance de LSA s'améliore en utilisant soit un anti-dictionnaire soit un racineur, soit les deux. Néanmoins pour certaines requêtes les traitements linguistiques n'améliorent pas la performance de la LSA. Ceci est dû au racineur qui échoue à traiter certains pluriels et verbes. Par exemple pour les pluriels irréguliers des noms comme « طفل, enfant » « أطفال, enfants » qui ne sont pas une combinaison des formes singulières. Pour les verbes irréguliers comportant des consonnes particulières dites faibles (و, ا, ي) qui sont soit conservée, soit remplacée ou éliminée lors de leur déclinaison, exemple «قال , il a dit » «يقول , il dit » . Vu aussi la petite taille de nos corpus, par conséquent on peut dire que la LSA reste sensible dans le cas où on a peu de données à traiter.

6 Conclusion

Nous avons proposé une approche pour améliorer la méthode de l'analyse sémantique latente (LSA) en intégrant les paramètres linguistiques et de pondération. L'évaluation a montré l'intérêt d'appliquer conjointement le traitement linguistique et la pondération des unités lexicales pour pouvoir améliorer la performance de la LSA. Dans la suite de nos travaux, nous envisageons d'étendre cette étude à l'utilisation de la lemmatisation.

Références

Buckwalter T.(2002), Buckwalter Arabic Morphological Analyzer Version 1.0, http://www.ldc.upenn.edu/Catalog/CatologEntry.jsp?catologId=LDC2002L49.

Darwish K. (2002), Building a Shallow Arabic Morphological Analyzer in One Day, Proceedings of the workshop on Computational Approaches to Semitic Languages in the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02) , pp. 47-54.

Deerwester S, Dumais S.T., Furnas G.W., Landauer T.K., Hrashman R. (1990), Indexing by latent semantic analysis, Journal of the american society for information science, Vol.41, pp. 391-407.

Grefenstette G. (1994), Explorations in automatic thesaurus discovery, New York, Kluwer Academic Publishers.

Salton G. (1989), Automatic text processing the transformation analysis and retrieval of information by computer, New York, Addison-Wesley.

Salton G. (1983), An Introduction to Modern Information Retrieval, New York, McGraw-Hill.

Etiquetage morpho-syntaxique des textes arabes par modèle de Markov caché

Abdelhamid EL JIHAD (1), Abdellah YOUSFI (2)
(1),(2) Institut d'études et de recherches pour l'arabisation
Université Mohamed V, Rabat, Maroc
(1) eljihad@ifrance.com
date de soutenance prévue : 2007
(2) yousfi240ma@yahoo.fr
date de soutenance : 19 juin 2001

Mots-clefs – Keywords

Corpus, jeu d'étiquettes, Etiquetage morpho-syntaxique, texte arabe, modèle de Markov caché
Corpus, the set of tags, the morpho-syntactic tagging, arabic text, Hidden Markov Model

Résumé - Abstract

L'étiquetage des textes est un outil très important pour le traitement automatique de langage, il est utilisé dans plusieurs applications par exemple l'analyse morphologique et syntaxique des textes, l'indexation, la recherche documentaire, la voyellation pour la langue arabe, les modèles de langage probabilistes (modèles n-classes), etc.
Dans cet article nous avons élaboré un système d'étiquetage morpho-syntaxique de la langue arabe en utilisant les modèles de Markov cachés, et ceci pour construire un corpus de référence étiqueté et représentant les principales difficultés grammaticales rencontrées en langue arabe générale.
Pour l'estimation des paramètres de ce modèle, nous avons utilisé un corpus d'apprentissage étiqueté manuellement en utilisant un jeu de 52 étiquettes de nature morpho-syntaxique. Ensuite on procède à une amélioration du système grâce à la procédure de réestimation des paramètres de ce modèle.

The tagging of texts is a very important tool for various applications of natural language processing : morphological and syntactic analysis of texts, indexation and information retrieval, vowelling of arabic texts, probabilistic language model (n-class model).
In this paper we have used the Hidden Markov Model (HMM) to tag the arabic texts. This system of tagging is used to build a large labelled arabic corpus. The experiments are carried in the set of the labelled texts and the 52 tags of morpho-syntactic nature, in order to estimate the parameters of the HMM.

1 Introduction

Le développement des corpus électroniques a bénéficié ces dernières années d'un appui vigoureux et un soutien financier important, de la communauté du traitement automatique des langues naturelles, qui voit là une étape indispensable pour la mise au point de systèmes de TAL robustes. Aujourd'hui de vaste corpus de textes électroniques étiquetés sont disponibles et sont majoritairement de langue anglaise. Ceci a permis l'essor considérable des traitements automatiques concernant cette langue; des outils d'interrogation de ces corpus ainsi que des outils d'annotations proprement dits (étiqueteurs, analyseurs syntaxique, etc.) se répandent. Leurs équivalents en français commence à apparaître également [Habert et al 1997].

Pour la langue arabe, il n'existe pas à ce jour de corpus étiqueté aisément disponible. Par conséquent les recherches linguistiques qui ont recours à des corpus étiquetés sont donc encore rares. Motivé par ce manque, l'Institut d'Etudes et de Recherches pour l'Arabisation (IERA) a entrepris un projet de recherche dont l'objectif est la constitution d'un corpus de référence étiqueté et représentant les principales difficultés grammaticales rencontrées en langue arabe générale. La disponibilité de ce corpus à l'institut, va donner le coup d'envoi aux divers travaux de recherches linguistiques qui utilisent les corpus étiquetés. Un corpus étiqueté est un corpus dans lequel on associe à des segments de textes (le plus souvent des mots) d'autres informations de quelque nature qu'elle soit morphologique, syntaxique, sémantique, prosodique, critique, etc [Veronis 2000][Vergne et al 1998].

En particulier, dans la communauté du traitement automatique des langues naturelles, quand on parle de corpus étiqueté on fait référence le plus souvent à un document où chaque mot possède une étiquette morpho-syntaxique et une seule.
L'étiquetage morpho-syntaxique automatique est un processus qui s'effectue généralement en trois étapes [Minh et al 2003][Rajman et al 2000]: la segmentation du texte en unités lexicales, l'étiquetage à priori, la désambiguïsation qui permet d'attribuer, pour chacun des unités lexicales et en fonction de son contexte, l'étiquette morpho-syntaxique pertinente.
La taille du jeu d'étiquettes, la taille du corpus d'apprentissage sont autant de facteur importants pour une bonne performance du système d'étiquetage [Chanod 1995][Claud 1995].
En général, il existe deux méthodes pour l'étiquetage morpho-syntaxique :
- Méthode à base de règles [Claud 1995][Bril 1992].
- Méthode probabiliste.
Dans cet article nous avons utilisé la deuxième approche.

2 Méthode probabiliste

Le choix de l'étiquette la plus probable en un point donné se fait au regard de l'historique des dernières étiquettes qui viennent d'être attribuées. En général cet historique se limite à une ou deux étiquettes précédentes. Cette méthode suppose qu'on dispose d'un corpus d'apprentissage qui doit être d'une taille suffisante pour permettre une estimation fiable des probabilités [Habert et al 1997].
Soit $Ph = w_1...w_T$ une phrase constituée des mots $w_1, ..., w_T$, $E = \{et_1, ..., et_N\}$ un jeu d'étiquettes.
L'étiquetage morpho-syntaxique de la phrase Ph par des étiquettes appartenant à E et s'appuyant

sur l'approche probabiliste , consiste à trouver l'ensemble d'étiquettes $et^*{}_1...et^*{}_T$ associés à la phrase Ph tel que :

$$et^*{}_1...et^*{}_T = arg \max_{et_1...et_T} Pr(w_1...w_T, et_1...et_T) \quad (1)$$

Pour faciliter la résolution de ce problème on utilise les modèles de Markov cachés d'ordre 1.

3 Etiquetage morpho-syntaxique par modèle de Markov caché d'ordre 1

Un modèle de Markov caché d'ordre 1 est un double processus $(X_t, Y_t)_{t\geq 1}$ avec :
• X_t est une chaîne de Markov d'ordre 1 à valeur dans un ensemble d'états fini $Q = \{q_1, ..., q_N\}$, X_t vérifie :
$Pr(X_{t+1} = q_j / X_1 = q_1, ..., X_t = q_i) = Pr(X_{t+1} = q_j / X_t = q_i) = a_{ij}$.
$Pr(X_1 = q_i) = \pi_i, i = 1, ..., N$.
a_{ij} est la probabilité de transition entre les états q_i et q_j.
π_i est la probabilité que l'états q_i est un état initial.
• Y_t est un processus observable à valeurs dans un ensemble mesurable Y, Y_t vérifie :
$Pr(Y_t = y_t / X_1 = q_1, ..., X_t = q_i, Y_1 = y_1, ..., Y_{t-1} = y_{t-1}) = Pr(Y_t = y_t / X_t = q_i) = b_i(y_t) = b_{it}$.
b_{it} est la probabilité d'émission de l'observation y_t à partir de l'état q_i.
Dans la suite on supposera que le double processus :
$X_t = et_{i_t}$ représentant les étiquettes appartenant à l'ensemble E,
$Y_t = w_t$ représentant les mots de notre vocabulaire $V = \{w_1, ..., w_L\}$,
est un modèle de Markov caché d'ordre 1.
Remarque :
Ce modèle est défini entièrement par un vecteur de paramètres noté $\lambda = (\Pi, A, B)$.
• $\Pi = \{\pi_1, ..., \pi_N\}$ l'ensemble des probabilités initiales.
• $A = (a_{ij})_{1\leq i,j\leq N}$ la matrice des probabilités de transition entre les étiquettes.
• $B = (b_{it})_{1\leq i\leq N}$ et $1 \leq t \leq L$: la matrice des probabilités d'émission des mots à partir des étiquettes.

4 Procédure d'apprentissage (Estimation des paramètres)

L'apprentissage est une opération nécessaire pour un système de reconnaissance de formes (en particulier le système d'étiquetage), il permet d'estimer les paramètres du modèle $\lambda = (\Pi, A, B)$. Un apprentissage incorrect ou insuffisant diminue la performance du système d'étiquetage. Pour préparer le corpus d'apprentissage, on procède par approximations successives. Un premier corpus d'apprentissage, relativement court, permet d'étiqueter un corpus beaucoup plus important. Celui-ci est corrigé, ce qui permet de réestimer les probabilités, il sert donc à un second apprentissage, et ainsi de suite.
En général il existe trois méthodes d'estimation de ces paramètres[1] :
• L'estimation par maximum de vraisemblance (Maximum Likelihood Estimation), elle est réalisée par l'algorithme de Baum-Welch [Baum 1972] ou l'algorithme de Viterbi [Celeux 92].

[1] Pour plus de détaille sur ces formule voir [Yousfi 2001]

• L'estimation par maximum a posteriori [John Arice].
• L'estimation par maximum d'information mutuel [Bahl et al 86,87][Kapadia 93].
Dans notre cas nous avons utilisé l'estimation par maximum de vraisemblance car c'est la plus utilisée et la plus facile à calculer.
Alors si on prend un ensemble d'apprentissage $R = \{Ph_1, ..., Ph_K\}$, constitué des phrases $Ph_1, ..., Ph_K$ étiquetées manuellement, les formules d'estimation des paramètres du modèle $\lambda = (\Pi, A, B)$ sont données par :

$$a_{ij} = \frac{\sum_{n=1}^{K} \text{le nombre de fois où la transition } et_i\ et_j \text{ est dans la phrase } Ph_n}{\sum_{n=1}^{K} \text{le nombre de fois où l'état } et_i \text{ est atteint le long de la phrase } Ph_n}$$

$$\pi_i = \frac{\sum_{n=1}^{K} \delta[\text{l'étiquette } et_i \text{ est un état initial dans la phrase } Ph_n]}{K}$$

$$b_{it} = \frac{\sum_{n=1}^{K} \text{le nombre de fois où le mot } w_t \text{ à l'étiquette } et_i \text{ le long de la phrase } Ph_n}{\sum_{n=1}^{K} \text{le nombre de fois où l'état } et_i \text{ est atteint le long de la phrase } Ph_n}$$

avec :

$$\delta[x] = \begin{cases} 1 & \text{si l'événement x est vrai} \\ 0 & \text{sinon} \end{cases}$$

5 Etiquetage automatique par algorithme de Viterbi

Pour un calcul plus rapide du chemin optimal[2] dans la formule (1) nous avons utilisé l'algorithme de Viterbi [For 73].
On note par :

$$\delta_t(et_j) = \max_{et_{i_1} ... et_{i_t}} Pr(w_1 ... w_t, et_{i_1} ... et_{i_t})$$

avec $et_{i_t} = et_j$.
Cette formule devient [Yousfi 2001]:

$$\delta_t(et_j) = \max_{et_i} \delta_{t-1}(et_i).a_{ij}.b_j(w_t)$$

On calcule cette formule pour toutes les valeurs $t = 1, ..., T$ et $j = 1, ..., N$.
Enfin le chemin optimal est obtenu à l'aide d'un calcul récursif sur cette formule.

6 Expérimentation

6.1 Données d'apprentissage

Le travail expérimental a été réalisé en trois grandes étapes :
1) étape de définition du jeu d'étiquettes et de construction de corpus d'apprentissage.
La définition de notre propre jeu d'étiquettes morpho-syntaxique a été particulièrement délicate, cette phase a été réalisée en collaboration avec des linguistes pour satisfaire au besoin des projets en cours de réalisation à IERA. Ce jeu d'étiquettes est constitué de 52 étiquettes de nature

[2]Nous cherchons ce chemin dans un réseau d'étiquettes. Ce réseau est construit de tel façon à ce que pour une phrase donnée, chaque chemin de ce réseau correspond à la probabilité que cette phrase à les étiquettes de ce chemin ($Pr(w_1 ... w_t, et_{i_1} ... et_{i_t})$). Le chemin associé à la probabilité maximale est nommé chemin optimal.

morpho-syntaxique (comme par exemple ism-faail, ism-mafaoul, harf nasb,...).
Le corpus d'apprentissage est constitué d'un ensemble de phrases représentant les principales règles morphologiques et syntaxiques utilisées en langue arabe générale. Ce corpus a été étiqueté manuellement par un linguiste.
2) étape d'estimation des paramètres du modèle de Markov caché.
3) étape d'étiquetage automatique et réestimation des paramètres du modèle de Markov caché.
Pour réaliser ces deux dernières étapes, nous avons développé une application en langage C, comportant deux modules, module d'apprentissage et module d'étiquetage automatique qui permet d'étiqueter automatiquement un corpus brut, ce dernier est corrigé manuellement pour servir à une réestimation des paramètres du modèle de Markov caché.
Les programmes sont évalués sur deux versions de textes voyellé et non voyellé.

6.2 Résultats

Le taux d'erreur est mesuré sur deux ensembles :
Ensemble1 constitué des mêmes phrases que l'ensemble d'apprentissage mais sans étiquettes,
Ensemble2 constitué de phrases (sans étiquettes) différentes de celles de l'ensemble d'apprentissage.

	Ensemble1	Ensemble2
Textes voyellés	1,76%	2%
Textes non voyellés	2,5%	3%

Table 1: Les taux d'erreur d'étiquetage automatique.

On remarque que dans le cas des textes non voyellés le taux d'erreur augmente par rapport aux textes voyellés, à cause de l'augmentation de l'ambiguïté (un mot peut prendre plusieurs étiquettes). Pour le reste des erreurs, elles sont dues au manque de données d'apprentissage (il existe des mots et des transitions entre des étiquettes qui ne sont pas représentées dans le corpus d'apprentissage).

7 Conclusions et perspectives

En analysant les résultats trouvés, nous avons remarqué que la majorité d'erreurs d'étiquetage provient essentiellement du problème de manque ou d'insuffisance de données d'apprentissage.
Dans notre cas il existe deux type de problèmes de manque de données :
• un ou plusieurs mots, appartenant à la phrase à étiqueter par ce système, n'existent pas dans le lexique, c'est à dire nous n'avons pas une estimation des probabilités d'observation de ces mots dans tous les états.
• une ou plusieurs étiquettes n'ont pas de prédécesseurs dans la phrase à étiqueter automatiquement, c'est à dire nous n'avons pas une estimation des probabilités de transition de ces étiquettes vers tous les autres étiquettes du système.
Dans la suite de notre travail, nous allons proceder á deux solutions pour remedier à ces deux problèmes :

la première est d'introduire une sorte d'analyse morphologique qui s'appuit sur les formes morphologiques des mots pour pouvoir identifier les étiquettes des mots inconnus.
La deuxième est d'introduire une base de règles syntaxiques qui définie les transitions possibles entre les différents étiquettes.

Références

L.R. Bahl, P.F. Brown, P.V. de Souza & R.L. Mercer : "*Maximum mutual information estimation in hidden Markov model parameters for speech recognition* ", Proc. ICASSP, pp. 49-52, Tokyo, 1986.

L. R. Bahl, P. F. Brown, P.V De Souza and R. L. Mercer : "*Estimating HMM parameters so as to maximise speech recognition accuracy* ", Research Report RC-13121, IBM TJ Watson Research Center, 9/10/1987.

L. Baum : "*An inequality and association maximization technique in statistical estimation for probabilistic functions of Markov processes* ", Inequality, vol. 3, 1972.

G. Celux, J. Clairambault :"*Estimation de chaines de Markov cachées: méthodes et problèmes* ", Journées thématiques CNRS sur les approches markoviennes en signal et images, Septembre 1992.

Jean-Pierre Chanod and Pasi Tapanainen : "*Tagging French - comparing a statistical and a constraint-based method*", Proceeding of the seventh Conference of the European Chapter of the Association for Computatinal

Linguistics (EACL.95), Dublin, Ireland. pp.149-156, 1995.

Claude De Loupy : "*La méthode détiquetage d'Eric Brill*". Revue T.A.L, 1995, Vol.36, nř 1-2, pp.37-46

Eric Brill : "*A simple rule-based part of speech tagger*". Proceedings of the third Conference on Applied Natural Language Processing, Trento, Italy. pp.152-155. Avril 1992.

Fornay D. R. : "*The Viterbi Algorithm* ", Proc. IEEE, vol. 61, n 3, mai 1973.

Benoît Habert, Adeline Nazarenko, André Salem : "*Les linguistiques de corpus*" , Armand colin / Masson.Paris, 1997.

John Rice : "*Mathematical Statistics and data*

analysis ", page 511-540.

S. Kapadia, V. Valtchev & S.J. Young : "*MMI training for continuous phoneme recognition on the TIMIT database* ", Proc. ICASSP, pp. II.491-494, Minneapolis, 1993.

Thi Minh Huyen Nguyen, Laurent Romary, Xuan Luong Vu : "*Une étude de cas pour l'étiquetage morpho-syntaxique de textes vietnamiens*" , 5e conférence sur le traitement Automatique du Langage Naturel (TALN2003), Batz-sur-Mer, 11-14 juin, 2003.

Patrick Paroubek et Martin Rajman : "*Etiquetage morpho-syntaxique.*" , Ingénierie des langues. pp.131-150, Paris, HERMES Sciences Europe.

Jacques Vergne, Emmanuel Giguet: "*Regards théoriques sur le "Tagging"* " , 5e conférence sur le traitement Automatique du Langage Naturel (TALN98), Paris, France, 10-12 juin, 1998.

Jean Veronis : "*Annotation automatique de corpus : panorama et état de la technique*" , Ingénierie des langues. pp.111-128. Paris, HERMES Sciences Europe.

A. Yousfi : "*Introduction de la Vitesse d'élocution dans un modèle de reconnaissance automatique de la parole* " , Thèse de doctorat, 19 juin 2001.

Identification des composants temporels pour la représentation des dépêches épidémiologiques

Manal EL Zant[1], Liliane Pellegrin[1], Hervé Chaudet[1,2], Michel Roux[1]

[1]Laboratoire d'Informatique Fondamentale, UMR CNRS 6166
Équipe BIM, Faculté de médecine, 27 Bd Jean Moulin, 13005 Marseille
{el.zant, liliane.pellegrin, michel.roux}@medecine.univ-mrs.fr
[2]Unité de Recherche Epidémiologique, Département de Santé Publique,
IMTSSA, 13998 Marseille Armées
lhcp@acm.org
Date de la thèse (fin 2006)

Mots-clefs – Keywords

Analyse de textes, structure des évènements, extraction d'information, sémantique du temps

Text analysis, event structure, extraction information, temporal semantics

Résumé – Abstract

Dans le cadre du projet EpidémIA qui vise à la construction d'un système d'aide à la décision pour assister l'utilisateur dans son activité de gestion des risques sanitaires, un travail préalable sur la compositionalité des évènements (STEEL) nous a permis d'orienter notre travail dans le domaine de la localisation d'information spatio-temporelle. Nous avons construit des graphes de transducteurs pour identifier les informations temporelles sur un corpus de 100 dépêches de la langue anglaise de ProMed. Nous avons utilisé le système d'extraction d'information INTEX pour la construction de ces transducteurs. Les résultats obtenus présentent une efficacité de ces graphes pour l'identification des données temporelles.

EpidémIA project aims to the construction of a computerized decision-making system to assist user in his activity of medical risk management. A preliminary work on the events compositionality (STEEL) enables us to direct our work in the field of the space-time information localization. We have created some transducers graphs to identify temporal information on a corpus of 100 SARS ProMed English reports. We used the extraction information system INTEX to construct these transducers. The results obtained present an effectiveness of these graphs to identify temporal data.

1 Introduction

Les déplacements individuels dans un cadre professionnel ou de loisir et l'essor des nouvelles épidémies sur la planète ont conduit à la création d'un nombre croissant de dispositifs de veille sanitaire dont la liste de diffusion internationale PROMED (http://www.promedmail.org). Simultanément du fait d'Internet, nous assistons à ladématérialisation généralisée de l'information utilisable pour la veille épidémiologique. Dans ce cadre, le projet EpidémIA a pour objectif le traitement des dépêches de ProMed pour l'analyse, la modélisation, la gestion et la restitution de connaissances et des données épidémiologiques. Un travail préalable nous a permis de développer un langage formel de représentation des connaissances (STEEL) adapté à la problématique des informations épidémiologiques, tenant compte à la fois de l'orientation évènementielle des récits, de la compositionalité des évènements et de leur localisation spatio-temporelle (Chaudet, 2004). Le travail que nous présentons ici concerne l'aspect TALN de projet, et en particulier la mise au point d'une méthode automatique d'extraction d'information de ces composants afin de pouvoir les utiliser pour l'inclure dans le modèle STEEL. Dans ce travail, nous abordons en particulier le problème d'identification et d'extraction des expressions temporelles en créant des automates INTEX (Silberztein, 1993). Orienté par STEEL, nos graphes de transducteurs recherchent les séquences possibles, dans le corpus des dépêches, contenants les éléments sémantiques nécessaires à la construction d'une représentation temporelle des évènements.

Dans cet article nous présentons d'abord le modèle STEEL sur lequel reposera la configuration des transducteurs INTEX. Nous dressons ensuite un état de l'art sur les différentes approches d'annotation temporelles en les comparant à la nôtre. Une étude des expressions temporelles est ensuite effectuée sur notre corpus, formé de 100 dépêches (248670 mots) de l'épidémie de SRAS (partie 4). Ceci nous permettra de proposer un graphe temporel spécifique tout en présentant et commentant les résultats obtenus.

2 Le modèle STEEL

Dans le cadre de la modélisation des systèmes dynamiques, plusieurs formalismes adaptés au raisonnement actions/évènements et leurs effets ont été proposés. Le calcul d'évènement de Kowalski et Sergot (1986) est un de ceux-ci. En se fondant sur cette approche et celle de Cervesato et Montanari (2000), Chaudet (2004) a créé une adaptation pour la représentation des récits épidémiologiques, STEEL (Spatio-Temporal Extended Event Language), qui se caractérise par la possibilité de représenter des agrégats d'évènements spatio-temporellement localisés. Les entités temporelles et spatiales y sont réifiées et introduites comme arguments de prédicats spécifiques dans une théorie de premier ordre. De plus le langage STEEL intègre les primitives de manipulation des évènements. Pour simplifier, au lieu de la forme traditionnelle *Happens(action,temps)*, les évènements selon STEEL se produisent dans un lieu spécifique : *Happens(macro-évènements,temps,espace)*. Trois composantes du discours doivent donc être identifiées et représentées de façon coordonnée dans le langage de représentation : l'évènement (simple ou complexe), le temps et le lieu. Par exemple, pour la phrase *The total number of dengue-affected patients, according to the official account, stood at 4763, as of 16 Sep 2003. Of these, 45 have died so far,* STEEL donne:

happens (e1,<t1, Bangladesh>, <2002-09-16, Bangladesh>)
happens (e2, <t2, Bangladesh>, <2002-09-16, Bangladesh>)

instance (e1,macroevent); instance (e2, macroevent)
meventdef(e1, iterevent(infection,4763)) ; meventdef(e2, iterevent(death,45))
agent (e1, dengue)
experiencer (e2,a) → experiencer (e1,a)
$t1 \leq 2002\text{-}09\text{-}16 \wedge t1 \leq t2 \wedge t2 \leq 2002\text{-}09\text{-}16$

3 Localisation temporelle

L'annotation précise et détaillée des expressions temporelles a commencé avec les conférences MUC 5-7 (Message Understanding Conferences) pour l'identification et la classification des entités nommées EN (Chinchor, 1999). Dans la même vision, Ferro et al. (2001) décrivent un ensemble de consignes pour l'annotation des expressions temporelles à partir de plusieurs langues, et leur associent une représentation canonique du temps à laquelle elles se réfèrent. Cependant, une autre approche de l'annotation a aussi été utilisée. C'est le marquage temporel, qui vise à associer un temps du calendrier à certains ou tous les événements du texte. Filatova et Hovy (2001) décrivent une méthode pour fractionner des phrases en leurs événements constitutifs et leur assigner des marqueurs temporels. Le marquage utilise deux temps principaux : le temps de l'article et le dernier temps indiqué dans la même phrase. Dans cette approche Schilder et Habel (2001) ont développé un système d'étiquetage sémantique des expressions temporelles. Elles sont classifiées selon deux types : celles qui se rapportent à un temps du calendrier ou d'horloge et celles qui se rapportent à des événements. L'ensemble des relations temporelles proposées est équivalent aux relations d'Allen (1983). Une troisième approche (Setzer et Gaizauskas, 2000) se focalise sur les relations temporelles entre les événements et le temps ou entre les événements mêmes. Cette approche prend l'identification des relations temporelles comme but et repose sur la façon dont l'information temporelle se présente ainsi que sa relation avec le texte. Leur schéma permet de déterminer l'ordre relatif ou le temps absolu des évènements. Katz et Arosio (2001) à leur tour ont proposé une annotation des informations des relations temporelles en se basant sur les relations entre évènement. Notre approche se rattache à la deuxième catégorie de travaux. L'annotation est pratiquée en tenant compte de la date de la dépêche et de celle signalée dans le récit. L'association entre l'événement et le marquage temporel se fait ultérieurement au niveau de la représentation logique.

4 Méthodologie

Notre approche permet d'analyser des membres de phrases qui peuvent comporter une expression temporelle. Dans ce cadre, nous l'avons décomposée en 6 étapes :

1 Isoler les mots caractérisant la localisation temporelle des évènements épidémiques,
2 Construire des dictionnaires[1] spécifiques pour ces mots. Par exemple, pour *April* la forme sera : *Apr,N+Month* et pour *Hong Kong* : *Hong Kong,N+Location*,etc.
3 Sélectionner dans le corpus les membres de phrases comportant un élément temporel par une recherche automatique (INTEX) des mots typés à l'étape 2,

[1] Un dictionnaire INTEX est une liste de termes avec une étiquette et une liste d'information sémantique associée aux informations flexionnelles et lexicales

4 Analyser la configuration syntaxique et sémantique qui entoure l'élément temporel. INTEX construit les graphes syntaxiques correspondant aux membres de phrases sélectionnés. Mais, il donne plusieurs solutions. Les ambiguïtés sont nombreuses, dues aux dictionnaires non spécifiques du domaine. Il est donc nécessaire de construire des dictionnaires spécifiques,

5 Elaborer les graphes correspondants. Ces graphes sont utilisés pour les formes complexes où des phénomènes d'insertion et d'optionalité interviennent. Nous donnons en figure 1 un exemple de transducteur[2] reconnaissant les numéro du jour ainsi que le jour en caractère. Dans ce schéma, la première partie <N+DateJour> désigne dans les dictionnaires de temps les différentes formes des noms des jours de la semaine comme (Monday, Mon, mon, MONDAY, etc.). La deuxième partie Day_num est le sous graphe qui est constitué des numéros des jours d'un mois (figure 2),

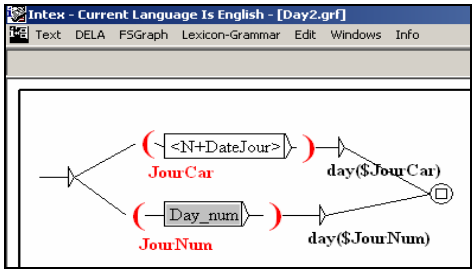


Figure1-Transducteur de Jour

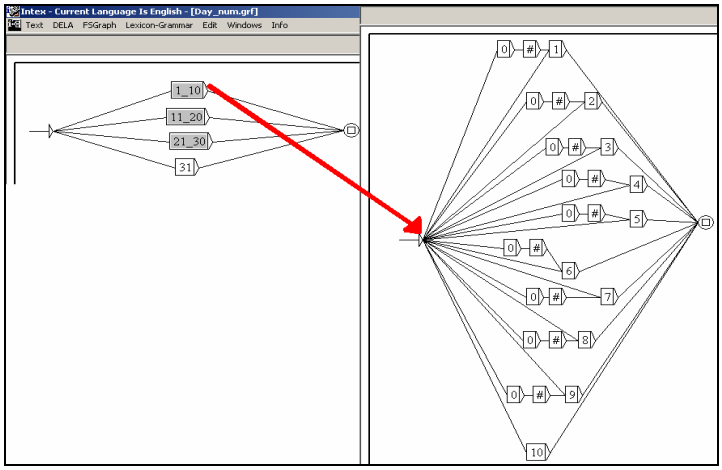


Figure 2-Numéro des jours du mois

6 Tester leurs performances en évaluant le taux d'erreur : rapport entre le nombre de séquences reconnues erronées sur le nombre total des séquences identifiées.

5 Résultats et discussion

Afin de créer les graphes de localisations temporelles, nous avons étudié les différentes formes présentes dans notre corpus formé de 100 dépêches de l'épidémie du SRAS. Elles ont été regroupées selon deux catégories principales.

Un graphe de transducteurs a été construit pour identifier ces expressions temporelles (Figure 3). En particulier, l'étiquette **"Month($MoisNum)"** provoque l'écriture du prédicat **"Month"** avec un argument: la valeur de la variable **"MoisNum"**, valeur trouvée dans le texte. Dans l'échantillon de 100 dépêches, 6284 séquences sont reconnues par ce graphe. Le tableau 1 présente quelques séquences reconnues, ainsi que leurs équivalents en mode de remplacement. Premièrement, des formes langagières spécifiques des dépêches ont été identifiées. Il s'agit de plusieurs formats non littéraires, de style rédactionnel abrégé comme les expressions suivantes, *10 Apr 2003, April 10, 2003, 10 April 2003, 2003_03_10, 2003-*

[2] Un transducteur est un automate qui reconnaît des séquences de mots et peut produire une nouvelle séquence

04/10, Friday [10 Apr 200, 10 Apr 2003, 2003/04/10, 10-Apr-2003, 10 April 2003, April 10, 2003, etc. Deuxièmement, les cas d'expressions temporelles présentant des formes littéraire plus classiques ont été identifiées dans notre corpus comme *In mid February, after several days, in recent days, during the second week of February, Monday evening, last Friday night.*

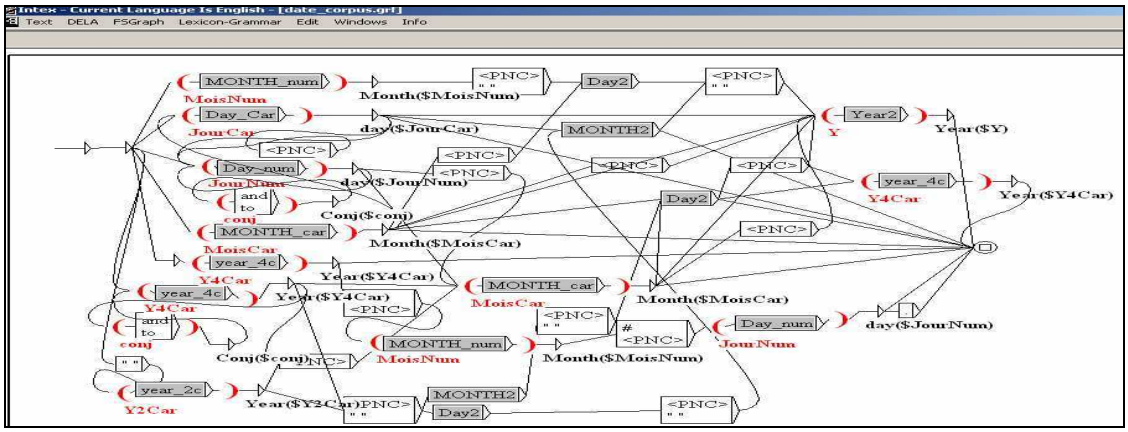


Figure 3- Annotation des expressions temporelles

Séquences reconnues	***Résultat en Mode de remplacement***
Worldwide 20030315.0637 …	Worldwide Year(2003)Month(03)day(15) 0637
Friday [18 Apr 2003]	day(Friday (day(18)Month(Apr)Year(2003)]
Monday [21 Apr 2003]	day(Monday)day(21)Month(Apr)Year(2003)]

Tableau 1. Exemples de séquences reconnues

Pour cela, nous avons bénéficié de la bibliothèque de graphes de Maurice Gross disponible sur le site d'INTEX. Appliqués sur ces dépêches, ces graphes identifient les formes littéraires des expressions temporelles dans ce corpus. Cependant, ils identifient à tort d'autres expressions comme *that **may** have identified, **from 16** countries, **in a second** Hong Kong hospital, **in terms** of industries, **on the 9th** floor.* Pour réduire ces cas d'erreurs, nous avons modifié les graphes concernés. Ils s'adaptent mieux au langage professionnel utilisé dans les dépêches.

Le graphe qui englobe finalement l'ensemble des formes associées aux expressions temporelles est composé des deux graphes décrits précédemment. 7087 cas ont pu être identifié dans notre corpus par l'application de ce graphe. Il reconnaît des expressions littéraires et non littéraires. Nous présentons dans le tableau 2 quelques séquences reconnues, ainsi que leurs résultats en mode de remplacement.

Nous avons décrit dans cet article une expérimentation d'utilisation d'INTEX afin d'extraire les expressions temporelles. Elle a permis d'extraire toutes les expressions temporelles, avec un taux d'erreur de 3 sur 6284 formes. Une solution serait, afin d'améliorer l'identification de ces différentes expressions, de passer à une forme générale, qui serait obtenue grâce à la fonction de génération des transducteurs. On substituerait à toutes séquences d'origine, des séquences générées à partir de marqueurs définis (figés) dans les graphes et de mots ou

marqueurs sémantiques récupérés par INTEX dans des variables, à partir du texte d'origine. Dans une étape ultérieure, cette forme devrait être traduite dans le langage STEEL.

Séquence Reconnues	***Résultat en mode de remplacement***
in the evening on 10 Apr 2003	ExpTemp(in the evening)day(10)Month(Apr))Year(2003)
the previous month [February]	ExpTemp(the previous month)Month(February)
yesterday (28 Apr 2003)	ExpTemp(yesterday)day(28)Month(Apr)Year(2003)

Tableau 2. Exemples de séquences reconnues

Références

Allen J.F. (1983), Maintaining Knowledge about Temporal Intervals, *Communications of the ACM,* vol 26(11), pp. 832-843.

Cervesato I., Montanari A. (2000), A Calculus of Macro-Events: Progress Report. *In 7th International Workshop on Temporal Representation and Reasoning*, 47-58.

Chaudet H. (2004), Une extension du Calcul des Evènements pour la représentation de récits épidémiologiques, *15ème journée francophone IC'2004*, 285-296.

Chinchor N., Brown E., Ferro L., Robinson P. (1999), Named Entity Recognition Task Definition, *MITRE.*

Ferro L., Mani I., Sundheim B., Wilson G. (2001), TIDES Temporal Annotation Guidelines Draft - Version 1.02, *MITRE Technical Report MTR 01W000004.* McLean, Virginia.

Filatova E., Hovy E. (2001), Assigning Time-Stamps to Event-Clauses, *In Proceedings of the ACL-2001 Workshop on Temporal and Spatial Information Processing.*

Katz G., Arosio F. (2001), The Annotation of Temporal Information in Natural Language Sentences. *In Proceedings of the ACL-2001 Workshop on Temporal and Spatial Information Processing*, 104-111.

Kowalski R., Sergot M. (1986), A Logic-based Calculus of Event, *New Generation Computing*, Vol.4 , pp.67-95.

Schilder F., Habel C. (2001), From Temporal Expressions To Temporal Information: Semantic Tagging Of News Messages, *In Proceedings of the ACL-2001 Workshop on Temporal and Spatial Information Processing,* 65-72.

Setzer A., Gaizauskas R. (2000), Annotating Events and Temporal Information in Newswire Texts. *In Proceedings of the Second International Conference on Language Resources and Evaluation*, 1287-1294.

Silberztein M. (1993), *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX.* Masson: Paris.

RECITAL 2005, Dourdan, 6–10 juin 2005

Utilisation de la Linguistique Systémique Fonctionnelle pour la détection des noms de personnes ambigus

Thomas Heitz
LRI - Université Paris Sud
91405 Orsay Cedex - France
heitz@lri.fr

Mots-clefs – Keywords

Linguistique Systémique Fonctionnelle, Détection des entités nommées, Fouille de textes
Systemic Functional Linguistics, Named entity recognition, Text mining

Résumé - Abstract

Dans cet article, nous nous proposons de construire un lexique étiqueté selon les principes de la Linguistique Systémique Fonctionnelle (LSF) et de l'appliquer à la détection des noms de personnes ambigus dans les textes. Nous ne faisons pas d'analyse complète mais testons plutôt si certaines caractéristiques de la LSF peuvent permettre de compléter les modèles linguistiques actuellement utilisés pour la détection des entités nommées. Nous souhaitons ainsi apporter une contribution à l'application du formalisme LSF dans l'analyse automatique de textes après son application déjà éprouvée à la génération de textes.

In this paper, we propose to build a tagged lexicon according to the Systemic Functional Linguistics (SFL) principles and to apply it to the recognition of ambiguous person names in texts. We do not achieve a complete analysis but rather test if some characteristics of the SFL could enable the completion of actual linguistic models used in named entity recognition. We want thus to bring a contribution to the application of the SFL model in automatic text analysis while it already proved its usefulness for text generation.

1 Introduction

Nous allons discuter de l'intérêt du formalisme de la Linguistique Systémique Fonctionnelle (LSF) pour la détection des noms de personnes ambigus dans les textes. Nous nous situons donc dans le domaine de la détection des entités nommées qui est une des tâches essentielles de la fouille de textes (Daille & Morin, 2000).

La LSF (Halliday & Matthiessen, 2004) offre une approche complémentaire des linguistiques habituellement utilisées. La LSF part du contexte social et dès le départ cherche à comprendre la fonction des mots alors que traditionnellement les linguistiques partent des mots pour arriver finalement au sens.

Nous souhaitons ici étudier l'intérêt de caractéristiques de la LSF dans le but de les incorporer dans les dernières étapes d'un système de détection d'entités nommées similaire à ceux utilisés lors de la compétition MUC-7 (Mikheev *et al.*, 1998).

Dans le domaine de la détection des entités nommées, il a été constaté que le problème de la poly-catégorisation d'une entité nommée nécessite une meilleure analyse du contexte pour la traiter (Daille & Morin, 2000). La poly-catégorisation existe, par exemple, pour le nom ambigu de personne *France*, qui peut aussi être un nom de pays.

Nous nous intéresserons ici aux verbes appartenant aux contextes des noms de personnes. Plus précisément, nous chercherons les processus de la LSF associés à ceux-ci puisque les processus sont portés par les verbes. Les processus d'une phrase étant les éléments qui désignent la ou les actions en cours entre les différents acteurs participant à l'action et pour des circonstances données. Puis, nous constaterons si des liens existent entre catégories de processus et contextes de noms de personnes. Ainsi, nous pensons pouvoir améliorer la discrimination entre les noms de personnes et les autres noms propres ou les mots communs dans les cas ambigus.

Nous allons établir une méthodologie pour la détection des noms de personnes ambigus à l'aide des processus de la LSF sur une première partie d'un corpus. Puis, nous validerons expérimentalement notre hypothèse d'amélioration de la détection des noms de personnes ambigus sur une autre partie du même corpus.

1.1 Extraction des noms de personnes

Comme la recherche sur les entités nommées l'a déjà établi (Poibeau, 2001), les noms de personnes sont souvent entourés par des mots inclus dans l'entité nommée. A savoir, les noms de fonction ou de titre qui peuvent se situer en préfixe comme *Mr.*, en infixe comme *de la* ou en suffixe comme *Second du nom*. Par exemple *Mr. Arnaud de la Tour Second du nom*. De nombreuses grammaires et lexiques ont été développés pour repérer ces formes.

En plus de ces mots inclus dans le nom de personne, que nous appellerons indices locaux, les indices contextuels ont été aussi étudiés. Par exemple, la préposition *chez* introduit presque toujours une personne. De même, des grammaires utilisant ces indices ont été développées. Mais ces indices contextuels restent la plupart du temps limités aux mots grammaticaux que sont les déterminants, conjonctions et prépositions pour les règles non apprises. Nous voulons donc l'étendre aux mots lexicaux que sont les verbes, noms, adjectifs et adverbes afin de mieux traiter les cas ambigus de noms de personnes. Dans cet article, nous nous limiterons aux verbes appartenant aux contextes des noms de personnes.

1.2 Processus dans la Linguistique Systémique Fonctionnelle

Les processus appartiennent à la fonction idéationnelle de la LSF (Halliday & Matthiessen, 2004) qui est un modèle d'analyse de la phrase qui divise celle-ci en processus, participants et circonstances. Dans ce modèle, les verbes peuvent être catégorisés dans des processus puisqu'ils les supportent. Il existe six principaux processus : matériel, mental, verbal, relationnel, comportemental et existentiel. Dans le tableau 1 sont caractérisés les participants des différents processus de la LSF.

Processus	Sens	Nombre de participant	Nature du premier participant	Nature du second participant
Matériel	faire et avoir lieu	1 ou 2	chose	chose
Mental	resentir	2	chose consciente	chose ou fait
Verbal	dire	1	source de symboles	
Relationnel	être et avoir	1 ou 2	chose ou fait	
Comportemental	se comporter	1	chose consciente	
Existentiel	exister	1 ou 0[1]	chose ou fait	

Table 1: Caractéristiques des participants des processus de la LSF

Les personnes, en tant que premiers participants, sont impliquées presque systématiquement dans les processus mentaux, comportementaux et verbaux comme chose consciente pour les deux premiers et source de symbole pour le troisième et d'une façon plus incertaine dans les autres processus. Nous allons donc tester l'implication de ces processus dans les contextes de noms de personnes sur un corpus afin de retenir les catégories pertinentes.

2 Méthodologie pour la détection des noms de personnes ambigus à l'aide des processus de la LSF

2.1 Extraction des relations syntaxiques et des noms de personnes

Nous avons utilisé un sous ensemble du corpus Aquaint d'une taille d'environ 10 mégaoctets provenant de la compétition TREC Novelty 2004 (Soboroff, 2004). Il contient des dépêches écrites en anglais en provenance d'agences de presse.

Un lexique de verbes étiquetés selon les processus de la LSF à été élaboré manuellement à partir de (Halliday & Matthiessen, 2004) et étendu en rajoutant l'intersection des hyponymes et synonymes contenus dans WordNet (Miller *et al.*, 1990) pour chaque verbe initial. Puis, nous avons extrait à partir du corpus une liste de noms de personnes de manière semi-automatique avec des listes de noms de personnes déjà connus et une correction manuelle.

Le logiciel Link Grammar 4.1 (Sleator & Temperley, 1993) a été utilisé afin d'extraire les relations syntaxiques des phrases du corpus. Seules les relations syntaxiques de type Sujet-Verbe dénommées S et SI dans la nomenclature de Link Grammar ont été prises en compte.

[1]Par exemple, dans la phrase : *There was a storm.* il n'y a pas de participant juste un processus.

Nous avons ensuite lemmatisé les verbes, étiqueté les verbes lemmatisés selon les processus de la LSF et étiqueté les noms de personnes. Une partie du corpus contenant des noms de personnes ambigus a été mise de côté afin de pouvoir tester les hypothèses faite dans la section 2.2.

2.2 Établissement de correspondances entre les processus de la LSF et les contextes de noms de personnes

Nous avons cherché les types de processus qui qualifient les verbes dont les sujets sont des noms de personnes à partir des relations Sujet-Verbe trouvées précédemment dans la section 2.1. Les résultats sont présentés dans le tableau 2.

Processus	Sens	Lexique	Relation Personne-Verbe	Comparaison
Matériel	faire et avoir lieu	46% (2462)	71% (6627)	+25
Mental	resentir	12% (653)	66% (6098)	**+54**
Verbal	dire	7% (353)	49% (4557)	**+42**
Relationnel	être et avoir	32% (1698)	71% (6609)	+39
Comportemental	se comporter	3% (166)	8% (345)	+5
Existentiel	exister	2% (129)	14% (1305)	+12
Total		5353 verbes	9285 relations	

Table 2: Occurrences des processus dans les relations Personne-Verbe. Il existe une catégorie par ligne dans le lexique mais plusieurs par ligne dans les relations.

Nous pouvons constater que les processus de types mental et verbal semblent les plus adaptés pour trouver les contextes de noms de personnes puisque ce sont les catégories de processus dont la proportion augmente le plus par rapport à celle du lexique utilisé. Ceci confirme en partie ce qui était attendu d'après le modèle théorique exposé dans la section 1.2. Cependant la catégorie relationnel augmente aussi beaucoup bien que les verbes obtenus ne soient pas très pertinents. Il sera donc utile dans une prochaine expérimentation de tenir compte des sous catégories pour obtenir des résultats plus précis.

3 Validation expérimentale

3.1 Ambiguïtés entre noms de personnes et autres noms propres

De nombreux noms de personnes sont aussi des noms d'organisations ou de lieux notamment. Plus généralement, ceci rejoint le problème de la poly-catégorisation des entités nommées.

Dans le tableau 3, sont présentés quelques cas d'ambiguïté avec les autres noms propres. Les résultats sont donnés sous forme de fractions comportant au numérateur le nombre de mots bien identifiés comme un nom de personne ou comme un autre nom propre et au dénominateur le nombre d'occurrences de ce mot pour lesquels les verbes dont il est le sujet ont été catégorisés. Les verbes sont catégorisés selon les processus de la LSF et **lorsqu'ils appartiennent à au moins deux types parmi mental, verbal et comportemental nous considérons que le mot**

est un nom de personne et non une organisation ou un lieu. Les modaux ne sont pas pris en compte.

Type d'ambiguïté	Nom	Précision
Personne - Organisation	Ford	24/35
	Bell	0/5
	Morris	0/0
Personne - Lieu	Virginia	0/0
	Madison	0/0

Table 3: Résultats pour la distinction entre noms de personnes et autres noms propres

3.2 Ambiguïtés entre noms de personnes et mots communs

Il existe de nombreux cas où la distinction entre noms propres et mots communs ne peut s'effectuer grâce à la casse de la première lettre du mot. Les principaux cas sont : premier mot de phrase ou de citation, titre, transcription de l'oral vers l'écrit tout en minuscules, langue sans différenciation de casse entre noms propres et mots communs. En allemand, par exemple, les mots communs s'écrivent avec une majuscule en première lettre comme pour les noms propres. Ceci justifie de traiter les cas d'ambiguïtés entre noms de personnes et mots communs.

Une phrase exemplaire tirée de notre corpus est la suivante : **Ray** *believes the electron beam process is superior to gamma* **ray** *irradiation because the technology is easier and cheaper to implement.*

On peut y voir le mot *ray* sous la forme de nom de personne et de mot commun. La majuscule de début de mot ne suffit pas ici à les distinguer car *Ray* est le premier mot de la phrase qui par définition possède toujours une majuscule en première lettre. Le verbe *to believe*, croire, qui le suit est classé dans les processus de type mental et correspond bien à une catégorie de processus appartenant à des contextes de noms de personnes comme vu dans la section 2.2.

Dans le tableau 4, trois des cas les plus fréquents d'ambiguïtés entre noms de personnes et mots communs sont présentés. Les résultats sont donnés sous la même forme que les résultats précédents et avec le même protocole.

Type d'ambiguïté	Nom	Précision
Personne - Nom commun	bill	49/66
	bird	0/2
	stone	3/3
	ray	4/4
Personne - Verbe	drew	0/0
	mark	0/0
Personne - Adjectif	brown	11/11
	gray	20/32

Table 4: Résultats pour la distinction entre noms de personnes et mots communs

Exemples de phrases traitées :

The **bill died** *in a Transportation subcommittee, but Moreno said it had widespread support in the house [...]* To die, processus mental impliquant un nom de personne, mal catégorisé.

Mel-Daniels was an assistant coach at State when **Bird led** *the team the NCAA final in 1979.* To lead, processus verbal impliquant un nom de personne, bien catégorisé.

On trouve notamment le problème de la poly-catégorisation comme pour le verbe *to die* qui ne devrait pas être classé mental mais matériel et très rarement comme processus mental. Beaucoup de mots communs sont ici personnifiés comme pour le mot *bill*, *projet de loi* en français, ce qui peut être source d'erreurs.

4 Conclusion et perspectives

Nous nous sommes ici intéressé aux verbes dont le sujet est un nom de personne. D'après ces premiers résultats, même s'il apparaît important de chercher quelle action subit ou effectue un être conscient pour découvrir tous les noms de personnes d'un texte, il semble nécessaire d'avoir des informations supplémentaires comme les autres participants des processus et les circonstances. L'utilisation des sous catégories de processus couplée à l'apprentissage de règles de classification semble aussi une piste à suivre.

Une comparaison avec d'autres méthodes d'expansion de termes pour notre lexique de base et d'autres lexiques catégorisant les verbes différemment serait utile. De même qu'une comparaison avec d'autres systèmes de détection d'entités nommées sur des corpus de référence une fois l'analyse des processus de la LSF intégrée dans un système de détection d'entités nommées.

Je remercie mon directeur de thèse, Yves Kodratoff, pour m'avoir fait découvrir la LSF.

Références

DAILLE B. & MORIN E. (2000). *Reconnaissance automatique des noms propres de la langue écrite : les récentes réalisations*, In *Traitement automatique des noms propres sous la direction de Denis Maurel et Franz Guenthner*, chapter 1, p. 601–621. Hermes. Collection : Traitement automatique des langues.

HALLIDAY M. A. K. & MATTHIESSEN C. M. I. M. (2004). *An Introduction to Functional Grammar*. Hodder Arnold. 3nd. edition.

MIKHEEV A., GROVER C. & MOENS M. (1998). Description of the ltg system used for muc-7. In *Message Understanding Conference Proceedings, MUC-7*.

MILLER G. A., BECKWITH R., FELLBAUM C., GROSS D. & MILLER K. J. (1990). Introduction to wordnet: an on-line lexical database. *International Journal of Lexicography 3 (4)*, p. 235–244. revised august 1993.

POIBEAU T. (2001). Deconstructing harry, une évaluation des systèmes de repérage d'entités nommées. *Revue de la société d'électronique, d'électricité et de traitement de l'information*.

SLEATOR D. & TEMPERLEY D. (1993). Parsing english with a link grammar. In *Third International Workshop on Parsing Technologies*.

SOBOROFF I. (2004). Overview of the trec 2004 novelty track. In *NIST Special Publication: The Thirteenth Text Retrieval Conference (TREC 2004)*, p. 57–70.

Durée des consonnes géminées en parole arabe : mesures et comparaison

Mohamed Khairallah KHOUJA (1), Mounir ZRIGUI (2)

(1) et (2) Laboratoire RIADI (unité de Monastir)

(1) khairallah_k@yahoo.fr

(2) Faculté des Sciences de Monastir,

mounir.zrigui@fsm.rnu.tn

Mots-clefs – Keywords

Analyse acoustique, arabe standard, gémination, durée, reconnaissance de la parole.

Acoustic analyze, standard Arabic, gemination, duration, speech recognition.

Résumé – Abstract

Dans ce papier, nous présentons les résultats d'une étude expérimentale de la durée des consonnes géminées de l'arabe. Nous visons à déterminer la durée, pour une séquence VCCV, de la consonne géminée CC ainsi que de la voyelle qui la précède. Nous comparons ces valeurs à celles mesurées pour une séquence VCV. Les résultats ont prouvé que la durée de la consonne simple était sensiblement différente de celle géminée, ainsi que la durée de la voyelle précédant la consonne. A la base, ce travail est entrepris dans un but d'étudier l'utilisation des durées de phonèmes comme une source d'information pour optimiser un système de reconnaissance, donc introduire des modèles explicites de durée des phonèmes, et mettre en application ces modèles comme partie du modèle acoustique du système de reconnaissance.

In this paper, we represent the results of an experimental study concerning the duration of geminated consonants in the Arabic language. We are seeking to determine the duration of a sequence VCCV, of the geminated consonant CC and of the vowel that precedes it. We are comparing these items with those measured for a sequence VCV. The results have shown that the duration of the simple consonant was so apparently different from that geminated and the same was true for the duration of the vowel that precedes such consonant. Basically, this work is undertaken with an aim of studying the use of the durations of phonemes like a source of information to optimize a recognition system, therefore to introduce explicit models of duration of the phonemes, and to apply these models like part of the acoustic model of the recognition system.

1 Introduction

L'arabe est une langue dans laquelle les durées des phonèmes jouent un rôle distinctif dans la reconnaissance automatique de la parole (RAP). L'information en durée est la plupart du temps négligée dans les systèmes de RAP dus à l'utilisation des modèles cachés de Markov qui sont déficients pour une modélisation correcte des durées des phonèmes (Selouani S-A., 2000).

2 La gémination dans l'arabe

La gémination est définie comme étant la succession de deux consonnes identiques prononcées consécutivement. En arabe, la gémination est exprimée à l'aide du symbole « ّ » (الشدّة). Ce symbole joue un rôle important dans la définition et le sens de certains mots.

Pour la langue arabe le paramètre de durée est très important tant au niveau sémantique qu'au niveau grammatical. Il caractérise non seulement les voyelles, mais également les consonnes géminées. Concernant ce trait, un double problème se pose en RAP de l'arabe : il faut déceler les phonèmes allongés tout en s'assurant que ce prolongement est pertinent, c'est à dire en le distinguant des allongements dus au débit d'élocution, à un accent particulier du locuteur, etc. En effet si l'on observe l'exemple du mot « صلّى » /salla:/ (prier) dérivé de la racine « صلا » qui ne s'oppose que par la gémination de la consonne « ل » au mot « صلى » /sala :/ (griller) dérivé de la racine « صلى », nous mesurons à la fois l'importance et la difficulté d'un système automatique à déceler ce trait. La gémination se manifeste par le renforcement de l'articulation et une prolongation de la fermeture de la plosive ou du continuant des autres consonnes (Barkat M., 2000).

Plusieurs études similaires à ce travail ont été présentées pour d'autres langues, ou la gémination est considérée comme un trait remarquable, notamment celle pour l'italien (Giovanardi M., Di Benedetto M-G., 1998), le grec (Arvaniti A., Tserdanelis G., 2000) et l'indien (Samudravijaya K., 2003). Pour l'arabe standard aucune étude n'est faite à nos connaissances dans le cadre de la parole spontanée. Les travaux de Jomaa (Jomaa M., 1993) et Allatif (Allatif O., Abry C., 2004), sont intéressés par l'effet de la quantité de contraste pour les séquences VVC et CCV et à la quantité vocalique pour l'arabe dialectale.

3 Données et mesures

3.1 Acquisition des données

Vu l'absence d'un corpus annoté, nous étions amenés à effectuer nos propres enregistrements à l'aide du logiciel PRAAT[1] et d'un microphone professionnel mis à la même distance de la bouche de chacun des quatre locuteurs (2 hommes et 2 femmes) participant aux enregistrements, tous dans la même salle pratiquement isolée. On a demandé aux locuteurs de

[1] Logiciel développé à l'institut des sciences phonétiques de l'université d'Amsterdam par P. Boersma et D. Weenink.

lire avec une vitesse moyenne tout en assurant une bonne articulation et en évitant les perturbations dus aux hésitations, les reprises, les respirations, …, dans de tel cas le locuteur est invité à reprendre sa lecture. La fréquence d'échantillonnage des enregistrements est 22050 Hz. Des phrases contenant les mots choisis pour l'étude sont lus les unes après les autres sans arrêt. La durée totale des enregistrements est de 1 h 27 min. Les mots sont choisie afin d'avoir au moins, si possible, deux fois la couverture de toutes les consonnes dans le cas de gémination et simple. Le nombre total des mots analysés a atteint 520 mots.

3.2 Analyse acoustique

Les données étudiées ont été extraites manuellement à l'aide du logiciel PRAAT. Le fait que tous les relevés aient été opérés par la même personne, garantie l'unicité de la méthode et des principes de base et donc l'homogénéité des données. Notre découpage a été fait manuellement sur la base d'indices visuels (spectre, amplitude, formant) le contrôle étant perceptif.

3.3 Mesures

La durée des différentes unités est considérée en générale comme le phénomène central pour la prosodie. En effet chaque variation de fréquence fondamentale ou d'intensité s'établit sur un certain laps de temps, durée mesurable. Etudier l'organisation temporelle de la parole est incontournable. Etudier la durée c'est observer et modéliser les durées d'unités bien déterminées. L'effort produit lors de l'articulation ne peut être infini, la séquence est toujours la même : effort et relâchement. Cet effort et ses variations sont véhiculés à travers le signal, ils permettent une segmentation temporelle de la parole en unités. Dans notre cas les unités à mesurer sont les phonèmes.

Pour chaque mot extrait des enregistrements on commence par en extraire les unités à mesurer donc la voyelle, la consonne où la consonne géminée. Chaque unité est éditer (signal + spectre + formants) séparément afin de pouvoir raffiner les limites et calculer la durée. Une moyenne de la durée pour chaque unité est calculée sur les quatre locuteurs.

4 Résultats

Les durées des diverses unités de la parole ont été calculées et employées pour étudier les caractéristiques en durée des consonnes et des voyelles qui les précèdes. Les changements observés des durées dues à la gémination sont présentés et discutés dans cette section.

4.1 Durée de la voyelle

La Figure 1 présente la durée moyenne pour les différents tests sur les trois voyelles courtes. Pour chaque couple de mots (simple et géminé) nous calculons la durée de la voyelle précédant la consonne sujet de la gémination. La durée pour les voyelles longues est calculée sur d'autres mots composant le corpus, elle est donnée à titre indicatif.

Nous constatons que la durée de la voyelle précédant une gémination diminue par rapport à son homologue (consonne simple) pour 86% des mots étudiés. En moyenne on a une diminution de 11 ms pour le /a/, 13 ms pour /u/ et 12 ms pour /i/. Cette diminution peut être expliquée par une tendance du locuteur à insister sur la gémination plutôt que sur la voyelle qui la précède. Ce résultat est trouvé pour d'autres langues tels que l'italien et l'indien (diminution de 10 à 15 ms) et il confirme les résultats présentés dans (Zeki M. Hassan 2002).

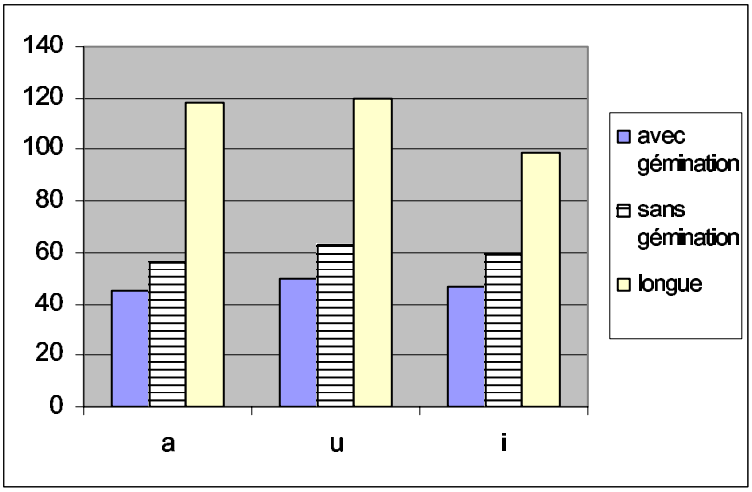


Figure 1 Représentation des durées moyennes (ms) des voyelles précédant une consonne simple ou géminé et celle d'une voyelle longue.

4.2 Durée des consonnes

L'un des exercices les plus périlleux du traitement de la parole consiste à déterminer les frontières des différentes unités phonétiques contenues dans un énoncé. Cette difficulté tient à la nature même de la parole continue : les unités sont fortement coarticulées, et l'on passe souvent de l'une à l'autre de manière continue.

Comme précisé en 3.3, nous avons mesuré la durée pour chaque consonne simple et son homologue géminée en conservant la même séquence qui précède et qui suit la consonne. Cela est possible puisque nous partons d'une racine verbale (فعل) et nous la dérivons en (فعّل), en plus de quelques noms et leurs dérivés.

Les résultats des moyennes de quelques unes de ces durées sont donnés par la Figure 2. Les valeurs sont arrondies et exprimées en ms.

Le rapport (G/S) donne une idée sur la différence entre une consonne simple et la même géminée. La thèse classique considérant ce rapport égal à 2, (Bonnot J.F., 1979), reste comme une moyenne mais n'est pas vraie pour toutes les consonnes puisque les résultats expérimentaux obtenus donnent un rapport variant entre 1,48 et 2,24.

Pour mieux cerné les caractéristiques en durée des consonnes nous avons établis la Figure 3, qui présente les durées suivant le type des consonnes.

Consonne	Simple	Géminé	G/S	Consonne	Simple	Géminé	G/S
/b/ ب	58	99	1,71	/t/ ت	62	105	1,69
/θ/ ث	65	123	1,89	/ʒ/ ج	70	121	1,73
/x/ خ	65	110	1,69	/ħ/ ح	63	102	1,62
/d/ د	69	145	2,10	/ð/ ذ	68	139	2,04
/r/ ر	64	95	1,48	/h/ ه	59	116	1,97
/s/ س	70	130	1,86	/ʃ/ ش	72	131	1,82
/ṭ/ ط	69	134	1,94	/ʕ/ ع	61	113	1,85
/f/ ف	62	129	2,08	/q/ ق	52	101	1,94
/k/ ك	67	129	1,93	/l/ ل	68	141	2,07
/m/ م	62	116	1,87	/n/ ن	60	124	2,07

Figure 2 Extrait des moyennes des durées des consonnes simples et géminées et leurs rapports

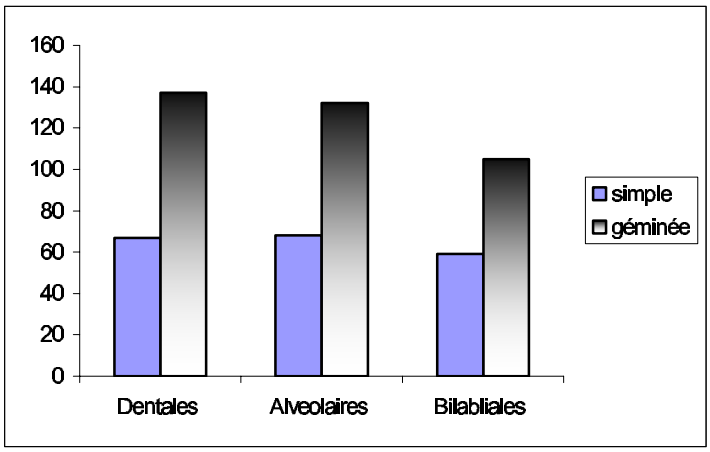


Figure 3 Représentation des durées (ms) pour certains types de consonnes dans les cas simples et géminés.

Le classement des consonnes suivant leurs types donne une vision plus claire sur la notion de durée. Cette classification montre bien la différence entre les différentes classes de consonnes puisque le rapport géminé/simple est plus caractérisant : 2,04 pour les dentales, 1,94 pour les alvéolaires et 1,78 pour les bilabiales. Ces résultats restent préliminaires puisqu'une étude plus complète sur un corpus plus important sera plus consistante.

5 Conclusion

Dans cet article nous avons présenté les résultats des mesures en durée des consonnes géminées pour la parole arabe. Les résultats statistiques obtenus montrent bien que nous pouvons caractériser une consonne simple d'une géminée d'un point de vu durée. Les résultats ont prouvé que la durée de la consonne géminée était sensiblement différente que celle simple avec les rapports présentés, ainsi que la durée de la voyelle précédant cette consonne. Pour pouvoir résoudre le problème de la gémination, dans un système de (RAP) continu pour la langue arabe, il faudrait différencier entre une gémination et une consonne simple suivie d'une voyelle longue. Cela passe par une bonne approche automatique acoustique pour la détermination des frontières entre consonne et voyelle.

Références

Allatif O., Abry C., (2004) Adaptabilité des paramètres temporels et spectraux dans l'opposition de quantité vocalique de l'arabe de Mayadin (Syrie), *JEP'2004*, 2004.

Arvaniti A., Tserdanelis G., (2000) On the phonetics of geminates: evidence from cypriot greek, *In proceedings of 6th International Conference on Spoken Language Processing*, volume 2, p: 559-562. Beijing, China, 2000.

Barkat M. (2000) Détermination d'indices acoustiques robustes pour l'identification automatique des parlers arabes, *Thèse de doctorat*, Université Lumière LYON 2, 2000.

Bonnot J.F. (1979) étude expérimentale de certains aspects de la gémination et de l'emphase en arabe, *travaux de l'institut phonétique de Strasbourg*, N°11, pp. 109-118, 1979.

Giovanardi M., Di Benedetto M-G. (1998) Acoustic Analysis of Singleton and Geminate Fricative in Italien, *The European Student Journal of Language and Speech*, 1998.

Jomaa M. (1993) Effect of quantity contrasts on the temporal regulation of mandibular movements in Arabic, *Actes du colloque du 2ième congrés : Langue arabe et technologies informatiques avancées*, pages : 141-169, 1993.

Samudravijaya K. (2003) Durational Characteristics of Hindi Stop Consonants, *EUROSPEECH 2003*, 2003.

Selouani S-A. (2000) Reconnaissance automatique de la parole par des techniques multi-agents, connexionnistes et hybrides, *Thèse d'état, Université des Sciences et Technologie Houari Boumedièn*ne, 2000.

Zeki M. Hassan (2002) Gemination in swedish & arabic with a particular reference to the preceding vowel duration. An instrumental & comparative approach, *TMH-QPSR* vol. 44 Fonetik 2002.

Vers une utilisation du TAL dans la description pédagogique de textes dans l'enseignement des langues

Mathieu LOISEAU

LIDILEM – Université Stendhal Grenoble 3
mathieu.loiseau@u-grenoble3.fr

Mots-clefs – Keywords

Corpus, ALAO, TAL, indexation pédagogique, ressources textuelles

Corpus, CALL, NLP, pedagogical indexation, textual resources

Résumé – Abstract

Alors que de nombreux travaux portent actuellement sur la linguistique de corpus, l'utilisation de textes authentiques en classe de langue, ou de corpus dans l'enseignement des langues (via concordanciers), quasiment aucun travail n'a été réalisé en vue de la réalisation de bases de textes à l'usage des enseignants de langue, indexées en fonction de critères relevant de la problématique de la didactique des langues.
Dans le cadre de cet article, nous proposons de préciser cette notion d'indexation pédagogique, puis de présenter les principaux standards de description de ressources pédagogiques existants, avant de montrer l'inadéquation de ces standards à la description de textes dans l'optique de leur utilisation dans l'enseignement des langues. Enfin nous en aborderons les conséquences relativement à la réalisation de la base.

Despite numerous works concerning corpus linguistics, the use of authentic texts in language teaching as well as corpora in language teaching (through concordancers), hardly any work deals with the creation of a teacher-friendly text base that would be indexed according to criteria relevant to the set of problems of language didactics.
In this article, we will first discuss the notion of pedagogical indexation. We will then present the principal pedagogical resource description standards, before showing that those standards are inadequate to handle our problem. Finally we shall give an overview of the consequences of these inadequacies on the implementation of the text base.

1 Indexation pédagogique pour l'enseignement des langues

Parmi les différentes plateformes d'apprentissage des langues assisté par ordinateur (ALAO), la plateforme MIRTO, en cours de développement à l'université Stendhal – Grenoble 3, se

démarque grâce à une approche résolument centrée sur la didactique des langues. Alors que de telles plateformes forcent, en règle générale, l'utilisateur enseignant – didacticien à reformuler sa problématique en termes informatiques, MIRTO reste avant tout un produit didactique : « *un programme qui met en oeuvre une solution didactique pour un problème de la didactique des langues sans altérer, ni la solution, ni, a fortiori, le problème.* » (Antoniadis et al., 2004). Ainsi MIRTO permet la conception, sans compétences informatiques préalables, d'activités didactiques. A l'heure actuelle, MIRTO est capable, à la donnée d'un texte brut et d'une série de paramètres entrée par l'enseignant, de générer divers types d'activités parmi lesquels des exercices lacunaires ou des exercices de traduction avec aide…

Cette philosophie est à l'origine du projet de création d'une base de textes indexée pédagogiquement pour l'enseignement des langues. En effet, grâce à son architecture, MIRTO est capable, de générer autant d'activités d'un type donné que l'on sera capable de lui fournir de textes. D'où l'idée d'intégrer un corpus à la plateforme. Cependant, deux textes différents ne sont pas strictement interchangeables pour une activité donnée. Les propriétés de chaque texte les rendront plus ou moins adaptés à une activité donnée. Un tel corpus ne pourra donc pas être une simple collection de textes, cette collection devra être organisée, indexée.

Depuis le lancement de ce projet, sa portée a été élargie, non seulement la base de textes devra être intégrée à la plateforme MIRTO, mais aussi pouvoir exister indépendamment de celle-ci et permettre à des enseignants de la consulter pour trouver des textes à utiliser en classe.

L'approche choisie pour la conception de cette base, s'inscrit dans la philosophie MIRTO, dans la mesure où ses trois fonctionnalités clés seront centrée sur l'enseignant :

- Exécution de requêtes selon des critères relevant de la didactique des langues.
- Ajout de textes dans la base, un processus qui doit être automatisé autant que faire se peut, pour ne pas rendre cette opération dissuasive de part sa complexité ou sa durée.
- Présence d'une interface qui permette aux enseignants n'ayant pas de connaissances particulières en informatique de pouvoir utiliser la base.

Ces fonctionnalités reposeront sur ce que l'on appellera l'indexation pédagogique de la base. On dira que des objets ont été indexés pédagogiquement, s'ils ont été indexés selon un système les décrivant en fonction de critères pédagogiques (relevant de la problématique de la didactique). Ici, les objets sont des textes et les critères pédagogiques relèvent de la didactique des langues. On parlera donc d'indexation pédagogique pour l'enseignement des langues.

Le travail d'indexation pédagogique en soi, reviendra donc, a priori, d'une part, aux utilisateurs du système et d'autre part au système lui-même, puisque lors de l'ajout d'un document dans la base, une partie du traitement sera vraisemblablement automatisée. C'est donc la définition du langage documentaire qui nous incombe. Elle repose sur trois problèmes distincts et interdépendants :

- L'utilisation des textes dans l'enseignement des langues, que ce soit avec ou sans l'intervention de l'ordinateur
- Le processus de recherche d'un texte par un enseignant de langue que nous modéliserons, à travers le recensement de critères de recherche.
- L'implémentation informatique de la base.

Nous détaillerons ici principalement une partie du troisième de ces aspects, même si aucun des trois aspects ne peut être traité complètement indépendamment des deux autres. Pour ce faire, nous utiliserons les résultats d'une étude préliminaire, qui nous a permis de commencer à mettre en évidence certains besoins des enseignants, comme le fait de pouvoir choisir un texte en fonction de son contenu lexical et grammatical ou encore celui d'avoir recours à des textes entiers et non à des portions de texte. Les résultats de cette étude ne sont pas définitifs, ils ont servi de base à la création d'un questionnaire à plus grande échelle. Ils permettent cependant d'évaluer l'adéquation des systèmes existants avec la notion d'indexation pédagogique de textes pour l'enseignement des langues.

Avant de proposer une solution *ad hoc* pour le langage documentaire à utiliser, nous devions nous assurer que l'existant ne répondait pas d'ores et déjà à notre problème, nous présenterons donc tout d'abord les principaux standards de description de ressources pédagogiques, après quoi nous les soumettrons à notre problématique avant de conclure sur le sujet.

2 Standards de description de ressources pédagogiques

2.1 Présentations des principaux standards

La Dublin Core Metadata Initiative (DCMI) constitue le standard le plus ancien parmi ceux que nous allons présenter. Bien que n'étant pas à proprement parler un standard de description « pédagogique », il a influencé les principaux standards de description d'objets pédagogiques existants, d'où sa présence ici. La DCMI est à l'origine du Dublin Core Element Set (DCES) qui permet de décrire des ressources à partir de quinze éléments, tous utilisables zéro, une ou plusieurs fois (Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights) (DCMI, 2005). Ces éléments très généraux ne permettront pas de décrire suffisamment précisément tout type de document mais le DCES n'est pas figé :

Ce modèle permet aux différentes communautés d'utilisateurs de se servir des éléments DC pour les informations descriptives fondamentales tout en autorisant les extensions spécifiques à un domaine et qui sont pertinentes pour un public moins large (Hillman, 2005) (Traduction de l'auteur)

Le Gateway to Educational Material (GEM, 2004) et Educational Network of Australia metadata (EdNA, 2002) constituent tous deux des exemples d'extensions / raffinements du DCES pour la description d'objets pédagogiques.

Enfin, le LOM (Learning Object Metadata), qui est probablement le plus utilisé et le plus influent de ces formats, est le fruit de la collaboration d'équipes du projet européen ARIADNE[1] et du consortium américain IMS Global Learning[1].

[1] http://www.ariadne-eu.org et http://www.imsproject.org respectivement

2.2 Caractéristiques communes

Tous ces standards restent très généralistes. C'est évident pour la DCMI, cependant, c'est le cas aussi pour les autres. Ils se veulent capable de décrire des ressources appartenant à n'importe quel domaine de l'enseignement : dans le LOM un objet pédagogique est « *N'importe quelle entité, numérisée ou non, qui pourrait être utilisée pour l'apprentissage, l'enseignement ou la formation* » (LOM, 2002) (traduction de l'auteur)

LOM utilise des niveaux d'agrégation censés permettre à partir du même ensemble d'éléments de décrire aussi bien un texte qu'une formation entière. Le fait que ces standards soient si généralistes nuira fatalement à la précision de la description qu'elles proposent, comme le fait remarquer Jean-Philippe Pernin lorsqu'il cite parmi les imprécisions ou ambiguïtés dont souffre le LOM : « *la volonté d'intégrer au sein d'un même modèle des entités de niveau conceptuellement très différent* » (Pernin, 2004). Notre problème est extrêmement spécifique : la description que l'on pourra faire d'un texte dans le cadre de l'enseignement des langues, ne sera pas nécessairement applicable à la description d'un texte dans une autre matière et ne le sera assurément pas pour décrire un problème de mathématique ou un cursus entier.

Les standards présentés sont donc trop généralistes pour être appliqués tels quels dans notre projet, mais elles présentent toutes la possibilité d'être adaptées à d'autres problématiques. La conformité au LOM s'exprime en les termes suivants (LOM, 2002) :

- il n'est pas obligatoire de renseigner tous les éléments LOM
- si l'on n'étend pas le standard l'instance sera strictement conforme
- sinon elle sera conforme à condition qu'aucun des éléments ajoutés ne remplace un élément LOM

On pourra donc tout à fait réutiliser le LOM, on sortira du cadre de la stricte conformité mais cela ne l'exclut pas pour autant. EdNA Metadata et GEM, constituant des extensions du DCES, ils pourront eux aussi être raffinés en respectant les principes de la grammaire DCMI.

Que ce soit dans LOM, le GEM ou EdNA metadata, les éléments « pédagogiques » sont séparés des autres éléments. Le GEM et EdNA suivent les directives DCMI et utilisent donc le DCES pour « *les informations descriptives fondamentales* ». Dans LOM, les objets pédagogiques sont décrits en 77 éléments répartis en 9 catégories dont l'une et dédiée à la composante pédagogique de la description (« Educational »).

2.3 Les éléments pédagogiques

Le GEM ajoute cinq éléments dits pédagogique (Audience, Duration, Essential Ressources, Instructional Method et Standards) (GEM, 2004), EdNa Metadata trois (Audience, Category Code et Review) (EdNA, 2002) et enfin le LOM dispose de onze éléments de ce type (5.1. Interactivity Type, 5.2. Learning Resource Type, 5.3. Interactvity Level, 5.4. Semantic Density, 5.5. Intended End User Role, 5.6. Context, 5.7. Typical Age Range, 5.8. Difficulty, 5.9. Typical Learning Time, 5.10. Description, 5.11. Language).

Quelle que soit le standard utilisé, chacun des champs concerne une propriété jugée intrinsèque à l'objet pédagogique. Pour le GEM, EdNA Metadata et LOM, un texte brut entre

dans le cadre de l'utilisation du standard. Pourtant les propriétés considérées comme intrinsèques à un objet pédagogique ne le sont pas forcément pour un texte brut.

Prenons l'exemple des éléments « *Audience* » de GEM et EdNA, qui pourraient correspondre à l'élément 5.5 ou au couple d'éléments 5.5 et 5.7 de LOM. Il est ressorti de l'étude préliminaire évoquée précédemment que selon le type d'activité, un texte donné pouvait être utilisé avec des publics très variés. Pour une activité de compréhension, il est moins important que les apprenants connaissent parfaitement les structures et le vocabulaire employés que dans un exercice de grammaire. Le texte ne peut donc pas être considéré comme intrinsèquement destiné à un public puisqu'en fonction du type d'activité, le public pourra être différent. De même les éléments « *Duration* » du GEM et 5.9 de LOM, que l'on peut considérer comme équivalents, n'auront aucun sens pour un texte brut. Un texte donné pouvant servir de support à un exercice de compréhension ou à l'introduction d'une nouvelle notion grammaticale sera utilisé beaucoup plus longtemps que le même texte servant uniquement de base pour un exercice lacunaire concernant les déterminants pour un public plus avancé. Nous terminerons ce tour d'horizon non exhaustif des champs pédagogiques en faisant remarquer qu'un champ comme l'élément 5.3 de LOM n'est pas seulement inadapté à notre problème, il est en outre complexe à renseigner puisque les valeurs acceptées sont à choisir parmi « *very low, low, medium, high et very high* » sans que le standard ne fournisse de réelles directives pour leur utilisation.

3 Conclusion

Mêle si ces différents standards, ne sont pas utilisables tels quels pour notre projet, il n'est pas exclu de réaliser un profil d'application de métadonnées : « *un assemblage d'éléments de métadonnées choisis à partir d'un ou plusieurs schémas de métadonnées et combinés dans un schéma composite* ». (Duval et al., 2002) (traduction de l'auteur) En effet la plupart des éléments non pédagogiques seront réutilisables. De plus, la création d'un profil d'application de métadonnées peut permettre d'améliorer les ensembles de valeurs proposés pour certains éléments, en fournissant des lignes directrices très précises quant à l'utilisation d'une norme.

Les remarques que nous avons pu faire sur le caractère intrinsèque ou non de certaines propriétés décrites dans les éléments pédagogiques nous ont amené à préciser l'architecture probable du système de gestion de la base de texte. Le fait que les éléments ne correspondent pas à des propriétés intrinsèques des textes ne signifie pas que certaines informations qu'ils pourraient contenir ne s'avéreront pas intéressantes pour les enseignants. La description des textes devra, pour être cohérente, contenir exclusivement des informations valables quelle que soit l'utilisation du texte. Nous nous proposons donc de séparer le processus en deux parties. Dans un premier temps on décrira les textes en fonction de traits pertinents pour les enseignants, mais qui ne varieront pas selon l'utilisation qui sera faite du texte. Les caractéristiques lexicales, syntaxiques (vocabulaire, structures grammaticales) du texte, par exemple sont des caractéristiques intrinsèques au texte, qui pourront dans une certaine mesure être relevées automatiquement grâce à l'utilisation d'outils TAL comme un lemmatiseur ou un analyseur morphologique. Mais les propriétés intrinsèques aux textes ne seront probablement pas suffisantes pour répondre à toutes les questions des utilisateurs de la base. Nous aurons donc recours, à une couche logicielle supplémentaire, un moteur d'inférence qui permettra de considérer plusieurs facettes de chaque texte. Le moteur devra analyser les caractéristiques de chaque texte en les combinant avec des informations fournies par

l'utilisateur dans sa requête afin d'en déduire certaines qualités d'un texte en fonction de l'utilisation qui en sera faite.

Il nous est actuellement impossible de faire la liste des outils TAL nécessaires à la réalisation de cette base. Les deux exemples ci-dessus (lemmatiseur / analyseur morphologique) paraissent correspondre à la demande des enseignants[2]. Cependant cela devra être confirmé par une analyse plus poussée des besoins des enseignants, via un questionnaire moins ouvert. En précisant les besoins, nous serons en mesure de dire si oui ou non ces outils peuvent répondre à la demande des enseignants. En outre, les résultats obtenus avec ces types d'outils sont suffisamment fiables pour que ces derniers puissent être employés. Car si l'utilité de l'outil dans la perspective de la description du texte pour l'enseignement des langues est une valeur primordiale pour décider de son utilisation, elle n'est pas la seule, une fiabilité minimum sera à définir plus tard afin que la précision des requêtes effectuées sur la base ne soit pas trop faible. Enfin, nous devrons probablement adopter une architecture modulaire pour le système, afin de pouvoir intégrer plus tard de nouveaux outils à la base.

Références

Antoniadis, G., Echinard, S., Kraif, O., Lebarbé, T., Ponton, C. (2004), Modélisation de l'intégration de ressources TAL pour l'apprentissage des langues : la plateforme MIRTO 57-70 actes de la journée TAL et Apprentissage des langues du 22 Octobre 2004, http://www.u-grenoble3.fr/lidilem/talal/actes/JourneeTALAL-041022-antoniadis.pdf.

DCMI Usage Board (10 janvier 2005). DCMI Metadata Terms. Consulté en janvier 2005 http://dublincore.org/documents/dcmi-terms/

Duval E., Hodgins W., Sutton S., Weibel S. L. (2002) Metadata Principles and Practicalities. D-Lib Magazine, 8 (4). Consulté en avril 2005. http://www.dlib.org/dlib/april02/weibel/04weibel.html

EdNA (septembre 2002), EdNA Metadata Standard V1.1. Consulté en septembre 2004 http://www.edna.edu.au/edna/go/pid/385

GEM (1er Juin 2004), GEM Top-Level Elements. Consulté en janvier 2005 http://www.thegateway.org/about/documentation/metadataElements/

Hillman D. (26 août 2003), Using Dublin Core Consulté en septembre 2004 http://dublincore.org/documents/usageguide/

LOM (juillet 2002), Final 1484.12.1 LOM Draft Standard Document. Consulté en avril 2005 http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf

Pernin J.P. (2004), A propos des objets pédagogiques, in "Entre technique et pédagogie : la création de contenus multimédia pour l'enseignement et la formation", Neuchâtel : IRDP.

[2] Données provenant d'un premier questionnaire, volontairement très ouvert de manière à laisser les enseignants s'exprimer sans qu'on leur impose un point de vue a priori, rempli par 130 enseignants.

Une méthode pour la classification de signal de parole sur la caractéristique de nasalisation

Luquet Pierre-Sylvain
GREYC - CNRS UMR 6072 - Université de Caen
Bd Maréchal Juin - F14032 Caen Cedex
psluquet@info.unicaen.fr
Date de soutenance prévue : décembre 2005

Mots-clefs – Keywords

Phonologie, phonétique, classifieur, réseaux de neurones
Phonology, phonetic, classifier, neural nets

Résumé - Abstract

Nous exposons ici une méthode permettant d'étudier la nature d'un signal de parole dans le temps. Plus précisément, nous nous intéressons à la caractéristique de nasalisation du signal. Ainsi nous cherchons à savoir si à un instant *t* le signal est nasalisé ou oralisé. Nous procédons par classification à l'aide d'un réseau de neurones type perceptron multi-couches, après une phase d'apprentissage supervisée. La classification, après segmentation du signal en fenêtres, nous permet d'associer à chaque fenêtre de signal une étiquette renseignant sur la nature du signal.

In this paper we expose a method that allows the study of the phonetic features of a speech signal through time. More specifically, we focus on the nasal features of the signal. We try to consider the signal as [+nasal] or [-nasal] at any given time. We proceed with a classifier system based on a multilayer perceptron neural net. The classifier is trained on a hand tagged corpus. The signal is tokenized into $30ms$ hamming windows. The classification process lets us tag each window with information concerning the properties of its content.

1 Appuis théoriques

La reconnaissance de la parole a, grâce aux techniques Markoviennes, fait un bond qualitatif énorme ces dernières années. Les décodeurs acoustiques, tels que ceux développés au LIMSI (Lamel & Gauvain, 1993), atteignent des taux de reconnaissance proches des 75%. Cependant, les limitations restent nombreuses et la critique la plus largement formulée vis-à-vis de ce type de système est la quasi absence de connaissances sur le langage dans les modèles sous-jacents (Plaut & Kello, 1999). Les travaux actuels s'articulent autour de deux axes. Le premier s'intéresse à l'amélioration des techniques de description du signal (Chetouani *et al.*, 2002). Le second est orienté vers la production : acquisition de connaissances concernant les gestes articulatoires des locuteurs (Vaxelaire *et al.*, 2002), leurs influences sur le signal (Montagu, 2004), et les processus cognitifs mis en jeu (Hawkins, 2003). Ces connaissances font l'objet de différentes études visant à leur intégration dans les systèmes de reconnaissance automatique de la parole (Wrench & Richmond, 2000).

Nous décrivons dans ces lignes une approche sensiblement différente. Nous cherchons à appuyer une technique de décodage acoustico-phonétique sur la *notion de différence*. Saussure affirme que « *dans la langue, il n'y a que des différences [...] sans termes positifs* » (Saussure, 1986). Coursil reprend à son compte cette notion dans (Coursil, 1992)[1] et affirme à son tour « *Pour tout phonème* x, *il existe un phonème* y *tel que* y = x *à une et une seule différence catégorique près* ». C'est à partir de cette dernière affirmation qu'il construit la *topique des phonèmes du français contemporain*[2]. Le but de la classification est de mettre en évidence cette différence dans le signal. Notons enfin que la classification automatique d'un signal de parole suivant un trait phonétique donné suppose que le phonème est une *substance*, hypothèse validée par la « Dispersion-Focalisation Theory » publiée dans (Schwartz *et al.*, 1997).

Le trait de nasalité. On distingue dans le français contemporain les phonèmes nasaux des phonèmes oraux, le tableau 1 en présente la partition[3]. Du point de vue de la mécanique articulatoire, la nasalité est décrite comme une connexion du conduit vocal avec le conduit nasal par le biais de l'abaissement du vélum. Les répercussions acoustiques de ce phénomène, sont décrites par Jakobson dans (Jakobson, 1980) en s'appuyant sur Fant et Delattre. Pour Fant, les consonnes nasales sont « caractérisées par un spectre où F2 est faible ou bien absent »[4] ; Delattre affine la description en précisant que pour les voyelles nasales, comparées aux orales, F1 perd une bonne part de son intensité au profit de F2. Plus récemment, Feng et Kotenkoff (Feng & Kotenkoff, 2004) ont mené à l'ICP[5] des observations basées sur une technique d'enregistrement du locuteur en différenciant les prises de son en provenance du conduit vocal et du conduit nasal. Ils ont constaté que l'abaissement du vélum a deux effets distincts : pour le conduit vocal le rétrécissement engendre le rapprochement des formants F3 et F4, et pour le conduit nasal sa connexion entraîne un rayonnement au niveau des narines caractérisé par une concentration dans les basses fréquences et aux alentours de 3000 Hz.

[1] Les travaux sur la phonologie de Coursil s'inscrivent dans un projet global dénommé *ANADIA*. On lira dans (Mauger, 1999) l'une des extensions de ce projet.

[2] Le format dans lequel cet article est accepté ne me permet pas d'expliquer plus avant cette notion de topique. Néanmoins, je mets à disposition de tout lecteur en faisant la demande une version étendue décrivant plus finement celle-ci.

[3] La notation employée ici pour désigner les phonèmes est le codage SAMPA (Speech Assessment Methods Phonetic Alphabet). Pour plus d'informations se reporter au site de l'UCL : `http://www.phon.ucl.ac.uk/home/sampa/home.htm`

[4] F1 et F2 désignent respectivement le premier et le second formants. Les formants sont des fréquences de résonance maximum de l'enveloppe spectrale du signal de la parole à un instant donné.

[5] Institut de la Communication Parlée - Grenoble

	nasal	oral
voyelles	/e∼, o∼, 9∼, a∼/	/u, y, i, E, e, O, o, 9, 2, @, a, A/
consonnes	/m, n, J/	/p, f, v, b, d, t, k, g, z, s, S, Z, R, l, w, H, j/

TAB. 1 – Nasal vs. oral

2 Corpus

2.1 Constitution

Nous faisons ici l'hypothèse que le corpus présentant le moins de difficultés pour réaliser la partition oralisé *vs.* nasalisé est constitué de paires minimales oral *vs.* nasal. De plus, nous nous sommes concentrés sur les phonèmes dont la production pouvait être maintenue. Nous avons retenu dans notre corpus les quatre paires oppositives suivantes : /o∼/ - /o/, /e∼/ - /E/, /9∼/ - /9/ et /a∼/ - /A/. Ces phonèmes sont associés aux mots prototypiques du tableau 2.

Phonèmes	/o∼/	/o/	/e∼/	/E/	/9∼/	/9/	/a∼/	/A/
Prototype	tronc	trot	bain	baie	un	neuf	pente	pâte

TAB. 2 – Phonèmes et mots prototypes

Nous disposons aujourd'hui des résultats sur 3 corpus[6] de test monolocuteur (voir le tableau 3). Le premier corpus, $C1$ est constitué d'une seule paire minimale (/o∼/ & /o/), dont la seule variation est le trait de nasalisation (N). Les corpus $C2$ et $C3$ sont plus complexes : sur les 7 caractéristiques mises en jeu 5 varient. Pour les orales (/o/ & /a/) les variations portent sur la laxité (L), la compacité (C) et la bémolisation (B) ; la hauteur (H) intervient en plus pour les nasales (/o∼/ & /a∼/)[7].

	Phonèmes	Nb. Phonèmes	Nb. fenêtres	Maintenus	Variations
$C1$	/o∼/-/o/	22	2250	oui	N
$C2$	/o∼, a∼/-/o, a/	20	2000	oui	N, L, C, B, H
$C3$	/o∼, a∼/-/o, a/	24	470	non	N, L, C, B, H

TAB. 3 – Corpus

2.2 Paramètres

L'outil principalement utilisé dans cette expérience est le logiciel d'analyse acoustique PRAAT[8]. **Corpus.** Nous travaillons en utilisant la technique classique de fenêtrage du signal. Chaque signal de parole à analyser est segmenté en tranches de $30ms$ avec un décalage de $10ms$. A chaque

[6]Pour chacun de ces phonèmes, il a été demandé au locuteur de le prononcer dans un mot prototypique, puis de le répéter de façon isolée. Cette façon de procéder permet à l'utilisateur de « calibrer » le phonème qu'il doit ensuite prononcer isolément. Nous demandons au locuteur de répéter une dizaine de fois ce processus par couple prototype/phonème.

[7]Les quantités du tableau 3 (nombre de phonèmes et nombre de fenêtres) données ici correspondent pour chaque corpus aux versions de test. Les versions d'apprentissage sont du même ordre de grandeur.

[8]Pour plus d'informations voir la page web `http://www.fon.hum.uva.nl/praat/`

fenêtre est appliquée une fonction de $Hamming$. Chaque tranche de signal fenêtré constitue un $vecteur$ dont le nombre d'éléments est dépendant de la fréquence d'échantillonnage du signal (662 échantillons par tranche pour du signal échantillonné à $22kHz$). Chaque vecteur est labellisé par sa caractéristique acoustique ("nasal" ou "oral"). Ces vecteurs sont concaténés en matrices, qui selon le corpus servira soit à l'apprentissage, soit à la phase de test[9].

Classifieur. Nous utilisons des réseaux de neurones type perceptron à une couche cachée. L'entrée du réseau comporte autant de cellules que nous avons de valeurs par vecteur de signal, soit 662 cellules. La sortie est composée de deux cellules correspondant aux classes activables. La couche cachée est composée de 331 cellules. Lors des phases d'apprentissage l'évaluation de l'erreur est calculée suivant la méthode *minimum squared error*.

3 Résultats

3.1 Variation restreinte

Les résultats du tableau 3 concernent le corpus $C1$ et ont été obtenus au terme d'un apprentissage de 400 cycles. Les phonèmes sont maintenus et la seule variation phonologique mise en jeu est la nasalisation. Nous voyons ici que sur une quantité restreinte de corpus il est possible de classifier le signal avec de bons résultats. En effet nous obtenons un taux d'erreurs faible ($2,8\%$), mais nous voyons surtout que le nombre de fenêtres continues incorrectement classifiées est très faible (8) en regard du nombre de fenêtres par phonème (102). Le risque de mal classifier un phonème est donc minime.

fenêtres	2252	fenêtres par phonème	102
erreurs	63	groupes d'erreur	17
taux	**2,8%**	erreurs par groupe	3,71
erreurs consécutives maximum			**8**

TAB. 4 – Résultats $C1$

3.2 Augmentation de la dissemblance

Les résultats donnés ici concernent le corpus $C2$. Le tableau 5 donne les résultats obtenus pour 400 cycles d'apprentissage, tandis que le tableau 6 nous donne les résultats au bout de 600 cycles. Comme précédemment les phonèmes sont maintenus mais plusieurs variations phonologiques sont ici mises en jeu (voire 2.1). La focalisation du classifieur sur la caractéristique de nasalisation est donc rendue plus complexe en raison du bruit apporté par les autres variations. Cependant, les résultats obtenus montrent qu'une classification est toujours possible. Avec 400 cycles (tableau 5), nous obtenons un taux d'erreurs qui reste faible ($5,3\%$). Le nombre de fenêtres contiguës incorrectement classifiées l'est aussi (12 fenêtres mal classifiées). Néanmoins, si nous augmentons d'un tiers le nombre de cycles (tableau 6), le taux d'erreurs retombe à $2,3\%$.

[9]Dans les deux cas, les valeurs des échantillons sont décalées et mises à l'échelle pour être dans le domaine de définition de notre classifieur. Les valeurs d'origine varient dans l'intervalle $[-1, 1]$. Nous les réduisons d'un facteur $1/2$ puis les décalons de 1 pour qu'elles soient comprises dans l'intervalle d'entrée du classifieur : $[0, 1]$.

fenêtres	1996	fenêtres par phonème	100
erreurs	106	groupes d'erreur	47
taux	**5,3%**	erreurs par groupe	2,3
erreurs consécutives maximum			**12**

TAB. 5 – Résultats $C2$ - Apprentissage : 400 cycles

fenêtres	1996	fenêtres par phonème	100
erreurs	47	groupes d'erreur	16
taux	**2,3%**	erreurs par groupe	2,9
erreurs consécutives maximum			**9**

TAB. 6 – Résultats $C2$ - Apprentissage : 600 cycles

3.3 Phonèmes non maintenus

L'expérience menée sur le corpus $C3$ est similaire à l'expérience précédente, mais concerne des phonèmes non maintenus. Les résultats obtenus (tableau 7 et 8) sont nettement en retrait, mais restent néanmoins très intéressants. Au terme d'un apprentissage de 300 cycles, nous observons un taux d'erreur de 20% que nous pouvons réduire à $15,8\%$ au terme de 600 cycles d'apprentissage (soit une réduction de ce taux de $21,5\%$). En revanche, le doublement du nombre de cycles d'apprentissage n'apporte rien ici en terme de réduction du nombre d'erreurs contiguës (7 fenêtres mal classifiées[10]). Néanmoins ce nombre reste acceptable, dans le cas d'une stratégie de classification *winner-takes-all* dans la mesure où un phonème compte en moyenne 20 fenêtres. Notons que la taille de notre corpus d'apprentissage (412 fenêtres) pose ici un problème ; le nombre de patrons étiquetés limite la capacité de classification. Enfin, un dernier cycle long d'apprentissage (2000 cycles) ne nous a pas permis d'améliorer sensiblement le taux d'erreurs et a également confirmé qu'au delà de 600 cycles, la réduction de l'erreur est faible pour un coût très élevé ; dans notre cas le nombre de cycles a été plus que triplé pour un gain de 2 erreurs seulement sur le corpus de test.

fenêtres	469	fenêtres par phonème	20
erreurs	94	groupes d'erreur	37
taux	**20,0%**	erreurs par groupe	2,5
erreurs consécutives maximum			**7**

TAB. 7 – Résultats $C3$ - Apprentissage : 300 cycles

4 Perspectives et conclusion

Les résultats présentés dans cet article sont prometteurs, cependant certains aspects sont à approfondir. D'autre types de descripteurs sont envisagés : techniques d'extraction de type MFCC (Mel Frequency Cepstral Coefficients), LPC (Linear Predictive Coding) ou plus encore PLP (Perceptual Linear Predictive coding). Par ailleurs, la limite en terme de fréquence d'échantillonnage en deçà de laquelle l'apprentissage n'est plus réalisable n'est pas connue. Qu'en

[10]Le nombre donné ici correspond au nombre maximal de fenêtres contiguës mal classifiées dans un phonème.

fenêtres	469	fenêtres par phonème	20
erreurs	74	groupes d'erreur	30
taux	**15,8%**	erreurs par groupe	2, 5
erreurs consécutives maximum			**7**

TAB. 8 – Résultats $C3$ - Apprentissage : 600 cycles

est-il d'un signal de qualité téléphonique échantillonné à $8kHz$?
Nous envisageons également d'augmenter la complexité du corpus : nombre de locuteurs et nombre de phonèmes présents. L'augmentation du nombre de locuteurs a pour but de tester l'indépendance de l'apprentissage du classifieur. Pour valider notre méthode sur du signal de parole continue, une nouvelle série d'expériences est envisagée. L'augmentation du nombre de phonèmes doit permettre de multiplier les caractéristiques prises en considération.
En outre, nous faisons l'hypothèse que le croisement de résultats issus de plusieurs classifieurs (avec un apprentissage sur des catégories phonétiques différentes) permettra de situer le signal dans l'espace topique et de déterminer ainsi la classe phonétique à laquelle il appartient.

Références

CHETOUANI M., GAS B. & ZARADER J. (2002). Coopération entre codeurs neuro-prédictifs pour l'extraction de caractéristiques en reconnaissance de phonèmes. In *Reconnaissances des formes et intelligence artificielle*.

COURSIL J. (1992). *Essai d'intelligence artificielle et de linguistique générale*. PhD thesis, Université de Caen.

FENG & KOTENKOFF (2004). Vers un nouveau modèle acoustique des nasales basé sur l'enregistrement bouche - nez séparé. In *Journées d'Etude sur la Parole*.

HAWKINS S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*.

JAKOBSON R. (1980). *La charpente phonique du langage*. Paris : Editions de Minuit.

LAMEL L. & GAUVAIN J. (1993). High performance speaker-independent phone recognition using cdhmm. In *European Conference on Speech Communication and Technology*.

MAUGER S. (1999). *L'Interprétation des Messages Énigmatiques. Essai de Sémantique et de Traitement Automatique des Langues*. PhD thesis, Université de Caen.

MONTAGU J. (2004). Les sons sous-jacents aux voyelles nasales en français parisien : indices perceptifs des changements. In *Journées d'Étude sur la Parole*, p. 385–388.

PLAUT D. C. & KELLO C. T. (1999). *The Emergence of Language*, chapter The Emergence of Phonology from the Interplay of Speech Comprehension and Production : A Distributed Connectionist. Lawrence Erlbaum Assoc : Mahwah.

SAUSSURE F. (1986). *Cours de linguistique générale*. Paris : Mauro Payot.

SCHWARTZ J.-L., BOË L.-J., VALLÉE N. & ABRY C. (1997). The dispersion-focalization theory of vowel systems. *Journal of Phonetics*.

VAXELAIRE B., FERBACH-HECKER V. & SOCK R. (2002). La perception auditive de gestes vocaliques anticipatoires. In *Journées d'Etude sur la Parole*.

WRENCH A. A. & RICHMOND K. (2000). Continuous speech recognition using articulatory data. In *International Conference on Spoken Language Processing*.

De la linguistique aux statistiques pour indexer des documents dans un référentiel métier

Wilfried Njomgue Sado (1,2), Dominique Fontaine (1)

(1) UMR CNRS 6599 Heudiasyc, Université Technologie de Compiègne BP 20529, F-60205 Compiègne
{wilfried.njomgue-sado, dominique.fontaine}@hds.utc.fr
(2) Suez Environnement CIRSEE Pôle Informatique Métier
38, rue du Président Wilson, F-78230 Le Pecq

Mots-clefs – Keywords

Linguistique, indexation, recherche d'information, statistique

Linguistics, statistics, indexing, information processing

Résumé – Abstract

Cet article présente une méthode d'indexation automatique de documents basée sur une approche linguistique et statistique. Cette dernière est une combinaison séquentielle de l'analyse linguistique du document à indexer par l'extraction des termes significatifs du document et de l'analyse statistique par la décomposition en valeurs singulières des mots composant le document. La pondération des termes tire avantage de leur contexte local, par rapport au document, global, par rapport à la base de données, et de leur position par rapport aux autres termes, les co-occurrences. Le système d'indexation présenté fait des propositions d'affectations du document à un référentiel métier dont les thèmes sont prédéfinis. Nous présentons les résultats de l'expérimentation de ce système menée sur un corpus des pôles métiers de la société Suez-Environnement.

This article presents an automatic method of documents indexing based on a hybrid, linguistic statistical approach. The proposed approach combines a linguistic analysis of the document by the extraction of the significant terms of the document in conformity with the referential; and a statistical analysis of the same document decomposed into separed words. Innovating weighting of terms is set to take judiciously advantage of both their position with respect to other terms (co-occurrence) and their local and global context. An application was developed in order to assign referential-based topics to documents. Finally, we will present experiments results and evaluation carried out on documents of Suez-Environnement Company.

1 Introduction

Les opérations de stockage et la diffusion des documents sur différents supports exigent au préalable une indexation qui consiste à réduire le contenu sémantique de chaque document. La Direction Technique et de Recherche de Suez-Environnement a initié un projet de gestion des connaissances dont l'objectif principal est de concevoir un outil qui permette à tout utilisateur d'introduire de nouveaux documents dans la base de données de l'Intranet du groupe. La particularité de ce projet réside dans l'existence d'un référentiel métier qui a été élaboré il y a quelques années et est constamment mis à jour. Il s'agit d'une taxonomie qui décrit l'ensemble des activités et métiers de l'entreprise. Avant de mettre un document dans la base de donnée, son auteur doit identifier au mieux le sujet du document, en fonction des activités qui en caractérisent la sémantique. Cette tâche d'indexation s'avère fort fastidieuse : en effet, l'auteur est d'abord censé connaître la plupart des activités de l'entreprise, hypothèse fort risquée, puis élaborer sa propre représentation du document, et enfin choisir certains métiers du référentiel parmi la multitude des possibilités. Il est donc essentiel de réduire le temps nécessaire à l'accomplissement de cette tâche, en l'automatisant partiellement ou totalement. La plupart des systèmes n'indexent pas de façon totalement autonome les textes numérisés, on parle alors d'indexation semi-automatique. Notre système propose une liste ordonnée d'affectations possibles du nouveau document à des métiers du référentiel afin que l'auteur puisse opter pour une ou plusieurs d'entre elles.

Le présent article a pour objectif principal de présenter un processus d'indexation qui permet d'analyser chaque document et de déterminer les affectations possibles en fonction des métiers du référentiel. Il présente d'abord la méthode d'indexation automatique, ensuite évalue la méthode sur une collection de documents, et conclut sur quelques perspectives.

2 Une méthode d'indexation automatique

L'indexation présentée ici a pour but d'extraire les concepts identifiant au mieux le document, puis de le rattacher aux métiers prédéfinis au sein du référentiel. L'indexation du document est faite relativement aux activités de l'entreprise, et non relativement aux mots du document. La contrainte supplémentaire est la suivante : nous ne sommes pas responsables de l'intégrité et de la pertinence de ce référentiel qui comporte des relations entre concepts dont la sémantique est pour le moins variable voire au pire indiscernable. En outre, il ne nous est pas permis de modifier cette structure. Nous sommes en mesure d'évaluer systématiquement les résultats produits par le système. En effet, nous les comparons à ceux fournis manuellement par l'auteur du document tout en faisant l'hypothèse que les propositions faites par l'auteur sont pertinentes et donc qu'elles ne sont pas à remettre en cause. Cette contrainte est extrêmement forte car la diversité des auteurs fait qu'ils n'ont pas toujours la même compréhension du référentiel, d'où la nécessité de concevoir un système semi-automatique.

Pour accomplir cette tâche, le processus d'indexation, considéré dans sa globalité, s'appuie à différents moments sur le référentiel, et comporte trois phases principales, où s'enchaînent successivement des traitements linguistiques, statistiques et sémantiques. Nous n'abordons dans cet article que les deux premières phases, la phase de traitement sémantique, basée sur l'exploitation d'une ontologie du domaine, étant encore en cours de finalisation.

3 Analyse et traitement linguistique

Il s'agit ici d'extraire automatiquement les termes composant un document (Séguéla, 2001 ; Bourigault, Jacquemin, 2000). Parmi la multitude d'outils ayant de très bonnes performances (Ana, Termino, Syntex, Intex, Acabit, etc.), Intex (Silberztein, 2001) a retenu notre attention car il permet d'intégrer des dictionnaires spécialisés, des grammaires de reconnaissance des syntagmes, répondant en particulier à nos besoins. Le traitement linguistique comprend alors séquentiellement des analyses morphologique, puis syntaxique, et sémantique, celui-ci se réduisant à un regroupement morphologique et/ou synonymique des termes clés). Dans le but de compléter l'analyse linguistique, nous avons inséré un dictionnaire spécialisé, et des grammaires locales afin de détecter, modifier ou réduire certaines abréviations afin de ne pas perdre l'information associée aux abréviations. Pour réduire la taille des mots non lemmatisés, nous avons construit des grammaires de reconnaissances de certains lemmes. Des grammaires de corrections de certains mots erronés ont également été mises sur pied afin de ne pas biaiser l'information du document.

Au terme de cette étape, nous obtenons deux fichiers texte : un fichier «taggé » F_{tag} , fichier où les lemmes non ambigus sont écrits entre accolades et un fichier F_{lemme} , fichier des termes lemmatisés et des occurrences associées. Afin d'affiner les lemmes ainsi obtenus, nous appliquons premièrement un « stemming » sur le fichier F_{lemme} pour obtenir le fichier que nous noterons $F_{Stemming}$. Le stemming est la réduction des formes de surfaces similaires à un seul concept, par exemple « inintéressant », « intéressant », « intérêt », « intéressé » seront réduits en « intérêt ». Ce « stemming » s'est basé sur la liste des mots du référentiel métier ayant préalablement subie un « stemming » manuel. L'objectif de cette méthode est de donner plus d'importance aux termes métiers du domaine. Ensuite, nous obtenons le fichier issu de la phase linguistique par application de la technique dite de « stop-list » sur le fichier $F_{Stemming}$ (Table 1). Celle-ci consiste souvent à dresser une liste de termes non soumis à l'indexation. Ici, plutôt que d'utiliser une liste prédéterminée, nous avons fait le choix d'écarter les termes dont le nombre de caractères est inférieur à une valeur fixée empiriquement (3 est la valeur optimale obtenue par expérience), considérant que ce sont des termes de poids sémantiquement faibles (« le », « la », etc.), et donc peu intéressants à être indexés.

Table 1. Résultats des applications des techniques de « stemming » et de « stop-list »

Fichier des lemmes F_{Lemme}	Fichier « stemming » $F_{Stemming}$	Fichier linguistique
{S} 3	{S} 3	
{activation} 1		
{activer} 1	**{activer} 2**	{activer} 2
{ainsi} 1	{ainsi} 1	
{aire} 1	{aire} 1	{aire} 1
{alimentation} 2	{alimentation} 2	{alimentation} 2
{analyser} 1		
{analyse} 2	**{analyse} 3**	{analyse} 3

4 Un traitement statistique

Nous présentons le processus statistique à travers successivement la méthode de pondération des termes existantes dans le document, la recherche des proximités entre les termes fondée sur la notion de co-occurrence, et enfin l'application de la méthode du « latent semantic indexing ».

4.1 Méthode de pondération

Pour mettre en valeur un terme par rapport à un autre, le système le pondère. Avant de procéder au choix de notre modèle de pondération, il nous a paru utile de faire une rapide synthèse des différentes méthodes existantes. De toutes les pondérations existantes, (Singhal et al., 1996) affirment que la pondération $P_{1\alpha}$ est la plus intéressante parmi celles ne prenant pas en compte la composante globale pour la recherche d'information. Pour celles qui prennent en compte cette composante, $P_{2\alpha}$ est intéressante d'après (Faraj et al., 1996). La composante globale est le facteur qui permet d'accorder un poids plus important aux termes discriminants qui apparaissent moins fréquemment dans la collection des documents.

Nous noterons par $P_{f\alpha}$ le poids du terme α du profil lexical P_{lex} ; $C_{i,j}$ le poids de co-occurrence du couple des termes (u_i, u_j) ; P_i le poids du terme u_i dans le document ; n_c le nombre de fois où les termes u_i, et u_j apparaissent ensemble ; n_i le nombre de fois où le terme u_i apparaît seul, N le nombre de documents dans la base de donnée ; tf_i l'occurrence du mot i dans un document ; df_j le nombre de documents dans lequel le terme j apparaît.

$$P_{1\alpha} = [1+\log(tf_\alpha)] \times \left[\frac{1}{\sqrt{\sum_{\alpha=1}^{n}[1+\log(tf_\alpha)]^2}}\right] \quad \text{et} \quad P_{2\alpha} = [1+\log(tf_\alpha)] \times \left[\log\left(\frac{N}{df_\alpha}\right)\right]$$

Nous définissons alors, à partir de ces deux pondérations, une méthode de pondération par étude de cas (Table 2). Cette pondération est normalisée dans [0,1]. On dira que la valeur de la pondération est respectivement faible, moyenne et forte si cette valeur est incluse dans [0 ;0.25],] 0.25 ; 0.75[, [0.75 ; 1]. Tous les termes du fichier linguistique sont soumis à cette méthode.

Table 2. Pondération définitive des termes. $\text{Max}(P_{1\alpha}, P_{2\alpha})$ et $\text{Min}(P_{1\alpha}, P_{2\alpha})$ désignent respectivement le maximum et le minimum entre les pondérations $P_{1\alpha}$ et $P_{2\alpha}$.

Pondération définitive	**Pondération $P_{2\alpha}$: elle met plus en avant la composante globale que ne le fait $P_{1\alpha}$**			
Pondération $P_{1\alpha}$: elle met plus en avant la normalisation que ne le fait $P_{2\alpha}$		Forte	Moyenne	Faible
	Forte	$\text{Max}(P_{1\alpha}, P_{2\alpha})$	$P_{1\alpha}$	$P_{1\alpha}$
	Moyenne	$P_{1\alpha}$	$P_{1\alpha}$	$P_{1\alpha}$
	Faible	$P_{2\alpha}$	$P_{1\alpha}$	$\text{Min}(P_{1\alpha}, P_{2\alpha})$

4.2 Approche matricielle : adaptation du « Latent Semantic indexing »

Plusieurs applications dans le domaine de la recherche d'information (RI), de la classification des documents, du filtrage d'information (Deerwester et al., 1990 ; Dumais et al., 1996) ont été développées selon l'approche matricielle du « Latent Semantic Indexing » (LSI) qui forunit de meilleurs résultats par rapport aux méthodes standards. Ici, on suppose qu'il y a une structure « latente », à caractère « sémantique », dans l'usage des mots d'un document qu'on révèlera par la décomposition en valeur singulière du LSI. Le plus souvent, la matrice (« unités lexicales » x « unités textuelles ») est le point de départ de cette méthode. Dans le cas présenté, nous utilisons respectivement les thèmes de ce référentiel et les termes du profil lexical pour définir les unités

textuelles et les unités lexicales et obtenir ainsi une matrice Termes-Thèmes notée $X_{Termes,Thèmes}=(X_{i,j})$. Ainsi, la sémantique d'un document est considérée comme une combinaison linéaire (Dumais et al., 1996) du contenu des thèmes du domaine ainsi que du sens des termes associés. Notre but étant d'affecter un document par son contenu au référentiel, il est pertinent de représenter un document en liaison avec les thèmes du domaine, un thème étant un chemin de l'arborescence qu'est le référentiel. Alors :

$$X_{i,j} = \begin{cases} P_i + \sum_{k=i+1}^{n} C_{i,k} & \text{si } Terme_i \subseteq Thème_j \\ 0 & \text{sinon} \end{cases} \quad \text{avec} \quad C_{i,j} = (P_i + P_j) \times Proxi(u_i, u_j)$$

$$Proxi(u_i, u_j) = \frac{n_c(P_i + P_j)}{n_i P_i + n_j P_j} \; avec \; n_c \leq \min(n_i, n_j)$$

$X_{i,j}$ est la contribution du terme i du document au thème j, relativement au document à indexer. Les colonnes de cette matrice représentent la distribution du sens de chaque thème pour le document. Tout document est une combinaison linéaire de la contribution sémantique des thèmes représentant le domaine. Après décomposition en valeurs singulières (Husbands et al., 1996), la matrice réduite contient seulement les premiers composants linéaires indépendants k de $X_{Termes\ -\ Thèmes}$ avec $\sigma_1 \geq ... \geq \sigma_k > 0$. Nous chiffrons numériquement chaque thème afin d'en extraire les plus représentatifs, en calculant la norme des colonnes de la matrice projetée des thèmes obtenue.

5 Expérimentations

Pour réaliser les expériences, nous disposons d'emblée de documents issus du groupe écrits en langage naturel libre : le français ; et des affectations au référentiel proposées par les auteurs de ces documents. Dans notre expérience, nous avons à notre disposition un ensemble de près de 450 documents. Il y a 165 thèmes possibles pour l'indexation d'un document. L'auteur fait un choix parmi les 15 thèmes proposés par le système semi-automatique. Il s'avère, dans cette étude, que la moyenne de mots utiles dans un document est d'environ 700 mots. Le système d'indexation étant semi-automatique, le silence, (i.e.) le fait que le système n'extraie pas un thème suggéré par l'auteur, nous préoccupe davantage que le bruit. Il nous est alors apparu judicieux d'évaluer le système (Table 3) en terme de rappel plutôt que de précision.

Table 3. Résultats de l'expérience : rappel

	Mots importants	Thèmes proposés par l'auteur	Rappel (système)
Minimum	35	1	0%
Maximum	1537	10	100%
Moyenne	700	4	77.09%

Le *rappel* est le nombre de documents pertinents retournés par rapport au nombre total de documents pertinents.

A ce stade, ces résultats sont plutôt satisfaisants (77,09% de rappel) eu égard à la difficulté de la tâche : un référentiel assez hétérogène et imposé, une grande diversité de documents, et un jeu de tests dont la pertinence n'est pas toujours avérée mais auquel il faut se conformer. Bien sûr dans l'absolu, quelques problèmes de silence demeurent, notamment dans les cas où aucun terme du

thème n'apparaît dans le document, où le document est trop technique (présence de vidéo au détriment du texte), et enfin où l'information au sein du document est implicite.

6 Conclusion et perspectives

Le bien fondé de l'approche mixte à savoir linguistique puis statistique est confirmé à travers cette évaluation. D'abord, l'utilisation des grammaires, des techniques de lemmatisation, de stemming, de stop-list permet de réduire le document à indexer à un ensemble de mots jugés intéressants. Ensuite, un traitement statistique discrimine ces mots en raison de leurs occurrences et de leurs co-occurrences, puis estime la proximité entre le document et les thèmes du référentiel par le biais du LSI.

Cette évaluation révèle quelques insuffisances en terme de précision et de silence. Ces insuffisances ont été signalées dès le début du projet et nous nous proposons de les résorber certes par des améliorations au traitement effectué, mais surtout par un traitement sémantique. A cet effet, nous mettons en place une ontologie du domaine de l'eau, spécialisée de façon à répondre à nos besoins en matière d'indexation. Les experts et auteurs de l'entreprise sont actuellement sollicités dans cette phase de construction et au-delà d'exploitation de l'ontologie.

Références

BOURIGAULT D., JACQUEMIN C. (2000): « Construction des ressources terminologiques, In Ingénierie des langues », pp 215-230, 2000, ed. J.M. Pierrel. Hermes Sciences

DEERWESTER S., DUMAIS S., FURNAS G., LANDAUER T., HARSHMAN R. (1990): « Indexing by Latent Semantic Analysis », *Journal of Society for Information Science*, Vol.41, n.6, pp. 391-407, 1990

DUMAIS S., LETSCHE T., LITTMAN M., LANDAUER T. (1996): « Automatic Cross-Language Retrieval using Latent Semantic Indexing », *SigIR Multilingual IR Workshop*, Aug. 22, 1996

FARAJ N., GODIN R., MISSAOUI R., DAVID S., PLANTE P. (1996): « Analyse d'une méthode d'indexation automatique basée sur une analyse syntaxique de texte », *Canadian Journal of Information and Library Science / Revue l'information et de bibliothéconomie*, 1996

HUSBANDS P., SIMON H., DING H. (1996): « On the use of Singular Value Decomposition for Text Retrieval », *Proceeding's of SIAM Comp. Information Retrieval Workshop, 2000.*

SEGUELA P. (2001): Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques, 2001, Université Toulouse III, France

SINGHAL A., SALTON G., BUCKLEY C. (1996): « Length normalization in degraded text collections », *Proceedings of Fifth Annual Symposium on Document Analysis and Information Retrieval*, April 15-17, 1996, pp. 149-162.

SILBERZTEIN,M. (2001), *Intex @ manual* , 2000-2001. ASSTRIL - LADL, 201p

RECITAL 2005, Dourdan, 6-10 juin 2005

Vers un Système d'Ecriture Informatique Amazighe : Méthodes et Développements

Ali Rachidi[1] & Driss Mammass[2]

[1] Ecole Nationale de Commerce et de Gestion, B.P. 37/S Hay Salam Agadir, Maroc
Laboratoire de Traitement d'Images et Systèmes d'Information (LTISI)
Email : rachidi.ali@caramail.com
[2] Unité de Recherche en Traitement d'Images et de l'Information (URT2I), Faculté des Sciences, Université Ibn Zohr, Agadir, Maroc,
E-mail : driss_mammass@yahoo.fr

Mots-Clés – Keywords

Amazighe, Alphabet Tifinaghe, Traitement des langues naturelles.

Amazighe, Tifinaghe Alphabet, Natural Language Processing.

Résumé – Abstract

L'intégration des technologies de l'information et de communication (TIC) à l'apprentissage de la langue Amazighe est absolument nécessaire pour qu'elle ait droit de cité plein et entier sur le Web et dans le monde informatisé.
Nous présentons quelques réflexions sur les stratégies et méthodes d'informatisation de l'amazighe qui est une langue peu dotée informatiquement. Ces réflexions visent surtout l'optimisation de l'effort d'informatisation. En effet, les méthodes proposées tiennent en compte non seulement l'alphabet proposé par l'IRCAM[1] et confirmée par l'ISO (format Unicode) le 21 juin 2004 (IRCAM, 2004 a) mais aussi le contexte francophone des populations berbères.

Learning the Amazighe language would require the introduction of Information and Communication technologies so that the language would have full entire freedom of a mentioning on the Web and in the computerized world.
We present some reflections on strategies and methods of the amazighe computerization which is a language little dowered. These reflections aim particularly the optimisation of the effort of computerization. Indeed, proposed methods hold in account no only the proposed alphabet by the IRCAM[1] and confirmed by the ISO (Unicode format) the 21 June 2004 but therefore the French speaking context of Berber populations.

[1] Institut Royal de la Culture Amazighe, Rabat, Maroc

1. Introduction

Cet article s'inscrit dans un large mouvement international qui vise à ce que chaque peuple puisse disposer de tous les moyens pour communiquer dans sa langue. L'Amazighe fait partie des langues très peu dotées informatiquement. Par conséquent, des recherches scientifiques et linguistiques sont lancées dans ce sens pour améliorer la situation actuelle. La conception et la réalisation d'applications capables de traiter de façon automatique des données linguistiques, exprimées dans la langue naturelle amazighe, sont parmi les objectifs prioritaires de nos travaux de recherche en collaboration avec l'IRCAM. Dans ce contexte, nous proposons quelques réflexions, sur des méthodes et des stratégies à mettre en œuvre pour produire un outil de 1) traitement de texte amazighe sous codage ASCII et un autre sous format Unicode après l'intégration du format Unicode amazighe dans les applications informatiques par les firmes responsables et 2) traduction automatique et de gestion d'une base lexicale Amazighe. Cet article est composé de trois parties. Dans la première partie, nous présentons le contexte linguistique et le système d'écriture de la langue Amazighe. La deuxième partie présente les différentes méthodes d'informatisation. Enfin, la dernière partie est consacrée à la présentation de la mise en œuvre de certaines méthodes et des outils impliqués.

2. Amazighe : langue naturelle

La famille élargie des Berbères qui utilise le Tifinaghe comme écriture traditionnelle et commune est de près de vingt millions. Au Maroc, le berbère (« amazighe ») marocain englobe les trois grandes variantes : le *Tarifit*e, le *Tamazighte* et le *Tachelhit*e. Plus de 40% de la population marocaine est berbérophone. Nous présentons un petit aperçu sur l'évolution de la langue amazighe tout au long de l'histoire et sur son système d'écriture.

2. 1. Historique

L'alphabet berbère ou Amazighe a subi des modifications et des variations depuis son origine jusqu'à nos jours. Du libyque jusqu'au néotifinaghe en passant par le tifinaghe saharien et le tifinaghe touareg. Nous retraçons ci-dessous les aspects les plus importants de chacune de ces étapes.

2. 1. 1. Le libyque

Il s'agit des variétés du Tifinaghe les plus anciennes. Il existe deux formes du libyque, l'oriental et l'occidental.

2. 1. 2. Le Tifinaghe saharien

Cette variété est également appelée libyco-berbère ou touareg ancien. Elle contient des signes supplémentaires par rapport au libyque, plus particulièrement un trait vertical pour noter la voyelle finale /a/. Cette variété fut utilisée pour transcrire le touareg ancien mais ses inscriptions sont incomprises (IRCAM 2004 b).

2. 1. 3. Le Tifinaghe touareg

Il existe au sein du Tifinaghe touareg quelques divergences dans la valeur attribuée aux signes qui correspondent aux variations dialectales touarègues. Si d'une région à une autre, la forme et le nombre des signes peuvent changer, les textes restent en général mutuellement compréhensibles.

2. 1. 4. Le néotifinaghe

Le néotifinaghe désigne les systèmes d'écriture développés pour représenter les parlers berbères (amazighes) du Maghreb. La première variante fut celle proposée à la fin des années 60 par l'Académie Berbère (AB) sur la base de lettres Tifinaghes touarègues, elle est largement diffusée au Maroc et en Algérie (surtout en Kabylie).

2. 2. Tifinaghe : L'alphabet amazighe

2. 2. 1. Caractères Tifinaghe

L'Alphabet Tifinaghe a été proposé, par l'IRCAM (IRCAM 2004 a), à l'Organisation de Standardisation Internationale qui l'a confirmée. Cette proposition comprend quatre sous-ensembles de caractères Tifinaghes : a) le jeu de base de l'IRCAM ; b) le jeu étendu de l'IRCAM ; c) d'autres lettres néotifinaghes en usage et d) des lettres touarègues modernes dont l'usage est attesté.

La figure 1 montre la liste de l'alphabet Tifinaghe et le plan Unicode associé attribuée par l'ISO :

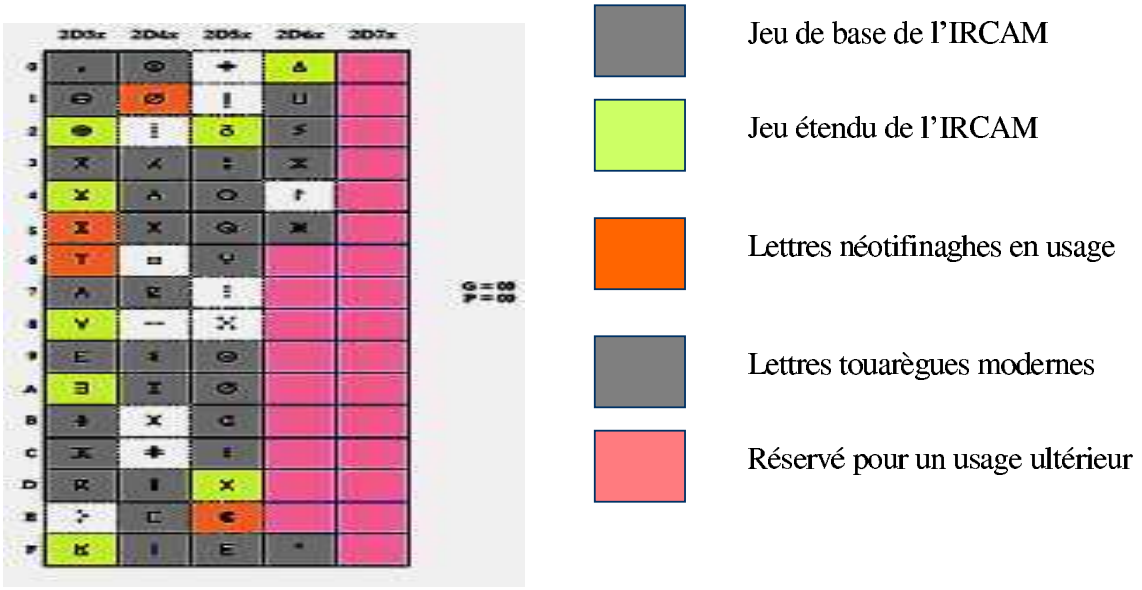


Figure 1: rangé 2D de Unicode : Tifinaghe

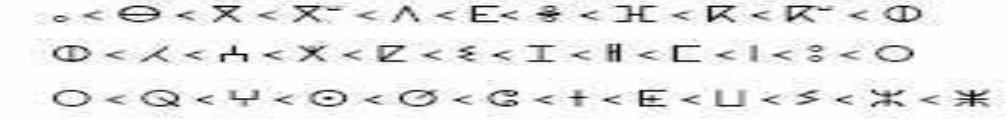
Pour plus de détails sur la phonétique/phonologie de la langue, sur les règles orthographiques et sur les éléments de morphosyntaxe, consulter (IRCAM 2004 c).

2. 2. 2. Ponctuation

Nous ne connaissons pas de signe de ponctuation particulier au Tifinaghe. L'IRCAM a préconisé l'emploi des signes conventionnels qu'on retrouve dans les écritures latines : « » (espace), « . », « , », « ; », « : », « ? », « ! », « … », etc. En conséquence, cette proposition ne présente aucun signe de ponctuation Tifinaghe.

2. 2. 3. Tri

L'IRCAM a défini un ordre précis décrit par l'expression ci-dessous (a < b, signifie que *a* est trié avant *b*



2. 2. 4. Chiffres
L'IRCAM a retenu les chiffres « arabes » occidentaux (0, 1, 2, 3, 4, 5, 6, 7, 8, 9) pour l'écriture Titifinaghe. Cette proposition n'introduit donc aucun nouveau chiffre ou nombre.

2. 2. 5. Directionnalité
L'IRCAM a retenu la direction horizontale de gauche à droite pour l'écriture Tifinaghe.

3. Méthodes pour l'informatisation de l'amazighe

3. 1. Trouver des solutions adaptées : idées forces

3. 1. 1. Bénéficier de développements faits pour des langues liées
L'Amazighe est une langue voisine de l'arabe et du français. Par conséquent, il est très possible de bénéficier des liens qui unissent cette langue au français et à l'arabe pour utiliser leurs ressources et faciliter les développements. Par exemple, on peut intégrer une barre d'outils pour l'écriture Amazighe qui se charge automatiquement avec Ms WinWord version Arabe ou Française. Cette idée sera l'objet d'une section prochaine.

3. 1. 2. S'intégrer à des projets et environnements génériques, open source et à pivot
Projet UNL (www.undl.org): Il est de plus en plus nécessaire de créer des documents multilingues qui intègrent l'Amazighe (Rachidi A. 2004). L'idée actuelle est de faire de la traduction manuelle collaborative sur le Web à l'aide d'une mémoire de polyphrases multilingues (MPM), outils en construction par l'équipe de C. Boitet (GETA, CLIPS, IMAG - Grenoble, France) (Berment V. 2004), puis d'intégrer le résultat (les phrases en Amazighe) dans le document en UNL-XML. Enfin, il faudra construire un déconvertisseur UNL-Amazigh et un enconvertisseur.
Projet Papillon (www.papillon-dictionary.org): L'amazighe intégrera le projet dés que le format Unicode sera bien installé dans les plates formes logicielles. En attendant, il faut préparer une liste exhaustive de catégories (part of speech) qui seront utilisées dans ce dictionnaire.

3. 2. Appliquer une gestion adaptée

1) Déterminer quel produit réaliser ? 2) Déterminer qui réalise les logiciels et ressources et 3) Etablir un plan de développement. En effet, au sein de notre laboratoire et en collaboration avec l'IRCAM, les développements se dérouleront en deux phases distinctes ayant nécessité chacune un plan de développement spécifique : 1) la réalisation d'un traitement de texte bien adapté à l'écriture Amazighe et 2) le développement d'outils d'aide à la traduction et d'une base lexicale en intégrant les projets UNL et le papillon).

4. Mise en œuvre et outils impliqués

4. 1. Phase actuelle (avant le format Unicode)

4. 1. 1. Clavier Amazighe
L'alphabet Tifinaghe est composé de trente trois caractères. Le centre CEISIC de L'IRCAM a proposé un clavier sous format ASCII (police, pilote) comme l'illustre la figure 2 (IRCAM 2003 a) (IRCAM 2003 b). Les 26 premiers caractères sont accessibles directement. Les caractères emphases s'obtiennent en utilisant la case Noir (le « ^ » en clavier latin) de la

même façon qu'on utilise le « ^ » en français (pour taper le « â »). L'IRCAM a proposé un projet qui normalise deux groupes de claviers en précisant deux niveaux de conformité, l'un pour la saisie stricte des trente-deux lettres de l'alphabet Tifinaghe de base tel qu'enseigné dans les écoles marocaines, l'autre pour la saisie de l'alphabet de base, plus les vingt-deux lettres de l'alphabet Tifinaghe étendu et des ligatures (lorsque la technologie sous-jacente permet de traiter deux caractères de commande permettant de former ou d'empêcher la formation de ligatures). Ce projet sera confirmé dans le prochain amendement à la norme internationale ISO/CEI 14651.

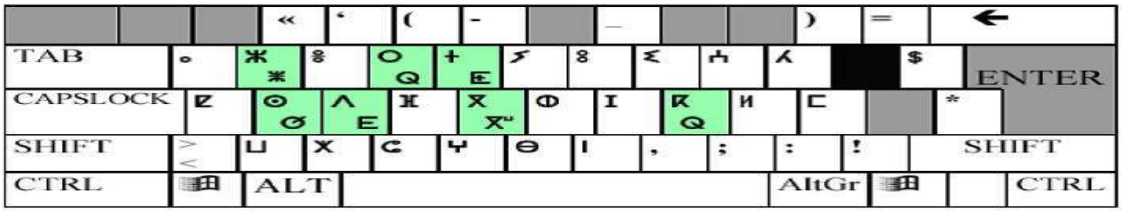


Figure 2 : Clavier Amazighe sous format Ascii

4. 1. 2. Réaliser un traitement de texte pour l'amazighe

On peut réaliser une plate forme logicielle qui permet de traiter un texte Amazighe sous format Ascii avec C / C++ et utilisant uniquement le SDK (Software Development Kit) de Windows, qui visait les premiers niveaux du service de traitement du texte (TALN 2003). Il inclut les fonctionnalités suivantes : 1) saisie de textes amazighe indépendant de la police utilisée et utilisant un clavier intuitif, 2) changement de police (donc transcodage), 3) mise en forme canonique du texte sélectionné (standardisation, saisie non univoque), 4) facilité de sélection à la souris et au clavier des syllabes et des mots Amazighes, 5) formatage de textes Amazighes, pour les rendre utilisables par des traitements de texte commerciaux, 6) export aux formats TeX et RTF, 7) construction d'un lexique à partir de textes (ajouter, modifier, supprimer une entrée dans un lexique local), 8) traduction en français de mots Amazighes et 9) transcription phonétique du texte sélectionné. L'interface proposée de cette plate forme est illustrée dans la figue 3 suivante :

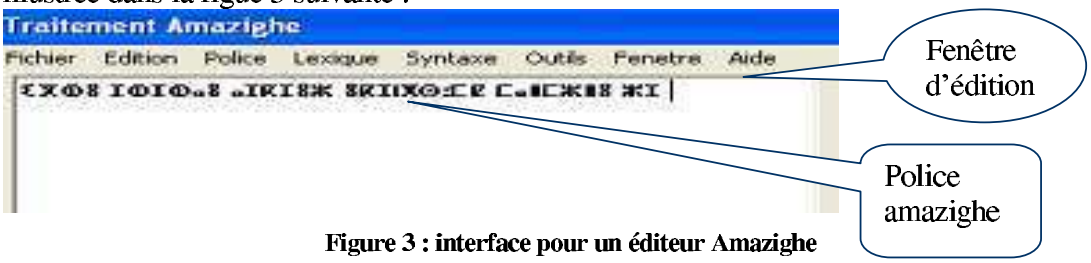


Figure 3 : interface pour un éditeur Amazighe

4. 2. Phase après Unicode : réalisation d'un traitement de texte

4. 2. 1. En utilisant Word Pad

Word Pad peut servir comme base pour une version Amazighe (Amazighe Pad). Ces sources sont disponibles dans C++ de MicroSoft. Word Pad est développé dans C++ et s'appuis sur les classes de la Bibliothèque MFC (Microsoft Fondation Classes) (Berment V. 2004): a) Classe de la fenêtre d'édition CRich Edit Ctrl et b) Format Rich text.

4. 2. 2. En utilisant WinWord

On peut Utiliser le kit du développeur Word (Word's developer's kit). C'est un Ensemble de modules C qui permet de s'interfacer avec Word. Le principe c'est de développer une librairie dynamique qui se charge avec Word et qui appelle les fonctions de Word à travers une API appelée CAPI (Berment V. 2004).

Le format proposé de cette barre est la suivante : 
De la gauche vers la droite, les fonctions des boutons sont : a) configuration AmazigheWord ; b) changement de police Amazighe ; c) tri de tableaux en Amazighe ; d) transcriptions Amazighe ; e) dictionnaire électronique ; f) mise en forme du texte ; g) choix Amazighe-latin pour la saisie des textes et h) l'aide en ligne.

5. Conclusion

Née avec l'informatique, l'informatisation des langues a évolué, offrant de plus en plus de services pour de plus en plus de langues. C'est cependant un processus coûteux qui ne bénéficie actuellement qu'à une faible partie des langues du monde (moins de 1 %). Informatiser l'Amazighe demande un effort de plusieurs chercheurs et l'utilisation des leviers permettant d'obtenir rapidement des logiciels de qualité. L'existence du standard Unicode, qui vient d'intégrer l'Amazighe ces derniers mois, a récemment permis la réalisation de systèmes d'exploitation et de logiciels couvrant de nombreux systèmes d'écriture tout en évitant la multiplication des incompatibilités entre plates-formes. L'Amazighe bénéficie ainsi d'outils d'édition performants. Le codage et les principes de base étant communs aux différents systèmes d'écriture. Notre projet, en collaboration avec l'IRCAM, œuvre pour la réalisation d'un traitement de texte bien adapté à l'écriture Amazighe et le développement d'outils d'aide à la traduction et d'une base lexicale.

Références

Berment V. (2004) méthodes pour informatiser des langues et des groupes de langues «peu dotées', thèse de Doctorat de l'université Joseph Fourier, Grenoble 1, UFR d'informatique et mathématiques appliquées, 18 mai 2004.

IRCAM (20004 a), 'Proposition d'ajout de l'écriture Tifinaghe au répertoire de l'ISO/CEI 10646 (format unicode)', 21/06/2004, centre CEISIC, IRCAM, Rabat, Maroc

IRCAM (2004 b), 'Graphie de la langue amazighe' Centre de l'Aménagement Linguistique Coordinateur El Mehdi Iazzi, Publications de l'IRCAM, Rabat, 2004.

IRCAM (2004 c), 'Initiation à la langue amazighe', Centre de l'Aménagement Linguistique, publication de l'IRCAM, manuel N° 1, Rabat, 2004.

IRCAM (2003 a), 'Élaboration d'une première version du clavier amazighe', Centre des études informatiques et des systèmes d'information et de communication, Rabat, plan d'action 2003.

IRCAM (2003 b), 'Conception et mise au point des polices Amazighes', Centre des études informatiques et des systèmes d'information et de communication, Rabat, plan d'action 2003.

Rachidi A. (2004), ' Les Graphes UNL : un concept unificateur pour l'intégration de l'Amazighe dans des Documents Multilingues', séminaire international : La typographie entre les domaines de l'art et de l'informatique, IRCAM, Rabat septembre. 2004.

TALN (2003), 'Traitement automatique des langues minoritaires et des petites langues', l'atelier associé à TALN 2003.

Un système de lissage linéaire pour la synthèse de la parole arabe : Discussion des résultas obtenus

Tahar SAIDANE (1), Mounir ZRIGUI (2), Mohamed BEN AHMED (3)

(1) Centre de production de Sousse, Société Tunisienne d'Electricité et du Gaz, Tunisie
saidane.tahar@planet.tn
(2) Labaoratoire RIADI, Unité Monastir
Faculté des Sciences de Monastir, Tunisie
mounir.zrigui@fsm.rnu.tn
(3) Labaoratoire RIADI, Ecole Nationale des Sciences de l'informatique, Tunis, Tunisie
Mohamed.BenAhmed@riadi.rnu.tn

Mots clés – Keywords

Synthèse de la parole arabe, Phonèmes, Diphones, Triphones, Unités acoustiques, Dictionnaire de polyphones.

Arabic speech sythesis, Phoneme, Diphones, Triphones, Acoustic units, Polyphones dictionary.

Résumé – Abstract

Notre article s'intègre dans le cadre du projet intitulé "Oréodule" : un système embarqué temps réel de reconnaissance, de traduction et de synthèse de la parole. L'objet de notre intérêt dans cet article est la présentation de notre système de synthèse hybride de la parole arabe. Nous présenterons, dans ce papier, les différents modules et les différents choix techniques de notre système de synthèse hybride par concaténation de polyphèmes. Nous détaillerons également les règles de transcription et leurs effets sur le traitement linguistique, les règles de syllabation et leurs impacts sur le coût (temps et difficulté) de réalisation du module acoustique et nous poursuivrons par l'exposé de nos choix au niveau du module de concaténation. Nous décrirons le module de lissage, un traitement acoustique, post concaténation, nécessaire à l'amélioration de la qualité de la voix synthétisée. Enfin, nous présenterons les résultats de l'étude statistique de compréhension, réalisée sur un corpus.

This research paper is within the project entitled "Oreillodule" : a real time embedded system of speech recognition, translation and synthesis. The core of our interest in this work is the presentation of the hybrid system of the Arabic speech synthesis and more precisely of the linguistic and the acoustic treatment. Indeed, we will focus on the grapheme-phoneme

transcription, an integral stage for the development of this speech synthesis system with an acceptable quality. Then, we will present some of the rules used for the realization of the phonetic treatment system. These rules are stocked in a data base and browsed several times during the transcription. We will also present the module of syllabication in acoustic units of variable sizes (phoneme, diphone and triphone), as well as the corresponding polyphones dictionary. We will list the stages of the establishment of this dictionary and the difficulties faced during its development. Finally, we will present the results of the statistical survey of understanding, achieved on a corpus.

1 Introduction

Notre étude porte sur la conception et la réalisation d'un système de synthèse de la parole arabe qui donne la voix la plus naturelle possible tout en tenant compte des particularités de la langue. Cet objectif a nécessité l'étude de toutes les étapes de la synthèse de la parole et le choix des solutions les plus adaptées à chaque tâche. Le résultat de ces études nous a guidé vers un système de synthèse hybride utilisant la concaténation d'unités acoustiques de tailles variables tout en utilisant des règles établies. Cet article présentera les modules de ce système de synthèse à savoir le transcripteur, le module de syllabation, le dictionnaire d'unités acoustiques et le module de concaténation muni de son système de lissage (Dutoit, 1993).

2 LA TRANSCRIPTION

L'analyse linguistique nous a permis d'établir un ensemble de 133 règles. Il est à noter que l'ordre d'application de ces règles est très important et influe sur le résultat final. En ce qui suit la description de quelques règles élaborées (Saidane, 2004) :

1. [CC]={ّ}+{C}

Lorsqu'une consonne est suivie par la ّ, elle est doublée, on obtient alors le phonème [CC]. Exemple : وَدَّ , رَوَّجَ.

2. {CL}+{ل}+{V}+{CL}={CL}+{ال }+{CL}
3. {CL}+{ل}+{V}+{CS}={CL}+{ال }+{CS}

Lorsque le ال est entre suivi par une consonne lunaire, il est équivalent à la non présence du أ. Exemple : مُنِعَ الأكل, أكلَ الأكلُ (Zrigui 1991).

3 LA SYLLABATION

Les unités acoustiques de notre système de synthèse sont de trois types : les triphones, les diphones et les phonèmes. On a établi un ensemble de règles de concaténation à partir desquelles les différentes occurrences de trois phonèmes pouvaient se transformer en : un triphone, un diphone suivi d'un phonème, un phonème suivi d'un diphone, ou éventuellement trois phonèmes. La sélection dynamique des unités se traduit alors par la recherche de la séquence optimale de représentants, visant à minimiser les discontinuités au point de concaténation (Boula, 2001). Le schéma suivant présente un exemple de syllabation pour l'expression «أَيْنَ المُسَافِرُونَ » (eaj.na.lmusaafiruuna[1]: Où sont les voyageurs) (Saidane, 2004):

[1] Suivant l'alphabet phonétique internationale IPA 96

أَيْنَ الْمُسَافِرُونَ ➡ eaj.na.lmusaafiru ➡ ea j. na .l mu saa fi ru

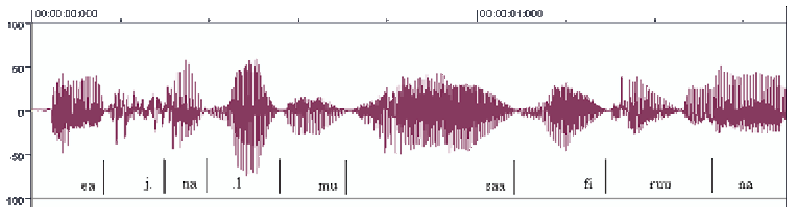


Fig. 1. Exemple de syllabation

La problématique de la sélection des unités a été formalisée en utilisant des règles. Ces règles de syllabation peuvent se résumer en ce qui suit :

1. [CVV] ={V}+{V}+{C} : lorsqu'une consonne est suivie de deux voyelles les trois graphèmes constituent une unité acoustique de notre système.
2. [CV]={C}+{V}+{C} : lorsqu'une consonne est suivie d'une voyelle puis d'une consonne les deux premiers graphèmes constituent une unité acoustique.
3. [CC]={C}+{C}+{C} : lorsque nous avons une succession de trois consonnes les deux premiers graphèmes constituent une unité acoustique.
4. [C]={V}+{C}+{C} : lorsque nous avons deux consonnes suivies par une voyelle seul le premier graphème constitue une unité acoustique.
5. [VV]={V}+{V} : lorsque nous avons une succession de deux voyelles, les deux constituent une unité acoustique.
6. [V]={V} : lorsque nous avons une voyelle isolée elle constitue une unité acoustique.

Il est à noter que l'ordre d'application de ces règles ainsi établies est très important pour une bonne syllabation et donc une meilleure concaténation sonore (Emerard, 1977). Ces six règles de syllabation élaborées vont imposer les types d'unités acoustiques à utiliser pour la synthèse de la parole. Le dictionnaire ainsi établi contient 196 unités acoustiques suffisantes pour la réalisation des différentes occurrences possibles. Le nombre de phonèmes est de 28, le nombre de diphones est de 84 et le nombre de triphones est de 84.Néanmoins, la pratique et l'étude de la langue arabe ont permis de dégager une dizaine d'autres unités dues principalement aux contraintes de la langue.

Le module de concaténation a besoin de la totalité des unités acoustiques sous la forme d'enregistrements sonores (Lemmety, 2000). Ces enregistrements constituent le dictionnaire de notre système. Le dictionnaire d'unités acoustiques ainsi établi a une taille de 9 MØ (en moyenne un phonème prend 20 kØ, un diphone 40 kØ et un triphone 60 kØ).

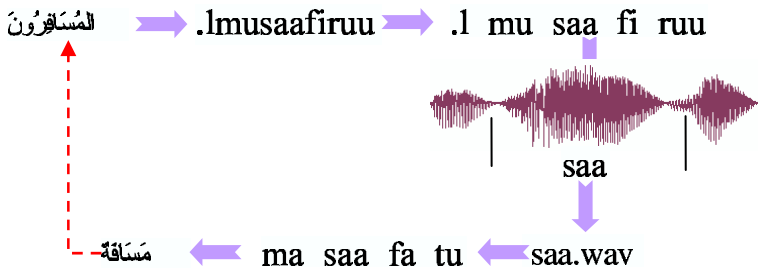


Fig. 2. Un exemple de traitement pour l'obtention du triphone « saa »

4 LA CONCATENATION

Pour notre système nous avons voulu commencer par un traitement de lissage temporel pour mesurer l'effet d'un post traitement sur la qualité de la parole obtenue. Après l'analyse des différentes unités acoustiques de l'arabe, il s'avère que celles-ci présentent une atténuation aux niveaux de leurs extrémités. L'idée retenue consiste alors à procéder, lors de la concaténation, à une accentuation aux niveaux d'un certain nombre de valeurs d'extrémités avant le collage en bout à bout. Ce traitement touchera évidement la fin de la première unité et le début de la suivante.Un signal numérique de la parole étant :

$$s(t) = \sum_1^N s_n \, \delta(t - nT) \qquad (1)$$

s(t) : signal numérisé de la parole (échantillonné), s_n = s(nT) : la valeur du signal à l'instant nT et δ(t) : impulsion de Dirac. La concaténation de deux unités sera :

$$s(t) = s_1(t) + s_2(t) = \sum_1^N s_{1n} \, \delta(t - nT) + \sum_1^M s_{2n} \, \delta(t - nT) \qquad (2)$$

L'idée consiste alors à isoler X valeurs du premier signal et Y valeurs du second. Ces valeurs subiront alors une atténuation proportionnelle définie par :

$$s_i^{\text{atténué}} = s_i \frac{K - i}{K} \quad i = 1 \,..\, K \qquad (3)$$

Le résultat se présentera sous la forme :

$$s(t) = \sum_1^{N-X} s_{1n} \, \delta(t-nT) + \sum_{N-X+1}^{N} s_{1n} \frac{N-n}{N} \delta(t-nT) + \sum_1^{Y} s_{2n} \frac{Y-n}{Y} \delta(t-nT) + \sum_{Y+1}^{M-Y} s_{2n} \, \delta(t-nT) \qquad (4)$$

La fonction d'atténuation ainsi définie a été appliquée pour un nombre de points représentants 10 % de la durée du signal de l'unité acoustique. Les résultats obtenus sont montrés en ce qui suit :

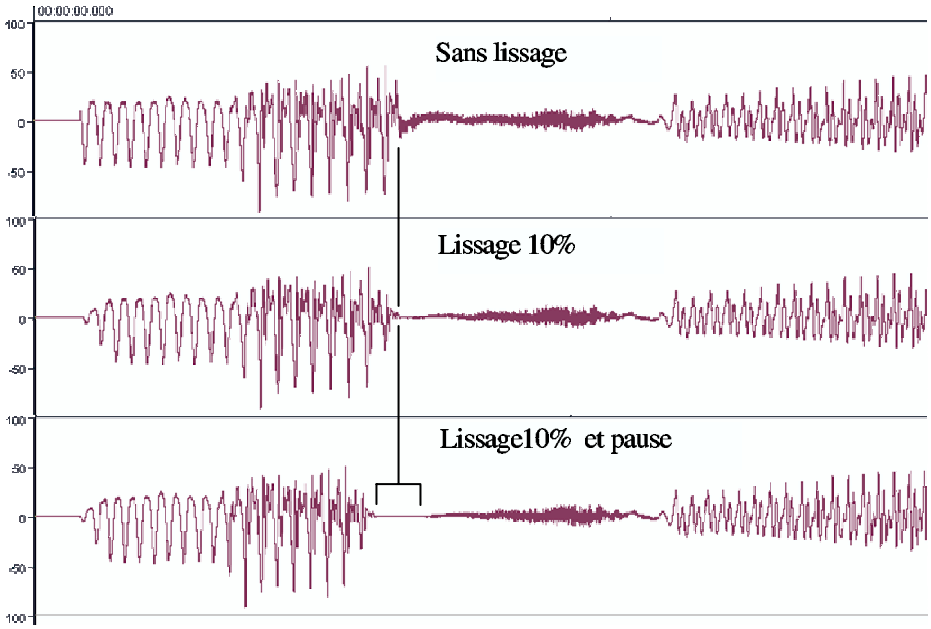


Fig. 3. Effet du lissage temporel sur la forme d'onde au niveau des points de discontinuités.

Les courbes précédentes montrent l'effet de ce lissage temporel sur un exemple de synthèse du mot « مَسَافَة » (masaafatun : distance). En effet, la première courbe montre une concaténation bout à bout nous y constatons une discontinuité flagrante aux niveaux des points de jointures. La courbe du bas introduit, quant à elle, le résultat d'une concaténation lissée et la fluidité aux niveaux des points de concaténation. Le résultat obtenu a sensiblement amélioré la qualité de la voix synthétisée. Néanmoins, nous constatons un chevauchement entre les unités. Pour éviter un tel problème nous avons introduit un temps de silence de 10 millièmes de seconde. L'insertion d'une pause entre les unités avec nous a alors permis d'obtenir une meilleure intelligibilité.

5 RESULTATS DES TESTS

Afin d'évaluer notre système, nous avons établi une procédure de test basée sur l'écoute et l'identification de phrases synthétisées. Nous avons utilisé 20 phrases, soit 53 mots, 211 unités acoustiques dont 73 différentes ce qui constitue 37.2 % de la totalité des unités acoustiques qu'utilise notre système. Nous les avons fait écouter à 8 personnes (4 femmes et 4 hommes) ce qui a permis une évaluation statistique réaliste du résultat. Chaque phrase est écoutée trois fois, à chaque passage le sujet doit orthographier ce qu'il entend. En ce qui suit les résultats obtenu :

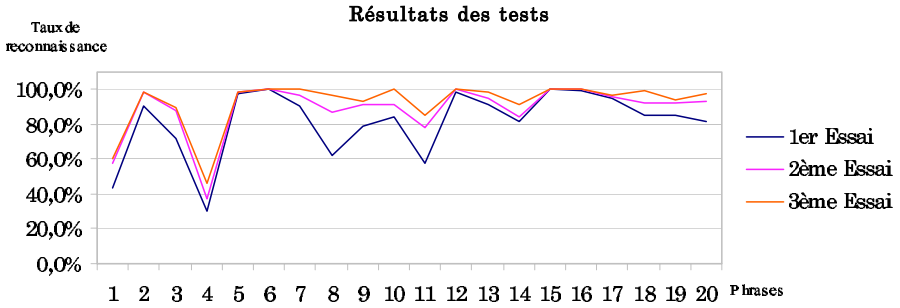


Fig. 4. Les résultats de la phase de test

Nous avons alors pu conclure à un pourcentage d'identification de plus de 81 % dès la première écoute, ce taux passe à plus de 92% pour la troisième phase. Par ailleurs nous avons remarqué qu'une phase d'adaptation de 2 à 3 phrases a été nécessaire pour avoir une stabilisation des taux de reconnaissance. De ces relevés nous avons aussi constaté que les mots non courants sont difficilement identifiables (exp : لَدَغَتْهُ phrase n° 4), et que quelques caractères sont plus difficiles que d'autres pour l'identification (exp :ذ phrase n° 3, 4 et 11).

6 CONCLUSION

Nous avons présenté dans cet article notre système de synthèse de la parole, ces différents constituants, les différentes phases de son élaboration et les choix techniques retenus pour chaque module. Le module de syllabation constitue à notre sens le point de départ pour une autre vision de la langue arabe, vue la rupture totale avec les méthodes jusque là utilisées en

langue arabe. Nous avons aussi exposé l'opération de concaténation ainsi que le poste traitement que nous avons choisi pour remédier aux problèmes de discontinuités.

La comparaison des résultats obtenus par rapport à l'existant demeure difficile. Les travaux sur les systèmes de synthèse de la parole arabe sont peu nombreux et les résultats d'évaluation ne font pas l'objet d'articles publiés. Néanmoins nous avons relevé que notre système a permis de se restreindre à trois types de syllabes seulement (CVV, CV et C) contrairement aux autres travaux préconisant cinq voir six types de syllabes différents (Ben Sassi, 2001). Nous n'utilisons que 196 unités acoustiques pour synthétiser n'importe quelle occurrence de l'arabe standard alors que le minimum jusque là était de 310 unités (Elshafei, 2002).

Références

1 Zrigui M., Mili A, Jemni M. 1991. Vers un système automatique de synthèse de la parole arabe, Maghrebin symposium on programming and system, Alger. p 180-197.

2 Saidane Tahar, Zrigui Mounir, Pr Ben Ahmed Mohamed. 2004. La Transcription Orthographique-Phonétique de la Langue Arabe. RÉCITAL 2004, Fès, Maroc.

3 Emerard Frençoise. 1977. Les diphones et le traitement de la prosodie dans la synthèse de la parole. Bulletin de l'institut de phonétique de grenoble.

4 Dutoit Thierry. 1993. High quality text to speech synthesis of the french language. Thèse. Faculté polytechnique de Mons.

5 Elshafei M., Al-Muhtaseb, H., Al-Gamdi M. 2002. Techniques for high quality Arabic speech synthesis, Information sciences, Vol.140, 255-267.

6 Ben Sassi S., Braham R., Belgith A. 2001. Neural speech synthesis system for Arabic language using celp algorithm, Proc. Conference on Computer Systems and Applications.

7 Saidane Tahar, Haddad Ahmed, Zrigui Mounir, Pr Ben Ahmed Mohamed. 2004. Réalisation d'un système hybride de synthèse de la parole arabe utilisant un dictionnaire de polyphones. JEP-TALN 2004, Traitement Automatique de l'Arabe, Fès, Maroc.

8 Boula de Mareuil Philippe, Célérier Philippe, Cesses Thierry, Fabre Serge, Jobin Carine, Le Meur Pierre-Yves, Obadia David, Soulage Benoît, Toen Jacques. 2001. Elan text to speech : un système multilingue de synthèse de la parole à partir du texte. Elan TTS Toulouse.

9 Lemmety Sami. 2000. Review of speech synthesis technology. Thèse. Helsinki University of Technology.

Clustering Web Pages to Identify Emerging Textual Patterns

Marina Santini

University of Brighton
Lewes Rd, Brighton, UK
Marina.Santini@itri.brighton.ac.uk

Mots-clefs – Keywords

genre textuels sur le Web, typologie des pages Web, analyse de groupement

Web genres, Web Text Types, Web pages, Cluster Analysis

Résumé – Abstract

Le Web a causé beaucoup de changements dans plusieurs domaines. Il a aussi influencé l'inventaire des genres textuels traditionnels. De nouveaux genres ont été créés, par exemple les *blogues* et *foires aux questions*. Il est probable que d'autres genres soient en train de se former, parce que le Web est un médium qui change constamment. Dans cet article, nous présentons une expérience qui vise à faire apparaître de façon inductive les plans textuels émergents, qui peuvent devenir un nouveau genre ou une nouvelle typologie textuelle dans peu de temps. Il s'agit de regrouper (analyse de groupement) les pages web en utilisant des traits linguistiques et de présentation. Les résultats sont encourageants et invitent à poursuivre la recherche dans ce domaine.

The Web has triggered many adjustments in many fields. It also has had a strong impact on the genre repertoire. Novel genres have already emerged, e.g. *blog* and *FAQs*. Presumably, other new genres are still in formation, because the Web is still fluid and in constant change. In this paper we present an experiment that explores the possibility of automatically detecting the emerging textual patterns that are slowly taking shape on the Web. Emerging textual patterns can develop into novel Web genres or novel text types in the near future. The experimental set up includes a collection of unclassified web pages, two sets of features and the use of cluster analysis. Results are encouraging and deserve further investigation.

1 The Web: An Evolution still in progress

The Web has had a strong impact on the genre repertoire. Novel genres have already emerged (Crowston and Williams 1997) such as *personal home pages, hotlists, FAQs*, and more recently *blogs, ezines*, *clogs*, etc. Presumably, other new genres are still in formation, because the Web is still fluid and in constant change. Three main well-established genre categories have been identified on the Web: reproduced/replicated, adapted/variant, novel/spontaneous (Crowston and Williams 1997; Shepherd and Watters 1998). Many genres on the Web are reproduced or replicated genres, i.e. they are traditional paper genres that have been transplanted into an electronic form, such as *academic papers*. However, most genres coming from the paper world have undergone some adjustments when moving on to the Web and variants have been created. For instance, *online newspapers* and *online manuals* show the adaptation of paper genres to the functionalities provided by the Web (cf. Crowston and Williams 1999, Shepherd and Watters 1999). Some genres are novel and spontaneous. They have become fully acknowledged and genre labels have been invented for them only in recent years, for instance *home pages* (personal, academic, organizational, etc.), *FAQs, newsletters, emails, weblogs*. However, there are many web pages that do not fall into one of these three categories. It is often hard to assign a genre label to a web page. Many web pages remain "unclassified" or labelled as "mixed" (Santini 2005a). We suggest that those web pages that are unclassifiable might represent an emerging textual pattern, i.e. a new textual organization, strongly influenced by the hypertextual structure and the functionalities provided by the Web.

In this paper we present an experiment that explores the possibility of automatically detecting emerging textual patterns that are slowly taking shape on the Web. Emerging textual patterns are interesting because they can reveal novel genres or novel text types in an embryonic form. The proposed approach consists in running cluster analysis (an inductive/unsupervised statistical algorithm based on similarity measures) in order to create groups of similar pages across a corpus of 1000 unclassified English web pages. The qualitative analysis of these groupings (the clusters) would reveal whether emerging patterns could be identified. This experiment does not include any classification tasks, because it would be hard to classify something which is not fully formed. The goal here is the analysis and the interpretation of new textual patterns, if any, brought about by the dynamism of the Web.

As mentioned earlier, emerging textual patterns are embryonic forms that are likely to develop into novel Web genres or novel text types in the near future. While emerging textual patterns are related to groups of documents that show new textual traits, genres and text types refer to fully formed categories. Many definitions of genre and text types have been formulated since Aristotle. Here, by genre, we refer to the socio-cultural connotation of a document together with the linguistic/discoursal devices enacted in the document itself. For instance, a *letter*, a *manual*, an *article*, an *academic paper* are genres. Web genres are genres that are used on the Web, and they range from plain electronic versions of paper genres, to genres more tailored to take advantage of the potentials of the Web. By text types, we refer to the purpose of the text, i.e. the reason for which a text has been written. Text types are related to the producer's intention towards the receiver(s). An advert is written to *persuade* customers to buy something; a car manual might *instruct* on how to fix a component in the car. Researchers involved in automatic genre classification rely on texts which are pre-classified by genres and use discriminant analysis or supervised learning to "learn" from exemplar texts belonging to a restricted set of genres and generalize this learning over unseen/unclassified documents. The quantitative approach to text types identification, instead, is mainly linked to the multi-dimensional analysis proposed by Biber (1988, 1989, 1995, 2004). His text types cut across genres (Biber 1988, 1989) or registers (Biber 1995). More recently, he has started

sketching a typology of web registers (Biber 2004) by using two main Google topical categories incorporating multiple subcategories. Biber's approach has strongly influenced two projects for French, TyPText (Folch et al. 2000 and Illouz et al. 2000) and TyPWeb (Beaudouin et al. 2001a and 2001b) (see Santini 2004 for a state-of-the-art of genre and text type identification).

The paper is organized as follows: section 2 briefly describes previous work on the same subject and identifies references to "unclassifiable" web pages in web user studies; section 3 describes the experiment and the results; section 4 draws some conclusions.

2 Related Work

So far, only a very recent exploratory study has been carried out to address the issue of automatic detection of emerging textual patterns on the Web, more specifically the identification of genres still in formation (Santini 2005a). From this preliminary study, it appears that although automatic clustering did not return any emerging genres, traditional rhetorical/discoursal types could be identified, despite some noise. The presence of this noise was ascribed to the use of shallow features, too shallow to highlight textual novelties.

Some references to "unclassified" web pages can be derived indirectly from the few surveys carried out so far on web pages. Crowston and Williams (1997) found that some of the web pages could not be classified because they did not have a recognizable genre. In these cases, the raters agreed that there was a genre, but did not know its name, and labelled the pages as "unclassified". Interestingly, one of the conclusions was that some of these unclassified pages could be interpreted as belonging to "emerging genres". Similarly, in Roussinov et al. (2001)'s exploratory user study on Web genres, a number of pages could not be classified, but no special conclusions were drawn.

Understandably, the main interest of web page surveys is to find what can be classified. However, web pages that are "unclassified" today, might become instantiations of a new web genre or a new text type tomorrow. In this respect, unclassified web pages could be seen as anticipations or forerunners of new textual categories, not fully formed yet.

3 Experiment

3.1 Web Page Collection

The SPIRIT collection is a random crawl with an initial seed of university websites carried out in 2001 by a Canadian university (Clarke et al. 1998). It contains single web pages rather than complete websites. This collection includes about 95 million web pages. It is multilingual and without any meta-information, except a short header including the original URL, the date and time when the pages were crawled from the Web, and a few other details. It represents a genuine slice of the real Web. 1000 random English web pages were extracted from this collection and used in the experiment.

3.2 Features

Two sets of features were used, each including three subsets. The first set of features includes 28 functional cues, 29 syntactic patterns, and 33 HTML tags (90 features). We will refer to this set as *sy_pat*. The second set includes instead 28 functional cues, 52 connectives and subordinators, and 33 HTML tags (113 features). We will refer to this set as *con_sub*. Functional cues and syntactic patterns are hand-crafted and parser-dependent features (the

parser used in this experiment is Connexor by Tapanainen and Järvinen (1997); Santini (2005b) contains the description and motivation of these features). Connectives and subordinators are lexical items. They represent an easy way to capture syntactic and discoursal information, even though their semantic interpretation is often ambiguous. Finally, HTML tags account for layout and functionalities, both important elements in a web page.

Two sets of features represent two views on the same data. As cluster analysis has a somewhat subjective nature, cluster solutions must be validated. Here we use the extent of the overlap between the two cluster solutions returned by the two sets of features as a measure of the stability of the final clusters.

3.3 Methodology

Cluster analysis is said to have the potential to reveal structures within the data by grouping homogeneous objects together on the basis of similarity measures. It can be used in an exploratory or confirmatory way. Here the aim is exploratory. The clustering algorithm chosen for the experiment is K-means, as implemented in SPSS. K-means is suitable for large datasets, it is easy to understand and very fast. However, it involves two hard decisions, one concerning the number of clusters that better represent the data, the other related to the set of initial seeds. Several alternatives are available (cf. Anderberg 1973). In this experiment, random seeds were used (default in SPSS), and the number of clusters was selected on the basis of the maximum distance between clusters and the minimum distance within each cluster. This approach ensures the highest distinctiveness and compactness of a solution. A 15-cluster solution was suggested for both sets of features.

The following steps were performed:

- Extraction of 1000 random English web pages from the SPIRIT collection.
- Parsing of the text-only version of the web pages.
- Extraction and frequency counts of the two sets of features.
- Normalization of the frequencies by the number of words in a web page.
- Transformation of the normalized frequencies into z-scores (z-scores represent the number of standard deviations that a raw score is above or below the mean; they represent the deviation from a "norm", and can be used as a way of weighing features within a corpus).
- Selection of the best cluster solutions, one for each set of features.
- Measure of the overlap between the two best cluster solutions.
- Qualitative analysis of the overlapping areas between the two best cluster solutions.

The two cluster solutions were compared using a method commonly used in document clustering, i.e. the comparison by pairs. First, all the possible unique combinations of pairs of the 1000 documents were computed (499,500 pairs). Then an algorithm was built to answer the two following questions: "does the pair get classified as "same" or "different" by the *sy_pat* cluster solution" and "does the pair get classified as same or different by the *con_sub* cluster solution"? The overlap between the two clustering solutions was computed using a two-by-two contingency table. The simple matching coefficient used to measure the overlap had a value of 0.61 (this coefficient ranges from 0=no overlap to 1=full overlap). This value shows an overlap of above 60%, i.e.10% more than the random baseline, and represents an acceptable degree of stability of the clusters returned by the two sets of features. For the qualitative analysis, we selected web pages shared between the two solutions and closest to the cluster centroids (intuitively, the most representative of each cluster).

3.4 Results and Discussion

There is almost a perfect distributional overlap between the minority clusters (see poster). The type of these web pages is easily recognizable. These pages comprises very short server messages, lists of names and extensions, a glossary, bibliographies, tables, summary lists; etc. We asked a web user to manually cluster the 24 web pages in the minority clusters and assign six labels (*server message, telephone directory, bibliographic references, glossary, tabular information, summary list*) to the pages. Then we compared the user's clustering with the automatic clustering and computed the K statistic as an inter-rater measure (Carletta 1996). The value returned was above 0.90, indicating a very good level of agreement. The almost complete agreement of the two cluster solutions with human assessment on the minority clusters is an important confirmation of the validity of the approach.

As for the majority clusters (see poster), web pages in "*con_sub* cluster 6" (368 cases) fall almost entirely (more than 98%) into "*sy_pat* cluster 1". This is a sign of stability. This cluster includes web pages with highly laid-out information, little text, centered (hot)lists, a photograph with personal details, such as address, phone, email, etc. They could be seen as ***contact web pages***. Web pages in "*con_sub* cluster 2" (39 cases) mainly fall into "*sy_pat* cluster 1" (72%). These web pages all share a highly laid-out textual organization, with a large number of hyperlinks, short schematic information, block language, many images. It seems that they share the common purpose of conveying information quickly and exhaustively, leaving to the user the decision whether to display more details by following the hyperlinks. This textual pattern is cross-genres. An e-shop, headlines, an art gallery, an animation website are some of the web genres that were gathered together by the clustering algorithm. The purpose seems to be a ***quick information delivery***. "*Con_sub* cluster 1" (508 cases) is spread across the following *sy_pat* clusters: cluster 1 (35%), cluster 3 (35%) and cluster 9 (25%). This large cluster is more heterogeneous and many of its web pages look like "containers", showing a multi-purpose intent. This cluster is still too diversified to inspire a single label.

In summary, minority clusters show extremely well-defined textual profiles, corresponding to recognized categories by human assessment. As highlighted earlier, these clear-cut clusters confirm that the approach is valid and the features are robust enough to show clear similarities among web pages. Two majority clusters can be seen as emerging textual patterns that we labelled as ***contact web pages*** and ***quick information delivery***. It will be interesting to see whether such textual patterns (still a little bit loose) will develop into a actual web genres or text types in the near future. The largest cluster, instead, shows a kind of mixed textuality and appears to be still too heterogeneous. The cluster is too big to have one single profile. Many of its web pages look like "containers" showing different communicative purposes.

4 Conclusions

Results are encouraging and the approach seems to be effective in providing hints about emerging textual patterns. With unsupervised techniques, deep linguistic features appear to be much more effective than shallow features (cf. Santini 2005a). Although vagueness and elusiveness are common conditions before a novel textual pattern becomes formally and functionally established and recognized, a couple of emerging textual patterns could be identified. In our opinion, they have a good chance of becoming novel web genres or text types once their traits become more tightly woven.

Although the objective evaluation of emergent textual patterns is a new and open issue (any discussions on this subject will be fruitful), as a whole the approach proposed to identify

emerging textual patterns seems to be valid. It may represent a starting point for further investigation on the textuality of web pages.

Références

Anderberg M. (1973), *Cluster Analysis for Application*, Academic Press, New York-London.

Beaudouin V., Fleury S., Habert B., Illouz G., Licoppe C., Pasquier M. (2001a), TyPWeb: décrire la Toile pour mieux comprendre les parcours, *Proc. of CIUST 2001*, France.

Beaudouin V., Fleury S., Habert B., Illouz G., Licoppe C., Pasquier M. (2001b), *Traits textuels, structurels et présentationnels pour typer les sites web personnels et marchands*, available at http://www.atala.org/je/010428/TyPWeb.ppt

Biber D. (1988), *Variations across speech and writing*, Cambridge University Press, UK.

Biber D. (1989), A typology of English texts, *Linguistics*, Vol. 27, 3-43.

Biber D. (1995), *Dimensions of register variation*, Cambridge University Press, UK.

Biber D. (2004), *Towards a typology of web registers: A multi-dimensional analysis*. Invited lecture, Conference on Corpus Linguistics: Perspectives for the future. Heidelberg University.

Carletta J. (1996), Assessing agreement on classification tasks: the kappa statistic, *Computational Linguistics*, Vol. 22, 2, 249-254.

Clarke C., Cormack G., Laszlo M., Lynam T., and Terra E. (1998), The Impact of Corpus Size on Question Answering Performance, *Proc. of the 25th Annual Intern. ACM SIGIR Conf. on Research and Development in IR*, Finland.

Crowston K., Williams M. (1997), Reproduced and Emergent Genres of Communication on the World-Wide Web, *Proc. of the 30 Hawaii Intern. Conf. on System Sciences*, USA.

Crowston K., Williams M. (1999), The Effects of Linking on Genres of Web Documents, *Proc. of the 32 Hawaii Intern. Conf. on System Sciences*, USA.

Folch H., Heiden S., Haber B., Fleury S., Illouz G., Lafon P., Nioche J., Prévost S. (2000), TyPText: Inductive typological text classification by multivariate statistical analysis for NLP systems tuning/evaluation, presented at *LREC 2000*, Greece.

Illouz G., Habert B., Folch H., Heiden S., Fleury Serge, Lafon S., Prévost S. (2000), TyPText: Generic features for Text Profiler, *RIAO 2000*, France.

Santini M. (2004), *State-of-the-art on Automatic Genre Identification*, Tech. Report ITRI-04-03, 2004, Brighton University, UK.

Santini M. (2005a), Genres In Formation? An Exploratory Study of Web Pages using Cluster Analysis, *Proc. of the CLUK 05*, UK.

Santini M. (2005b), *Linguistic Facets for Genre and Text Type Identification: A Description of Linguistically-Motivated Features*, Tech. Report, ITRI-05-02, Brighton University, UK.

Shepherd M., Watters C. (1998), The Evolution of Cybergenre, *Proc. of the 31 Hawaii Intern. Conf. on System Sciences*, USA.

Shepherd M.,Watters C. (1999), The Functionality Attribute of Cybergenres, *Proc. of the 32 Hawaii Intern. Conf. on System Sciences*, USA.

Tapanainen P., Järvinen T. (1997), A non-projective dependency parser, *Proc. of the 5 Conf. on Applied Natural Language Processing*, USA.

Memory-based-Learning et Base de règles pour un Etiqueteur du Texte Arabe

Yamina TLILI-GUIASSA
Laboratoire de Recherche en Informatique
Université Badji Mokhtar, Annaba Algerie
Mel : guiyam@yahoo.fr

Mots clés – Key words :

Etiquetage, Memory-Based Leaning, K-NN, Base de règles, Morphosyntaxique, Langue Arabe.

Tagging, Memory-based learning, K-NN, Based-rules, Morphosyntaxitic, Arabic language.

Résumé – Abstract

Jusqu'a présent il n'y a pas de système automatique complet pour l'étiquetage du texte arabe. Les méthodes qu'elles soient basées sur des règles explicites ou sur des calculs statistiques, ont été développées pour pallier au problème de l'ambiguïté lexicale. Celles-ci introduisent des informations sur le contexte immédiat des mots, mais font l'impasse sur les exceptions qui échappent aux traitements. L'apparition des méthodes Memory-Based Learning(MBL) a permis l'exploitation automatique de la similarité de l'information contenue dans de grandes masses de textes et , en cas d'anomalie, permet de déduire la catégorie la plus probable dans un contexte donné, sans que le linguiste ait à formuler des règles explicites. Ce papier qui présente une approche hybride combine les méthodes à base de règles et MBL afin d'optimiser la performance de l'étiqueteur. Les résultats ainsi obtenus, présentés en section 6, sont satisfaisants et l' objectif recherché est atteint .

Since now there is no complete automatic system for tagging an Arabian text. Methods based on explicit rules or on statistical calculations, have been developed to palliate problems of lexical ambiguousness. They introduce some information on the immediate context of the words but , make the dead end on the exceptions that escape to treatments. The apparition of the Memory-Based Learning(MBL) methods, that exploit automatically the similarity of information contained in big masses of texts and permit, in case of anomaly, to deduct the likeliest category in a given context, without the linguist has to formulate explicit rules. This paper presents an hybrid approach that combines methods based on rules and MBL, thus, in order to optimize the labeller's performance. Our objective is reached and the gotten results, presented in section 6, are satisfactory.

1 Introduction

L'étiquetage morphosyntaxique a pour but d'associer une étiquette grammaticale à chaque mot de la phrase. La première étape est alors de définir un jeu d'étiquettes, adapté au découpage en tronçons, l'étiqueteur morphosyntaxique doit fournir au module de découpage en tronçons toutes les informations grammaticales dont il a besoin pour mener à bien son processus. La majorité des publications sur l'étiquetage automatique font l'impasse sur les exceptions ignorées par les règles morphosyntaxiques, et la plupart des systèmes ont un comportement assez flou, voir inconsistant sur les cas épineux. Particulièrement, la langue arabe présente beaucoup de ces cas, cependant elle est très riche sur le plan morphologique. Certains préfixes, suffixes et des informations fournis par la prise en considération d'un contexte large jouent un rôle capital dans l'analyse morphosyntaxique. Les règles sont correctes pour la majorité des cas, mais la représentation des connaissances de ces règles n'est nécessairement pas parfaite pour tous les cas. Les exceptions ignorées par ces règles doivent être traitées par un autre processus qui prendra en charge les cas épineux avec une grande performance. Pour résoudre ce problème nous proposons une combinaison de la méthode à base de règles et la méthode basée sur l'algorithme des K-plus proches voisins (K-NN)[1]. L'approche proposée garde les K-NN pour chaque erreur commise par la règle. Le but de ce travail est de pallier aux problèmes posés au niveau de découpage en tronçons de la phrase arabe. Ainsi, fondée sur la combinaison de la méthode à base de règle et memory-based learning (MBL)[2], le type d'étiquette est déterminé par la première méthode et sera vérifié par la deuxième. Si le contexte courant est une exception de la règle alors le processus de calcul de similarité est déclenché, cette méthode est efficace pour la prise en charge des exceptions.
L'article présente brièvement l'état de l'art en section 2, traite l'étiquetage morphosyntaxique à base de règles dans la section 3, La section 4 décrit l'étiquetage par MBL et la section 5 explique l'approche proposée. La section 6 expose les résultats obtenus et se termine par une conclusion.

2 Etat de l'art

Le problème rencontré dans les analyseurs syntaxiques traditionnels est celui de la *combinatoire*, qui peut être d'origine lexicale[3] ou structurale[4]. De nouvelles méthodes ont été développées pour pallier à ce problème de combinaison. Fondées sur l'étiquetage morphosyntaxique ou *tagging,* ces méthodes permettent de réduire l'ambiguïté lexicale en introduisant des informations sur le contexte immédiat des mots. Le tagger vient ainsi se substituer à l'analyseur morpho lexical avec comme nouvelle ressource une base de connaissances contextuelles. Ces connaissances peuvent être soit de type probabiliste (si le tagger utilise des informations statistiques sur la contiguïté des mots), soit sous forme de règles explicites.

- bases de règles(TAGGIT de Greene, Rubin, 1971, Francis, Kucera, 1982).
- basées sur les données(Bahl, Mercer,1976, Debili, 1977, Leech et al., 1983, Church, 1988, DeRose, 1988).
- machine learning(TiMBL de Daelemans et al ., 2001).

[1] Les K plus proches voisins
[2] Memory-based learning
[3] plusieurs étiquettes pour un token
[4] plusieurs structures pour une phrase

- méthodes hybrides(Brill, Eric, 1995), Leech et al., 1994, Tapanainen, Voutilainen, 1994, Tzoukermann et al., 1995)

3 L'étiquetage morphosyntaxique à base de règles

L'étiquetage grammatical est l'opération qui consiste à attribuer à chacun des mots d'un texte la catégorie (nom, verbe, adjectif, article défini, etc.) qui est la sienne dans le contexte où il apparaît. Malgré l'apparente simplicité de la formulation, l'objectif n'est toujours pas atteint, ou seulement partiellement. La longévité de l'intérêt porté à l'étiquetage témoigne de la difficulté que celui-ci devait en fait receler. Et en même temps de son utilité au regard des applications mettant en œuvre le langage naturel. L'étiquetage de l'arabe hérite de cette situation de fait et voit même sa difficulté s'amplifier lorsque les textes visés se présentent sous leur forme non pas voyelle, mais partiellement seulement, ou encore totalement non voyelle, ce qui correspond au cas le plus courant (Van Mol, 2001).

3.1 Segmentation

Morphologiquement les langues riches, comme l'arabe, présentent des défis significatifs à des applications de traitement de langage naturel parce qu'un mot donne souvent plusieurs significations complexes. L'étiquetage par règles morphosyntaxiques utilise les préfixes et les suffixes comme des identificateurs de catégorie de mot : إسْ# تَفْعَل +وا :إستَفْعَلُوا . La méthode proposée par (L.Young-suk et al., 2002) présente plusieurs avantages pour notre approche ainsi que pour le jeu d'étiquettes adopté(S.Khoja et al., 2001).

3.2 Règles de déduction

1. Les noms
   Le processus pour identifier les noms et les noms propres s'appuie sur les travaux de (S.Abuleil et al., 2002) et (Abuleil, Evens, 1999) respectivement.

2. Les verbes
   La majorité des verbes arabes suivent des règles claires qui peuvent définir leurs morphologies et génère leurs paradigmes, la technique décrite dans (Beesley, Karttunen, 2000) est utilisée dans le système proposé.

3. Les outils
   Ils sont stockés dans une base.

4 L'étiquetage morphosyntaxique par Memory-based learning

Le MBL est décantant direct de l'algorithme K-NN qui utilise des structures de données complexes. Il a un nombre de propriétés intéressantes (W .Daelemans et al., 1996): i) pas de traitement additionnel de lissage pour les données rares. ii) les exceptions peuvent contribuer à une généralisation. (Zavrel, Daelemans, 2000). La similarité entre une instance x et les exemples stockés en mémoire est calculée en utilisant la métrique distance $\Delta(x, y)$:

$\Delta(x, y) = \sum \alpha_i \delta(x_i, y_i)$ ou α_i est le poids de $i^{éme}$ attribut , $\delta(x_i, y_i) = 0$ si $x_i = y_i$ et $\delta(x_i, y_i) = 1$ si $x_i \neq y_i$ (Zavrel, Daelemans, 2000, Park, Zhang, 2003)).

5 L'étiquetage morphosyntaxique

L'architecture générale de l'étiqueteur est donnée par la figure 1. Dans la phase apprentissage chaque mot est analysé par les règles, une étiquette est déterminée. L'étiquette déterminée doit être comparer à l'étiquette en entrée, en cas de non égalité alors le mot avec ces deux étiquettes sont stockées dans une liste appelée liste d'anomalies. Durant la classification l'étiquette du mot M_i est déterminée en regardant le mot et le contexte approprié C_i, le type de l'étiquette est ainsi déterminé par le calcul des similarités entre les instances stockés en mémoire et le mot cible (l'étiquette déterminer par les règles est alors écarté). Les informations utilisées pour calculer $\Delta(x, y)$ sont des valeurs qui représentent les mots et les catégories dans une fenêtre à 3 éléments (Hacioglu, Ward, 2003). Les plus importantes sont le mot en question, le tag du mot précédant (M. Diab et al., 2004).

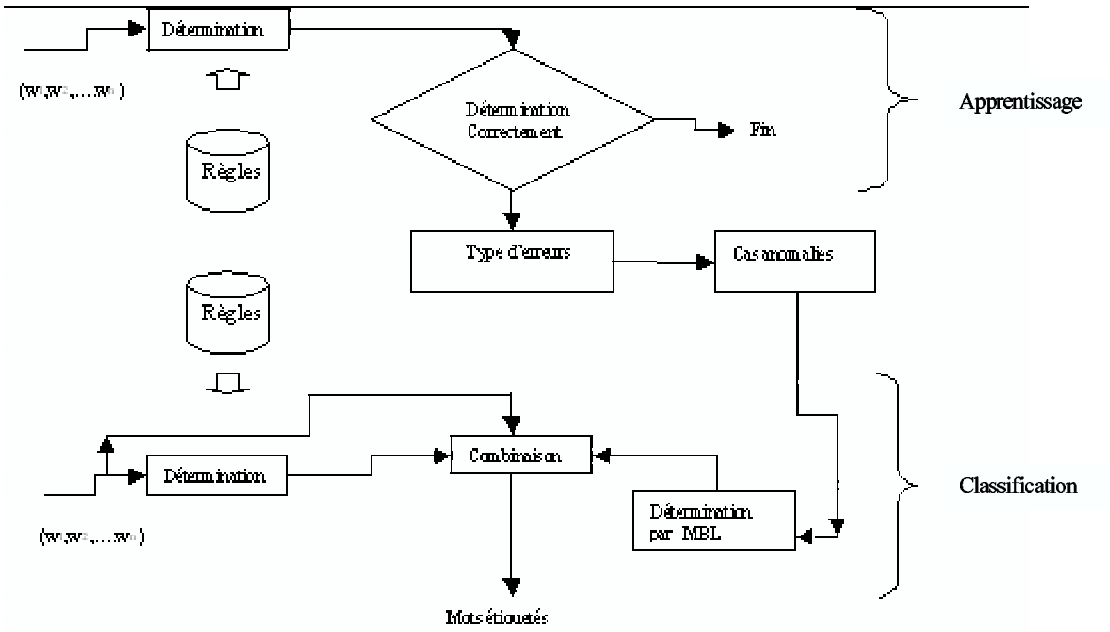


Figure 1. Architecture générale de l'étiqueteur

6 Résultats

Il suffit de comparer les différentes grammaires et dictionnaires, en particulier sur les classes de mots difficiles (adjectifs indéfinis, conjonctions, etc.), pour se convaincre de la difficulté d'établir une référence solide. La majorité des publications sur l'étiquetage automatique font l'impasse sur les cas épineux, ce qui à une conséquence directe sur l'interprétation des performances annoncées. L'application juste de l'étiquetage à base de règles donne un taux de performance de 91,87% à cause de l'ambiguïté(André, Veronis1999). Dans l'arabe les cas épineux sont nombreux, en particulier pour les noms qui ont un double rôle (nom où adjectif). Ainsi le mot de se type peut prendre une étiquette qui ne convient pas (Van Mol, 2001). Illustrons l'apport de la combinaison de la base de règles et le MBL par les exemples suivants (ces cas ambigus sont testés par le système proposé) :

Exemple 1: -جَمِيْلٌ يَشْرُبُ ici جَمِيْلٌ nom peut prendre l'étiquette suivante: NCSgMNI.
- جويٌ جَمِيلٌ ici جَمِيْلٌ adjective peut prendre l'étiquette suivante NACSgMNI.

Exemple 2: - دَخلتْ بِئْتُ ici بِئْتُ un nom mais en appliquant les règles morphosyntaxiques le nom est défini comme un verbe et prend l'étiquette suivante: VPSg1 (en langue Arabe beaucoup de noms peuvent être de ce type).

Exemple 3: مَا أَبْيضَ وَجهه ici أَبْيضَ est un nom adjectif mais par l'application des règles morphosyntaxiques il est identifié comme un verbe et prend l'étiquette suivante : VPSg3M

Exemple 4: مَدَارس, أقلام, قصور. Il existe une catégorie de pluriels qui ne peut pas être identifier comme tels (A.Goweder et al.,2002). En appliquant les règles morphosyntaxiques ce type de mot peut être identifié comme singulier.

Exemple 5 : La langue arabe est très riche en particules et notre base de particules est limitée, alors une particule peut être définit comme un nom si elle ne se trouve pas dans la base de particules et ne respectant pas les règles morphosyntaxiques exemple : شَنَّان, هَيهَات …etc.
Tout ces cas sont prisent par le système proposé et globalement les résultats attestent du gain apporté par le MBL, la figure 2 présente un pourcentage des cas anomalies, qui semble témoigner des limites de la base de règles(15% cas anomalies). Les résultats obtenus par la combinaison de base de règles et le MBL(figure 3) sont nettement supérieurs aux résultats obtenus par la méthode à base de règles. En particulier pour les cas de nom et de nom adjectif qui montrent que les exceptions des règles sont prises en charge par la méthode MBL. La remarque qu'on peut émettre, suite aux résultats obtenus, c'est que pour la langue arabe les informations lexicales jouent un rôle primordial dans la détermination du tag du mot en question.

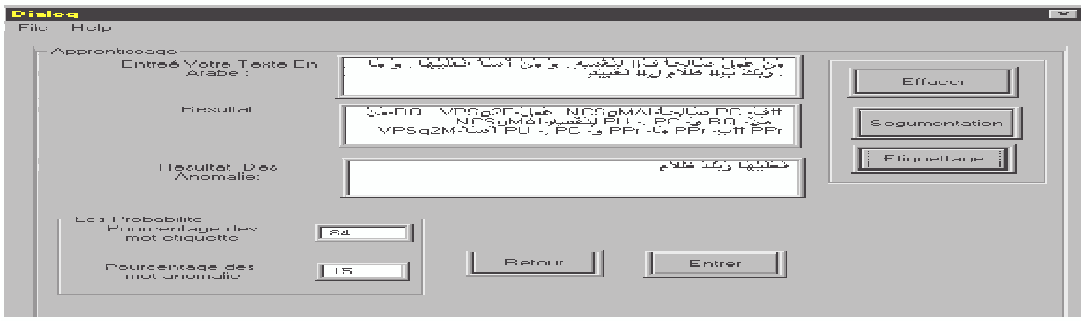


Figure 2. Résultats à base de règles

Figure 3. Résultats à base de méthode hybride

7 Conclusion

Les succès des applications de Memory-based learning dans différents domaines de recherche de bas niveau comme la reconnaissance des formes et la reconnaissance de la parole a été sans équivoque depuis l'apparition de cette approche. Néanmoins, leur utilisation pour des tâches cognitives de haut niveau telles que le traitement automatique du langage naturel a toujours suscité des remous. Dans ce contexte, le but du travail présenté dans ce papier était d'étudier la prise en charge des exceptions par les MBL plus exactement dans l'étiquetage morphosyntaxique de la langue arabe. L'étiqueteur morphosyntaxique proposé a été réalisé suite à des études approfondies sur les différentes formes d'étiquetages, qui ne sont malheureusement pas nombreux pour la langue arabe et ceux qui existe ne respectent pas les propriétés spécifique de cette langue, en appliquant à celle-ci les propriétés des langues étrangères indo-européen alors que c'est une langue sémitisée (voir MULTEXT). La comparaison nous amène à choisir l'étiquetage de sherine khoja, ce dernier est conçu spécialement pour l'arabe et l'approche hybride règle de déduction et Memory-based learning représente une performance considérable. Les perspectives envisagées pour faire évoluer le système actuel sont nombreuses. Dans un premier temps, il est possible de construire un modèle adaptatif (s'il existe des statistiques) afin de permettre le suivie de l'étiquetage et la détection avec correction automatique des erreurs. Aussi, il faut réfléchir sur l'intégration du système comme module dans un système de découpage de la phrase arabe en tronçons.

Références

Abduelbaset.,Goweder., Massimo.Poesto., Anne.De Roeck., Jeff.Reynolds.(2002); Identifying Broken Plurals in Unvowelised Arabic Text, in 2002.

Andrew.Roberts.(2003); Machine Learning in Natural language Processing, www.comp.Leeds.ac.uk.

Hacioglu K., Ward W.(2003); Target Word Detection and Semantic Role Chunking using Support Vector Machines, in *HLT-NAACL Proceedings*, pp. 25-27 , Edmonton, May 2003.

Jakub Zavrel., Walter Daelemans.(2000); Recent Advances in Memory-Based Part-of-Speech Tagging, in *Induction of Linguistic Knowledge* TSL 2000.

Mark Van Mol.(2001); The semi-automatic tagging of Arabic corpora, in *The Dutch language Union,* Amsterdam,Bulaaq, 2001.

Mona Diab., Kadri Hacioglu., Daniel Jurafsky.(2004); Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks, in *The National Science Foundation*, USA, 2004.

Saleem Abuleil., Martha Evens.(2002); Discovering Lexical Information by Tagging Arabic Newspaper Text, in *Computer and Humanities* 36(2):191-221, May 2002.

Seong-Bac Park., Byoung-Tak Zhang.(2003); Text Chunking by Combining Hand-Crafted Rules and Memory-Based Learning, in Proceedings of the *41st Annual Meeting of the Association for Computational Linguistics,* July 2003, pp 497-504.

Shereen Khoja., Roger Garside., Gerry Knowles.(2001); A tagset for the morphosyntactic tagging of Arabic, http://www.comp.lancs.au.uk/computing/users/khoja/cl2001.pdf.

Valli André., Jean Veronis.(1999) ; Etiquetage grammatical des corpus de parole : problèmes et perspectives, http://www.up.univ-mrs.fr/~veronis/pdf/1999rfla.pdf.

Walter Daelemans., Antal van den Bosch., Jakub Zavrel., Jorn Veenstra., Sabine Buchholz., Bertjan Busser.(1998), Rapid Development of NLP Modules with Memory-based Learnig, in Proceeding of *ELSNET in Wonderland,* March 1998, pp105-113.

Walter Daelemans., Jakub.Zavrel.(1996); Part-of-Speech Tagging of Dutch with MBT, in *Informatiewetenschap* 1996, pp 33-40, The Netherlands.TU Delft.

Young-suk Lee., Kishore Papineni., Salim Roukos., Langage Model Based Arabic Word Segmentation, www.acl.ldc.upenn.edu,

RÉCITAL 2005, Dourdan, 6-10 juin 2005

Cent mille milliards de poèmes et combien de sens ? Une étude d'analyse potentielle

Florentina Vasilescu Armaselu
Université de Montréal, Département de littérature comparée
armaselu@sympatico.ca

Mots-clefs – Keywords

Unité du discours, réseaux de cohésion, analyse thématique, littérature potentielle

Discourse unity, networks of cohesion, thematic analysis, potential literature

Résumé – Abstract

A partir du concept de *cohésion* comme mesure de l'*unité du texte* et du modèle oulipien de la *littérature par contraintes*, notre étude propose une méthode d'*analyse potentielle* sur ordinateur dans le cas des *Cent mille milliards des poèmes*. En s'appuyant sur un ensemble de contraintes initiales, notre programme serait capable d'analyser tous les textes potentiels produits par la machine en utilisant ces contraintes.

Using the concept of *cohesion* as a measure for the *unity of text* and the Oulipian model of the *literature by constraints*, our study proposes a computational method of *potential analysis* for *One hundred billions sonnets*. Starting from a set of initial constraints, our program would be able to analyze all the potential texts produced by the machine under these constraints.

1 Introduction

Les mécanismes de compréhension du *sens* d'un texte dans son ensemble restent encore peu connus. Des modèles du processus d'interprétation (comme compréhension, pas comme interprétation critique) ont été déjà proposés : la théorie des cadres (Minsky, 1975), la hiérarchisation des expressions anaphoriques (Lakoff, 1976), les réseaux de cohésion (Halliday, Hasan, 1976), l'hypothèse de la connectivité (Gentner, 1981). (Rastier, 1987) relie la notion de *sens* d'un texte à la modalité de perception de *l'unité du texte,* dans le processus d'interprétation. A partir de l'étude des réseaux de cohésion (Stoddard, 1991) et de la notion de cohésion lexicale (Morris, Hirst, 1991) nous proposons une nouvelle approche d'analyse sur ordinateur des réseaux de cohésion comme mesures de l'unité d'un texte, dans le cas des *Cents mille milliards de poèmes* (Queneau, 1961). Notre analyse s'appuie sur des relations d'ordre sémantique, syntaxique et cognitif qui contribuent à la perception d'un texte comme un tout cohésif. Ne disposant pas d'un dictionnaire électronique comportant ce type d'information, nous avons annoté manuellement (voir 4.1) les mots considérés significatifs (noms, verbes, adjectifs) des dix sonnets originaux de Queneau. A partir de cet ensemble d'annotations et en utilisant un mécanisme combinatoire permettant l'engendrement de nouveaux sonnets, notre programme de composition et d'analyse produit des diagrammes, des calculs estimatifs et des descriptions thématiques (voir 4.1, 4.2, 5), en simulant la « compréhension » d'un sonnet en termes d'unité thématique et de relations de cohésion établies entre les mots.

Le choix des *Cents mille milliards de poèmes* comme banc d'essai pour notre étude n'a pas été aléatoire. Premièrement, parce que le modèle oulipien de la création par contraintes et son mécanisme combinatoire (voir 2), représenteraient un point de départ pour une *analyse potentielle* du texte. Nous entendons par cela un programme qui, à partir d'un ensemble de contraintes initiales (dans notre cas, les dix sonnets originaux annotés), serait capable d'*analyser*, par des procédés combinatoires, tous les *textes potentiels* produits par la machine en utilisant ces contraintes. Deuxièmement, les dix poèmes originaux, doués, selon Queneau, d'un « thème » et d'une « continuité », joueraient le rôle de *base de comparaison* pour notre analyse. Troisièmement, en tenant compte que notre objet d'étude est une collection de sonnets, notre intérêt porte sur les enjeux d'un traitement automatique du *sens* dans le cas des textes littéraires. En d'autres mots, il s'agissait de reformuler en termes interrogatifs la bien connue citation de Turing placée par Queneau en tête de ses *Cent mille milliards de poèmes* : *Est-ce qu'une machine peut apprécier un sonnet écrit par une autre machine ?*

2 Une machine à fabriquer des sonnets

Les deux tendances majeures de la recherche oulipienne sont la reprise des œuvres du passé et l'invention de nouvelles règles de création littéraire, à partir de *contraintes formelles* (Le Lionnais, 1986). Selon (Motte W.F. Jr., 1986), les *Cent mille milliards de poèmes* représentent le modèle de l'entreprise oulipienne, par la reprise d'une forme poétique traditionnelle, le sonnet, et par l'invention d'une nouvelle forme poétique *combinatoire* permettant à chaque vers d'être intégré dans l'ensemble quasi-infini de sonnets potentiels. Comme l'affirme (Queneau, 1961), *Cent mille milliards de poèmes* est une « machine » à fabriquer 10^{14} poèmes différents. Le fonctionnement de cette machine s'appuie sur un ensemble de contraintes formelles, internes et combinatoires. Les *contraintes internes* concernent la forme de chaque sonnet (deux quatrains et deux tercets), les rimes qui ne doivent pas « être trop banales [...] trop rares ou uniques » et l'existence d'un « thème » et d'une « continuité » pour chacun des dix sonnets d'origine. Les *contraintes combinatoires* exigent une structure grammaticale invariante et l'absence des désaccords en genre et en nombre pour toute substitution de vers possible. En tenant compte de cette immense mais encore limitée *potentialité créative*, qu'est-ce qu'on pourrait dire alors sur la *potentialité de sens* de ce type de machine ?

3 Cohésion et unité du texte

Le concept de *sens* fait l'objet d'étude de plusieurs disciplines dans le cadre des sciences humaines. Comme nous avons déjà mentionné, notre démarche s'intéresse seulement à une partie plus restreinte de ce concept, reliée à la modalité par laquelle nous percevons *l'unité d'un texte* dans le processus d'interprétation (Rastier, 1987). Selon (Morris, Hirst, 1991), le texte ou le discours n'est pas une simple succession de mots et de phrases faisant référence à des choses différentes, mais un ensemble d'entités reliées l'une à l'autre qui portent sur un même sujet. C'est une propriété qui confère au texte la qualité d'*unité* et qui est appelée *cohésion.* La cohésion n'est pas pourtant une caractéristique inhérente au texte, elle dépend aussi de lecteur. Dans l'acception de (Stoddard, 1991), la cohésion est un mécanisme unificateur que nous construisons pendant le processus d'interprétation et qui nous aide à dériver beaucoup plus de sens du texte dans son ensemble que de la simple somme des sens des mots et des phrases qui le composent. La cohésion impliquerait ainsi la construction de liens mentaux entre les parties composantes d'un texte, dans le processus d'interprétation.

Il y a plusieurs types de relations déterminant la cohésion. Notre étude s'intéresse aux relations *sémantiques* existant entre les mots (partie/tout, co-occurrence dans des contextes similaires, appartenance à un même domaine) et déterminant la *cohésion lexicale* (Morris, Hirst, 1991). De plus, notre analyse s'appuie sur le modèle des *réseaux de cohésion* (Stoddard, 1991), utilisé dans l'analyse des articles définis, des pronoms et des dislocations d'agents verbaux. Le réseau de Stoddard comporte un *nœud* (le référent, par exemple *Abraham Lincoln*) et des *éléments de cohésion* (par exemple, les pronoms *he, him, his*) reliés au nœud par des *relations de cohésion* sémantico-syntaxiques. Stoddard fait ainsi une distinction entre la *cohésion*, une caractéristique sémantico-syntaxique, et la *cohérence* une mesure de « l'unité de sens d'un texte » qui entraîne « l'environnement cognitif » et « l'expérience » du lecteur. Nous avons adapté ce modèle, en considérant le réseau de cohésion comme une structure de *nœuds* (noms, verbes, adjectifs préalablement annotés) reliés entre eux par des *relations de cohésion* qui dépendent de la nature des attributs attachés aux nœuds et qui impliquent des connaissances d'ordre lexico-sémantique, syntaxique et cognitif (voir 4.1).

4 Ouvroir d'analyse potentielle

Le programme permet à la fois la construction et l'analyse d'un poème. Il y a deux modalités de construire un poème : choisir un des dix sonnets originaux en indiquant son numéro dans un champ de saisie ou composer un nouveau poème, en combinant les vers des dix sonnets à l'aide de 14 listes déroulantes. Le module d'analyse utilise les attributs attachés manuellement aux mots, comme des *contraintes initiales*. Après la composition d'un sonnet tel indiqué ci-dessus, le programme compare les attributs et construit un lien entre deux mots (nœuds), s'il y détecte au moins une valeur commune, indifféremment du type des attributs.

4.1 Les contraintes initiales

L'annotation des mots (en format XML) comporte un attribut obligatoire (le *lemme)* et un ensemble d'attributs optionnels[1] (voir Figure1), selon les quatre types de relations considérés :

1. Relations de type *sémantique,* décrites par l'attribut *domaine,* une classe sémantique reliée à « l'expérience d'un groupe » et qui encode une « pratique sociale » (Rastier, 1997). Ce type d'attribut permet de construire, par exemple, un lien entre *Tamise* et *bateaux* (domaine = navigation), *climat* et *bise* (domaine = météo), *Socrate et Platon* (domaine = philosophie), etc.
2. Relations de type *syntaxique* entre un nom et son complément (ou son attribut) et un verbe et son complément. Ces relations sont mises en évidence par l'attribut *relatif_à* associé au complément ou à l'attribut : *« climat londonien »* (londonien, relatif_à = climat) ; *« Sa sculpture est illustre »* (illustre, relatif_à = sculpture) ; « on transporte et le marbre …» (marbre, relatif_à = transporter).
3. Relations extratextuelles supposant des *connaissances du monde*, définies par les attributs *appartenance* et *allusion.* Le programme mettrait ainsi en relation *Grèce* (lemme = Grèce) avec *Platon* (appartenance = Grèce) ; *londonien* avec *Tamise* (appartenance = Angleterre) ; *Elgin* (allusion = Parthénon, Turquie) avec *Parthénon* (lemme = Parthénon) et *Turc* (appartenance = Turquie) ; *frissonner* avec *bise* (allusion = froid), etc. A la différence de (Stoddard, 1991) nous avons considéré ces connaissances du monde comme facteurs déterminant la cohésion du texte.
4. Relations *étymologiques* (*gaucho, pampa, maté* reliés par leur attribut *étymologie* = espagnol).

[1] Leurs valeurs ont été suggérées par *Le Petit Larousse, Le Grand dictionnaire terminologique* et *EURODICAUTOM.*

Comme dans le modèle oulipien, l'annotation s'appuie sur des contraintes initiales, internes et combinatoires (voir 2), supposant une interprétation au niveau de chacun des sonnets originaux et une sorte de méta-interprétation qui devrait tenir compte des combinaisons possibles des mots d'un sonnet avec les mots des autres. Il s'agissait ainsi de prévoir des valeurs d'attribut appropriées de façon que *Rameaux* (sonnet 3) soit par exemple relié à *cloche* (sonnet 1), *corne* (sonnet 1) soit relié à *taureau* (sonnet 1) et à *veau* (sonnet 5) mais pas à *chat* (sonnet 9) ou à *baleine* (sonnet 3), dans un sonnet potentiel. Un exemple de diagramme de cohésion produit par le programme est présenté ci-dessous :

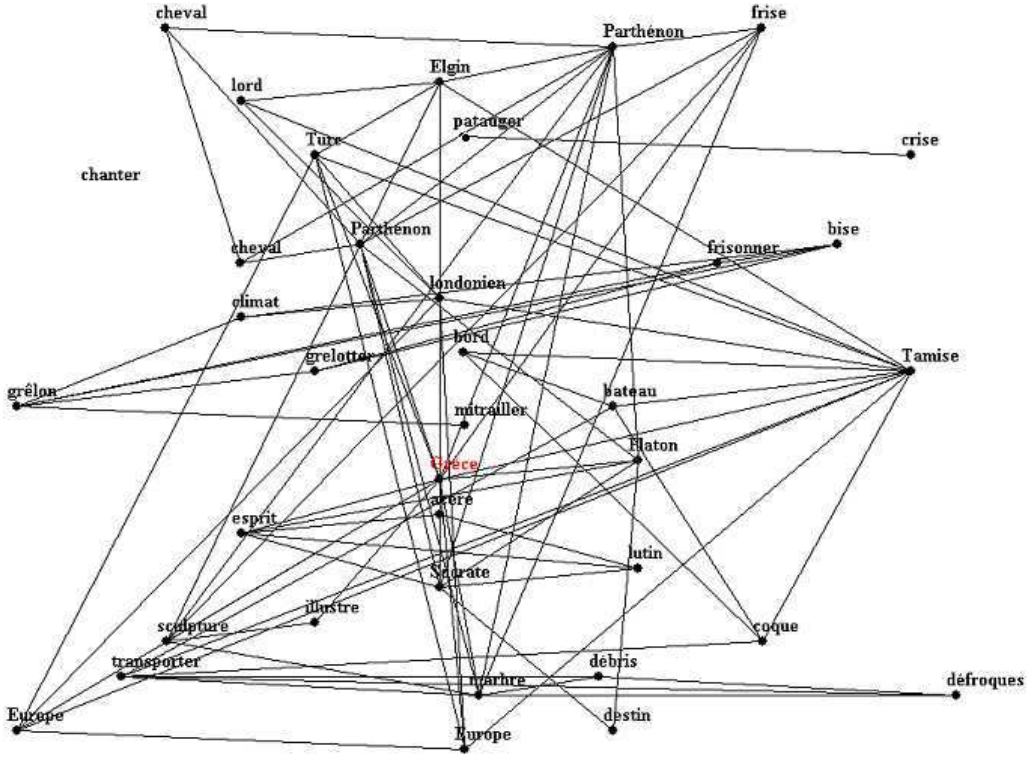


Figure 1. Réseaux de cohésion. Sonnet 2

4.2 Le banc d'essai

Les *Cent mille milliards des poèmes* nous ont servi également comme banc d'essai pour des estimations quantitatives (sur le contenu, la cohésion, le thème), considérées comme des indicateurs globaux de l'unité d'un poème. Le programme détermine la composition (en %) de chaque poème analysé, par rapport aux sonnets originaux (par exemple : 100% sonnet2 ; 50% sonnet1 +50% sonnet2, etc). C'est une mesure par laquelle on pourrait apprécier, en grandes lignes, le caractère homogène ou hétérogène d'un poème quant à sa composition. A partir de l'hypothèse qu'un texte nous paraîtrait plus unitaire si la plupart de ses mots sont reliés entre eux, nous avons défini le *coefficient de cohésion globale* (CCG) comme :

$$CCG = \frac{NRD}{TN} \cdot 100$$, où NRD = le nombre de Nœuds du Réseau de cohésion Dominant,

TN = le nombre Total des Nœuds pour le texte analysé.

Comme un texte pourrait renfermer plusieurs réseaux de cohésion indépendants (ce qui indiquerait une fragmentation de son unité), le réseau dominant regrouperait le plus grand

nombre de nœuds connectés entre eux par des liens de cohésion, pour le texte donné. Un coefficient de cohésion de 100% caractériserait ainsi un texte où tous les mots analysés soient reliés entre eux, en formant un seul réseau. Le sonnet 2 (voir Figure 1) présente un coefficient de cohésion globale assez élevé, puisque son réseau de cohésion dominant inclut presque tous les nœuds, sauf trois (*chanter, patauger, crise*). Une autre mesure, utilisée comme indicateur du degré de cohésion entre les mots, est la *densité moyenne* (DM), i.e. le nombre moyen de connexions par nœud pour un texte donné (Stoddard, 1991). Une valeur élevée de la densité signalerait une cohésion forte, une valeur basse indiquerait une cohésion faible (un texte ayant moins de relations entre ses éléments). Comme base de comparaison, le programme affiche les valeurs du CCG et de la DM des dix sonnets originaux, par ordre décroissant.

Le programme propose également un *thème,* i.e. l'entité à laquelle les mots du texte font le plus souvent référence et qui appartient au texte (valeur de l'attribut *lemme*) ou est extérieure au texte (valeur d'un attribut *domaine*, *allusion*, *appartenance*, etc). Le programme compte le nombre de fois qu'une valeur d'attribut apparaît dans l'annotation d'un sonnet, en choisissant comme thème la ou les valeurs les plus fréquentes. Pour une caractérisation plus détaillée, chaque thème est accompagné des *centres de focalisation* du sonnet, i.e. les mots comportant le plus grand nombre de connexions (par exemple, *Grèce* pour le sonnet 2, Figure 1). Cette description thématique serait une représentation condensée du sens global d'un texte (voir 5).

5 Observations sur les résultats

La Figure 2 présente les résultats d'analyse pour les dix sonnets de départ et pour dix sonnets composés. Le tableau indique une cohésion moins forte (densité plus basse) pour les sonnets composés, bien que des valeurs plus élevées soient également possibles (No 12). Le coefficient de cohésion globale présente des valeurs variables, comparables parfois avec les sonnets originaux. Les valeurs basses de cet indicateur montreraient l'existence de plusieurs réseaux de cohésion, dont aucun ne domine de façon marquante, et alors une fragmentation de contenu (14, 19). Cette caractéristique ne semble pas reliée nécessairement au caractère trop hétérogène d'un sonnet, une combinaison de tous les dix sonnets originaux pouvant déterminer des valeurs assez élevées (20). Comme le montre le tableau, les descriptions thématiques obtenues par le mécanisme combinatoire d'analyse peuvent reprendre les thèmes initiaux, avec un changement éventuel de centre de focalisation (11), engendrer de nouveaux thèmes (14, 15, 16, 19, 20) ou des thèmes composites (12, 13, 16, 17, 18). Dans ce dernier cas on pourrait avoir parfois une certaine compatibilité entre les éléments y impliqués (12, 16), parfois un caractère plus hétérogène, bien que pas tout à fait incompatible (13, 17, 18). D'un autre côté, les résultats d'analyse semblent fort dépendants de la description XML, i.e. de la subjectivité et du niveau de connaissances utilisés dans l'interprétation des sonnets originaux.

No	Sonnet/Contenu	CCG %	DM	Thème	Centres de focalisation (no. liens/centre)
			Sonnets originaux		
1	S1	*62.8*	*2.7*	*Amérique du Sud*	*Amérique du Sud; pampa (8)*
2	S2	*91.8*	*4.7*	*Europe*	*Grèce (14)*
3	S3	*72.5*	*9.5*	*mer*	*poisson; dorade; molve_lotte (20)*
4	S4	*72.5*	*2.8*	*noblesse*	*blason; baron (8)*
5	S5	*92.3*	*4.0*	*Europe*	*latin (13)*
6	S6	*96.6*	*2.8*	*urbanisme*	*escroc; provincial (6)*
7	S7	*62.0*	*3.9*	*généalogie, famille*	*généalogiste; adultérin; parent (11)*
8	S8	*92.5*	*6.5*	*langues*	*métromane (18)*
9	S9	*89.1*	*9.6*	*alimentation*	*turbot; requin (25)*
10	S10	*62.5*	*3.3*	*mort*	*mort (12)*

No	Sonnet/Contenu	CCG %	DM	Thème	Centres de focalisation (no. liens/centre)
			Sonnets composés		
11	**50%**S6,S10	*71.4*	*2.3*	*mort*	*croque-morts; tissu; pâlotte (6)*
12	**50%**S3,S9	*95.1*	*9.9*	*alimentation, zoologie*	*poisson; dorade; molve lotte (25)*
13	**50%**S1,S4	*62.1*	*2.4*	*géographie, Angleterre, noblesse*	*baron ; Malabar ; lord (6)*
14	**28.5%**S4,S8 ; **21.4%** S2,S6	*26.1*	*2.1*	*Angleterre*	*Tamise (8)*
15	**35.7%**S7,S8 ; **28.5%**S9	*70.5*	*2.5*	*psychologie*	*idiot ; métromane (6)*
16	**35.7%**S1; **21.4%**S3,S7,S2	*63.6*	*2.1*	*navigation, eau*	*marin (7)*
17	**28.5%**S5,S1 ; **21.4%**S2,S9	*65.8*	*2.6*	*Europe, zootechnie*	*taureau ; veau ; Grec (7)*
18	**14.2%**S1,S2,S3,S4,S8,S9,S10	*78.7*	*2.4*	*alimentation, emploi, transporter*	*chauffeur ; sel ; marbre ; marbrier ; pompier (5)*
19	**14.2%**S2,S4,S6,S7,S8 ; **7.1%**S3,S5,S9,S10	*35.4*	*1.9*	*art*	*aède ; poète (6)*
20	**21.4%**S4; **14.2%**S2,S5; **7.1%**S1,S3,S6,S7,S8,S9,S10	*70.2*	*2.3*	*boissons*	*Beaune, Chianti, anglais (6)*

Figure 2 : Résultats d'analyse. CCG – coefficient de cohésion globale, DM – densité moyenne

6 Conclusion

Notre étude du *sens* des *Cent mille milliards de poèmes* s'appuie sur la notion de *cohésion* comme expression de l'*unité d'un texte*. La question posée par le titre est rhétorique. L'étude complète du sens des *Cent mille milliards de poèmes* supposerait une analyse des 10^{14} sonnets potentiels, ce qui dépasse évidemment le but de notre projet. Notre démarche s'intéresserait plutôt aux enjeux d'une *analyse potentielle*, i.e. à la capacité d'un programme d'analyser un ensemble quasi-infini de textes, à partir d'un nombre fini de contraintes initiales.

Références

GENTNER D. (1981), Verb semantic structures in memory for sentences: Evidence for componential representation. *Cognitive psychology*, Vol. 13, pp. 56-83.

HALLIDAY M.A.K., HASAN R. (1976), *Cohesion in English,* London, Longman.

LAKOFF G. (1976), Pronouns and reference, *Syntax and semantics, Notes from the linguistic underground,* Vol. 7, pp. 275-335.

LE LIONNAIS F. (1986), Lipo. First Manifesto, In *Oulipo, A Primer of Potential Literature,* Translated and edited by Warren F. Motte Jr., Lincoln, University of Nebraska Press.

MINSKY M. (1975), A framework for representing knowledge. In *The Psychology of Computer Vision*, P.H. Winston Editor, New York, McGraw-Hill.

MORRIS J., HIRST G. (1991), Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of the Text, *Computational Linguistics*, Vol. 17, pp. 21-45.

MOTTE W.F. Jr. (1986), Introduction, In *A Primer of Potential Literature,* Translated and edited by Warren F. Motte Jr., Lincoln, University of Nebraska Press.

QUENEAU R. (1961), *Cent mille milliards de poèmes,* Paris, Gallimard

RASTIER F. (1987), *Sémantique interprétative,* Paris, Presses Universitaires de France.

RASTIER F. (1997), *Meaning and Textuality,* Toronto, University of Toronto Press.

STODDARD S. (1991), *Text and Texture: Patterns of Cohesion,* Norwood, Ablex Publishing.

RÉCITAL 2005, Dourdan, 6-10 juin 2005

Analyse informatique du roman proustien « *Du côté de chez Swann* »

Katia ZELLAGUI
Laboratoire le LASELDI – Université de Franche-Comté
30 Rue Mégevand - 25000 Besançon ;France;
katia.zellagui@univ-fcomte.fr

Mots-clefs – Keywords:

Automate fini, grammaire locale, dictionnaire électronique, étiquetage morpho-syntaxique, désambiguïsation, textes littéraires.
Finite state automata, local grammar, electronic dictionary, morpho-syntactic tagging, disambiguation, literary texts.

Résumé - Abstract

Dans le cadre du développement des environnements d'analyse linguistique, d'étiquetage de corpus et d'analyse statistique afin de traiter des corpus de grande taille, nous proposons de mettre au point des procédures nouvelles d'étiquetage morpho-syntaxique et sémantique. Nous présentons un ensemble de ressources linguistiques - dictionnaires et grammaires - dans le but d'étiqueter entièrement le roman proustien : « *Du côté de chez Swann* ». Notre recherche avance deux atouts majeurs : la précision des étiquettes attribuées aux formes linguistiques du texte ; et le repérage et étiquetage exhaustifs des mots composés.
To deal with a great amount of corpus data within the framework of environmental development of linguistic analysis of corpus' tagging and statistic analysis, we propose to establish new procedures of syntactic and semantic tagging. We present some general linguistic resources, such as dictionary and grammar built-in the way to entirely tag the novel of Proust «*Du côté de chez Swann*». Our research leads to two main advantages: precise tagging assigned to linguistic forms of the text and identification and exhaustive tagging of compound nouns.

Introduction

Face à l'expansion des ressources électroniques ;e-book, sites Internet, CD-Rom;, de plus en plus de données textuelles sont disponibles sur support électronique. Cette expansion implique inévitablement le développement des outils d'analyse de textes : que ce soit des analyseurs ou des logiciels d'étiquetage ;e.g. Brill, Cordial, Lexico, Hyperbase, INTEX;, et oblige les chercheurs en linguistique informatique à repenser la gestion et l'organisation des données textuelles ;e.g. Text Encoding Initiative;. Le problème des textes sur support électronique actuellement disponibles est qu'ils ne sont pas décrits suffisamment d'un point de vue lexical, syntaxique et sémantique. Dans le cadre de nos recherches, nous proposons de fournir aux chercheurs en littérature et en linguistique des procédés spécifiques d'étiquetage lexical et syntaxique de corpus littéraires, à partir desquels il sera possible de réaliser ;à court terme; des analyses linguistiques et littéraires fines et de construire ;à long terme; des hypertextes d'un nouveau type : les liens seront établis entre les unités linguistiques ;plutôt qu'entre les formes superficielles;.

1 Cadre théorique

Notre démarche s'inspire en grande partie des travaux du LADL et plus spécifiquement des travaux menés par Maurice Gross ;1975; sur le lexique- grammaire.

L'équipe qui s'intéresse essentiellement au traitement automatique et à la description à large couverture des langues naturelles constitue depuis les années 60 une base de données de descriptions linguistiques très importante sous la forme principalement de dictionnaires électroniques - les DELA ;Courtois, 1990; - et de grammaires locales ;Gross, 1997;.
Notre objectif est d'élaborer des ressources linguistiques afin de réaliser de façon automatique voire semi-automatique l'étiquetage du roman proustien « *Du côté de chez Swann* »[1]. L'étiquetage consiste à *« associer à des segments de texte, le plus souvent les 'mots', une ou plusieurs étiquettes, le plus souvent leur catégorie grammaticale voire leur lemme.* » ;Habert & al., 1997;. L'étiquetage que nous proposons de réaliser sera systématique ;toutes les formes du texte doivent avoir au minimum une étiquette lexicale et syntaxique. Une telle entreprise rencontre deux obstacles majeurs : le choix des unités linguistiques à traiter et le choix des outils d'étiquetage. Or, une des conditions sine qua non d'un étiquetage visant un taux d'erreur proche de zéro réside dans le repérage exhaustif et l'étiquetage des mots composés dès les premières phases du traitement du corpus. Nous avons donc besoin d'un outil capable d'une part, de traiter les mots composés ; et d'autre part, d'un outil permettant de créer rapidement et facilement des règles de grammaires. Notre choix s'est donc porté sur le système INTEX ;Silberztein, 1993;. Cet outil utilise les vastes bases de données de descriptions linguistiques fines développées au LADL et permet de développer des ressources linguistiques ;dictionnaires et grammaires locales;, puis de les appliquer au corpus. Après avoir présenté les problématiques liées à une telle entreprise, nous tenterons de proposer des solutions en décrivant les ressources que nous avons créées afin d'étiqueter le corpus Swann.

2 Unités linguistiques et ambiguïtés

2.1 - Les unités linguistiques

Nous choisissons de traiter deux types d'unités linguistiques : les mots simples ;ou formes simples; et les mots composés ;ou formes composées;.
En ce qui concerne la définition d'une forme simple, nous partons d'une définition formelle que nous empruntons à Courtois ;1990;. Les formes simples sont : « […] *des unités de texte définies sur l'alphabet des codes ASCII ou EBCDIC à 256 caractères, et ne comportant aucun séparateur.* » Les lettres ;éléments de base; appartiennent à un alphabet déterminé ;e.g. l'alphabet français comprend 41 lettres;.
Il n'existe pas de définition précise et admise de la composition. Gross ;1988; annonce clairement qu'il est inutile car impossible de proposer une définition unique et stable du mot composé. Il propose d'appréhender cette notion par le biais de contraintes linguistiques. Sa démarche consiste « *à montrer que le figement n'est pas une valeur absolue mais relève d'une gradation correspondante à des propriétés transformationnelles potentielles réalisées à des degrés différents* ». Il propose donc de calculer le degré de figement en terme de propriétés non observées ;Gaston, 1986;. Les listes des mots composés élaborées en grande partie à partir de ces contraintes par les chercheurs du LADL et du LLI[2] ont servi à l'élaboration du dictionnaires électroniques des mots composés : le DELAC.
La reconnaissance des mots composés ;au même titre que les mots simples; est nécessaire pour deux raisons principales :

1. ils forment une unité linguistique à part entière, e.g. *parce que, pomme de terre* ;
2. ils peuvent générer des erreurs lors de l'étiquetage s'ils ne sont pas reconnus.

[1] Le texte sur support électronique ;issu de la reconnaissance optique; a été fourni par les éditions Champions ;Paris;. Nous le nommons corpus Swann.
[2] LLI : Laboratoire de Linguistique Informatique, Paris 13.

2.2 Les ambiguïtés

La notion d'ambiguïté est fondamentale car elle pose des problèmes majeurs dans l'analyse de texte. Certaines unités de la langue ;mots simples ou mots composés; sont effectivement ambiguës et possèdent plusieurs étiquettes morpho-lexicales et syntaxiques dans les ressources d'INTEX, ce qui nécessite un travail de désambiguïsation. Par exemple, la forme simple **la** possède trois entrées dans le DELAF : *la, la.N+z1:ms:mp : **la** est un nom ; la, le.DET+z1:fs : **la** est un déterminant; la, le.PRO+z1:3fs : **la** est un pronom.*
L'ambiguïté que nous venons de décrire relève de la syntaxe. Il existe effectivement six grands types d'ambiguïtés : les ambiguïtés orthographiques, morpho-flexionnelles, morpho-dérivationnelles, syntaxiques, sémantiques, et les ambiguïtés pragmatiques ;Fuchs, 1996;.
Dans le cadre de notre recherche, nous traiterons deux types d'ambiguïtés :

1. les ambiguïtés syntaxiques qui se situent au niveau de la structure des énoncés;
2. les ambiguïtés morpho-flexionnelles qui se situent au niveau de la flexion des formes verbales, e.g. *j'aime / il aime* ;le verbe aimer est conjugué à la 1ère et 3ème personne du singulier du présent de l'indicatif ou du présent du subjonctif; ;

Une simple exploration des contextes gauche-droite de la forme suffit souvent à lever l'ambiguïté ;les grammaires locales sont alors très efficaces;. Mais certaines formes ne peuvent être désambiguïsées qu'en tenant compte du contexte syntaxique ou sémantique global, au niveau de la phrase ou même du discours ;e.g. *table ronde*;.

3 Les outils informatiques pour l'étiquetage

Nous allons à présent décrire les outils utilisés pour le traitement automatique et linguistique du corpus : les dictionnaires électroniques et les grammaires locales.
INTEX dispose de ressources lexicales. Pour le français, il s'agit des dictionnaires électroniques du LADL. Les deux dictionnaires principaux du système INTEX sont : le DELAF [dictionnaire des mots simples généré automatiquement à partir du DELAS ;Courtois, 1990;] et le DELACF [dictionnaire des mots composés généré semi-automatiquement à partir du DELACF ;Silberztein, 1989;]. Ces ressources ;une fois appliquées au corpus; permettent d'obtenir le vocabulaire complet du texte. Les ambiguïtés sont alors repérables ;i.e. une forme ambiguë possède plusieurs entrées dans le dictionnaire;.
Les grammaires locales s'avèrent être des outils très efficaces afin de gérer les ambiguïtés non résolues par consultation des dictionnaires. Elles se présentent sous forme d'automates. Dans le processus d'étiquetage, nous utiliserons des grammaires locales de désambiguïsation sous forme de transducteurs. Les transducteurs diffèrent des automates dans la mesure où ils produisent en plus de l'information « *les transitions sont étiquetées par des couples de symbole (Sr/Sp), où Sr est un symbole reconnu, et Sp un symbole produit.* » ;Silberztein, 1993;.

4 L'étiquetage du texte

La phase étiquetage se déroule en trois étapes successives : le pré-traitement du texte brut, l'étiquetage des formes composées et l'étiquetage des formes simples.

4.1 Pré-traitement

Avant la phase d'étiquetage, il faut réaliser l'étape dite de pré-traitement. Sous INTEX, cette étape consiste d'une part à segmenter le texte en phrases, ce qui passe par l'application d'une grammaire locale qui traite les ambiguïtés liées au point ;signe de ponctuation forte ou signe d'abréviation; et qui insère le symbole {S} après chaque fin de phrase ; et d'autre part, à reconnaître et étiqueter les mots composés non ambigus : e.g. *parce que, aujourd'hui* ;par le biais du dictionnaire Ucompound.dic intégré au système INTEX;.

Il s'agit ensuite d'appliquer les dictionnaires du système : le DELAF qui contient 746 214 entrées, et le DELACF qui compte 248 885 entrées. Une fois cette opération réalisée, INTEX fournit deux listes qui correspondent au vocabulaire du texte : la liste des mots simples ;47 592 entrées;, et la liste des mots composés ;3 408 entrées;. I1 est alors possible de procéder à la phase d'étiquetage, ce qui revient à gérer les ambiguïtés du texte. L'étape de désambiguïsation se déroule en deux temps : le traitement des formes composées puis celui des formes simples.

4.2 Traitement des formes composées

Le DLC contient donc 3 408 entrées, ce qui représente environ 7 240 mots composés dans le texte. Le problème majeur du traitement des mots composés réside dans le fait que certaines séquences reconnues par une consultation de dictionnaire sont en fait des séquences libres de mots simples ;ex. *bien que*;. C'est pourquoi INTEX traite tous les mots composés comme a priori ambigus. La désambiguïsation des mots composés ne peut être réalisée que par une analyse manuelle du contexte de leurs occurrences. En effet, une séquence peut être figée dans un certain contexte, et ne pas l'être dans un autre ;e.g. *Je me rappelle **bien que** je n'ai pas dormi ; cela lui ferait plus de **bien que** son lit ; […] il le disait **bien que** cela ne se fasse pas !*;.
Nous utilisons un programme interactif de convivialité d'étiquetage : DIATAG[3] afin de procéder à la levée d'ambiguïté des mots composés en contexte. DIATAG ;intégré au système INTEX; présente la forme avec son contexte ;gauche et droite; et la ou les étiquettes candidates. À partir de l'observation directe, l'utilisateur peut alors valider la bonne étiquette qui sera intégrée directement dans le texte.
Les séquences non validées seront ensuite traitées comme des séquences de mots simples.
À l'issue de cette opération, le DLC ne contient plus que **3 382** entrées et le texte possède **6 580** mots composés.

4.3 Traitement des formes simples

Une fois les formes composées étiquetées, il s'agit alors de traiter les formes simples. Ce traitement est entrepris en deux phases successives :

1. le traitement du plus grand nombre d'ambiguïtés est réalisé avec INTEX. La bibliothèque de grammaires locales de levée d'ambiguïté contient deux types de grammaires : les grammaires locales dites générales applicables à n'importe quel texte de la langue française, et les grammaires locales dites ad hoc ;i.e. grammaires spécifiques qui ne fonctionnent que sur le corpus *Swann)*. La bibliothèque ;dans son état actuel; contient une cinquantaine de grammaires locales. Ces grammaires locales présentées sous forme de transducteurs fonctionnent par reconnaissance d'information et production d'information. Nous présentons deux grammaires de désambiguïsation. La grammaire locale *forme s.grf* ;cf. Figure 1; fonctionne sur tous les textes de la langue française.

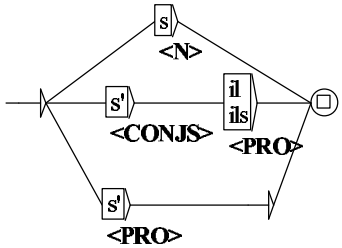


Figure1 : Grammaire locale de désambiguïsation : forme s.grf.

[3] http://www.nyu.edu/pages/linguistics/intex/#diatag

Le premier chemin analyse la forme *s* comme étant un nom ;et impose la contrainte <N> ; la forme *s'* suivie des pronoms *il* et *ils*, est systématiquement une conjonction ; dans tous les autres cas, le *s'* est un pronom. 944 ambiguïtés sont résolues en appliquant cette grammaire.
La grammaire locale *je plus V.grf* ;cf. Figure 2; appartient à l'ensemble des grammaires locales ad hoc. Elle permet de désambiguïser un grand nombre de pronoms ;e.g. *le, la, nous, lui*; et de préciser la flexion du verbe.

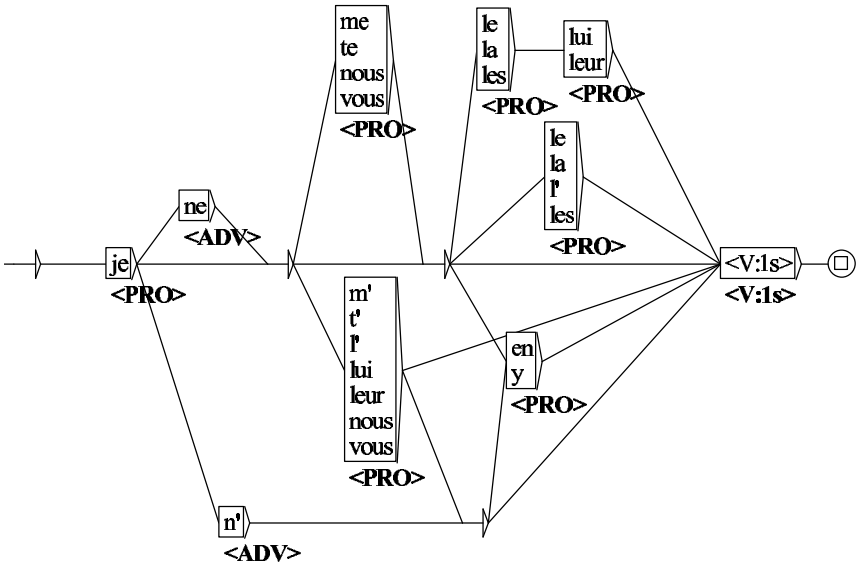


Figure 2 : Grammaire locale de désambiguïsation : je plus verbe.grf.

Le graphe *je plus verbe.grf* reconnaît 1 565 occurrences dans le corpus et permet de lever 1 378 ambiguïtés, par exemple : *{je,.PRO+PpvIL+z1} {me,me.PRO+z1} {suis,être.V+aux+z1:P1s} {couché,coucher.V+se+p+i+E+z1:Kms}*.
Les grammaires locales que nous avons construites ont permis de lever plus de 80% des ambiguïtés de façon automatique.

2. Traitement des ambiguïtés résiduelles avec DIATAG

Face à l'impossibilité de lever certaines ambiguïtés de façon automatique, nous entreprenons une partie de l'étiquetage des formes simples de façon manuelle à l'aide du programme DIATAG. La démarche est alors similaire à celle que nous avons suivie pour le traitement des formes composées. Prenons l'exemple de la forme **que** ;qui possède cinq étiquettes syntaxiques : pronom relatif ou interrogatif, conjonction de subordination, introducteur ou adverbe;. Compte tenu de la complexité de certaines phrases proustiennes, l'étiquetage de cette forme ne peut être réalisé qu'en se référant au contexte. Ce type d'ambiguïté est donc résolu au cas par cas avec DIATAG. Nous avons eu recours à ce procédé essentiellement afin de traiter les mots grammaticaux ;Dister, 2001;. À l'issue de ce traitement, le DLS ne contient plus que **16 372** entrées.

5 Résultats

Nous avons développé un environnement qui a permis de lever plus de 80% des ambiguïtés du texte de façon automatique. Ces ressources comprennent des dictionnaires électroniques et des grammaires locales de levées d'ambiguïté[4]. Le vocabulaire final du texte contient 164 060 mots

[4] Les grammaires locales de désambiguïsation sont disponibles sur le site INTEX.

simples et 6 580 mots composés. Afin de réduire les erreurs d'étiquetage, nous concevons une chaîne de traitements incluant des procédures de contrôle qualité ;permettant de détecter le maximum d'erreurs;. La qualité de l'étiquetage est ensuite évaluée en comparant le corpus étiqueté avec la version étiquetée proposée par FRANTEXT[5]. Notre objectif étant d'évaluer la qualité de l'étiquetage, nous précisons que nous comparons dans ce cas précis les résultats en ayant conscience du fait que les méthodes d'étiquetage sont différentes. Nous sélectionnons de façon aléatoire 10 paragraphes d'environ 200 mots chacun. Nous relevons dans la version FRANTEXT plusieurs erreurs dues essentiellement au fait que les mots composés n'aient pas été reconnus ;*« cette/DTN:sg femme/SUB:sg que/SUB$ j'/PRV:sg avais/ACJ:sg quittée/VPAR:sg '/' il/PRV:sg y/PRV:++ avait/ACJ:sg quelques/DTN:pl moments/SUB :pl* ***à/PREP peine/SUB:sg*** *:/: »*[6]*)*
Un objectif futur serait d'appliquer ces ressources et notre méthodologie à un corpus plus conséquent ; nous pensons bien entendu à la totalité de *La Recherche*. En s'appuyant sur l'expérience que nous avons acquise au cours de ce projet, nous pensons qu'un travail d'étiquetage de l'ensemble de *La Recherche* pourrait être effectué par un chercheur en quelques mois.

Références

COURTOIS B. ;1990;, « Un système de dictionnaires électroniques pour les mots simples du français », In *Dictionnaires électroniques du français*, Éditions Blandine Courtois et Max Silberztein, Langue Française, n°87, Paris : Larousse, pp. 11-22.
DISTER A. ;2001;, « Levée d'ambiguïté sur les mots lexicaux et grammaticaux », fascicule spécial, In *Description et levée des ambiguïtés*, Éditions Éric Laporte, Linguisticae Investigationes, Amsterdam/Philadelphia, John Benjamins, pp. 105-126.
FUCHS C. ;1996;, *Les ambiguïtés du français*, Collection l'Essentiel Français.
GROSS G. ;1988;, « Degré de figement dans les noms composés », In *Les locutions figées*, Éditions Laurence Danlos, Langages, n° 90, Paris : Larousse, pp. 57-72.
GROSS G. ;1990;, « Définition des noms composés dans un lexique-grammaire », In *Dictionnaires électroniques du français*, Éditions Blandine Courtois et Max Silberztein, Langue Française, n° 87, Paris: Larousse, pp. 84-90.
GROSS Gaston ;1996;, *Les expressions figées en français* : *noms composés et autres locutions*, Paris : Ophrys.
GROSS M. ;1975;, *Méthodes en syntaxe*. Paris : Hermann.
GROSS M. ;1997;, «The construction of local grammars », In *Finite-State Language Processing*, E. Roche and Y. Schabes ;eds.;, Cambridge, Mass/London, England: MIT Press, pp. 329-354.
GROSS M. ;2001;, « Les ambiguïtés », In *Description et levée des ambiguïtés*, Éditions Éric Laporte, Linguisticae Investigationes, vol. 24, fascicule 1, John Benjamins Publishing Company, pp. 3-41.
HABERT B., NAZARENKO A., SALEM A. ;1997;, *Les linguistiques de corpus*. Collection U Linguistique. Paris : Armand Colin/Masson.
SILBERZTEIN M. ;1990;, « Le dictionnaire électronique des noms composés », In *Dictionnaires électroniques du français*, Éditions Blandine Courtois et Max Silberztein, Langue Française, n° 87, Paris: Larousse, pp. 71-84.
SILBERZTEIN M. ;1993;, *Dictionnaires électroniques et analyses automatiques de textes* : *le système INTEX*, Paris : Masson.

[5] Version étiquetée par Josette Lecomte à l'INaLF – mars 2002.
[6] Nous avons souligné les erreurs d'étiquetage.

Un étiqueteur sémantique des énoncés en langue arabe

Anis Zouaghi (1), Mounir Zrigui (2) et Mohamed Ben Ahmed (3)

(1) et (2) Laboratoire RIADI (unité de Monastir) - Université du centre
Faculté des Sciences de Monastir - Tunisie
(1) Anis.Zouaghi@riadi.rnu.tn
(2) Mounir.Zrigui@fsm.rnu.tn
(3) Laboratoire RIADI - Université de la Mannouba
Ecole Nationale Supérieure d'Informatique de la Mannouba - Tunisie
Mohamed.BenAhmed@riadi.rnu.tn

Mots-clefs – Keywords

Modèles statistiques de langage – Modèles n-classes – Décodage sémantique – Approche componentielle et sélective.

Statistical models of language – Models N-classes – Semantic analyze – Componential and selective approach.

Résumé – Abstract

Notre article s'intègre dans le cadre du projet intitulé Oréodule: un système de reconnaissance, de traduction et de synthèse de la parole spontanée. L'objectif de cet article est de présenter un modèle d'étiquetage probabiliste, selon une approche componentielle et sélective. Cette approche ne considère que les éléments de l'énoncé porteurs de sens. La signification de chaque mot est représentée par un ensemble de traits sémantiques Ts. Ce modèle participe au choix des Ts candidats lors du décodage sémantique d'un énoncé.

The work reported here is part of a larger research project, Oréodule, aiming at developing tools for automatic speech recognition, translation, and synthesis for the Arabic language. This article focuses on a probabilistic labelling model, according to a componential and selective approach. This approach considers only the elements of the statement carrying direction. The significance of each word is represented by a whole of semantic features Ts. This model takes part in the choice of the Ts candidates at the time of the semantic decoding of a statement.

1 Introduction

Depuis quelques années, La tendance est vers l'utilisation des modèles de langages statistiques dans le domaine de la compréhension automatique de la parole spontanée (Bousquet, 2002), (Lefèvre, 2002), etc. Pour la langue arabe, l'utilisation de tels modèles à notre connaissance constitue une nouveauté. L'avantage principal de ces modèles statistiques par rapport aux modèles à syntaxe fixe (Bennacef et al., 1994) est qu'ils sont plus portables vers d'autres domaines (Minker, 1999), et nécessite moins de recours à un expert humain. Dans cet article, nous proposons un étiqueteur sémantique basé sur un modèle de langage probabiliste hybride [Zouaghi et al., 2005] pour l'interprétation d'une séquence de mots reconnue par le module de reconnaissance de la parole. Ce modèle participe au choix des ensembles de traits sémantiques Ts candidats, en tenant compte des données suivantes: le type d'acte illocutoire accompli par l'énoncé (demande, refus, excuse, etc.), le type de l'énoncé (demande de réservation, de tarifs, etc.), des mots déjà interprétés (les traits sémantiques utilisés), et de la probabilité d'interprétation d'un mot par un Ts candidat.

2 Modèle probabiliste

2.1 Corpus d'apprentissage

Le corpus d'apprentissage considéré décrit des demandes de renseignements ferroviaires, en langue arabe classique. Chaque mot significatif dans ce corpus se voit attribuer un ensemble de traits (Ts), tel que défini dans (Zouaghi et al., 2004). Le mot الذاهب (qui va) par exemple se voit attribuer Ts = (Transport_Ferroviaire, Mouvement, Destination). Les mots synonymiques ou possédant un même rôle sémantique sont interprétés via un même Ts. Pareil, pour les mots dérivés à partir d'une même racine morphologique et possédant un même sens (tels que الذاهب (qui va) et يذهب (va) qui sont dérivés à partir de la racine ذهب (dhahaba)). Nous avons utilisé une quarantaine de Ts différents pour l'étiquetage du corpus. En plus chaque énoncé de ce corpus se voit attribuer une étiquette permettant de préciser le type de l'énoncé. En tout, nous avons utilisé sept étiquettes.

Domaine	Taille (Mo)	Nombre d'énoncés	Nombre de mots	Nombre de locuteurs
Renseignements ferroviaires	*3,4*	*10000*	*85900*	*1000*

Figure 1 : Caractéristiques du corpus de point de vue de son volume.

Nature de la tâche	Renseignements sur les:				Réservations	autres
Taux de sa représentation	horaires	trajets	tarifs	durées	*10,41 %*	*40,64%*
	28,7 %	*9,37 %*	*16,66 %*	*3,12 %*		

Figure 2 : Caractéristiques du corpus de point de vue de son contenu.

Ce corpus a été collecté en demandant à cent personnes de formuler des énoncés relatifs aux renseignements ferroviaires. Donc c'est un corpus simulé et non pas réel. L'inconvénient de ce type de corpus est qu'il ne permet pas de décrire parfaitement l'application.

2.2 Principe du décodage sémantique

Nous entendons par décodage sémantique d'un énoncé, l'étiquetage de chacun de ses mots significatifs via un Ts (Zouaghi et al., 2004). Seulement les mots porteurs de sens parmi ceux qui sont reconnus sont interprétés.

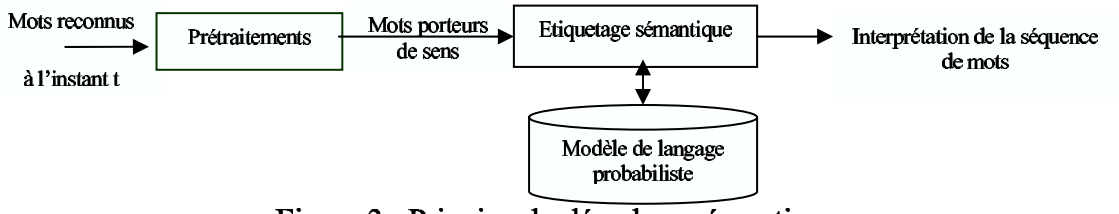


Figure 3 : Principe du décodage sémantique.

Soit la séquence de mots significatifs S = M1 M2 M3 M4 obtenus après la phase de prétraitement (figure 3). Soit Ts1, Ts2 et Ts3 les traits affectés respectivement aux mots M1, M2 et M3. A partir de ces données, nous voulons déterminer le Ts correspondant à M4. Pour atteindre cet objectif, nous utilisons un modèle de langage probabiliste hybride, permettant de tenir compte du type et de la nature de l'énoncé, ainsi que des mots déjà interprétés (figure 4).

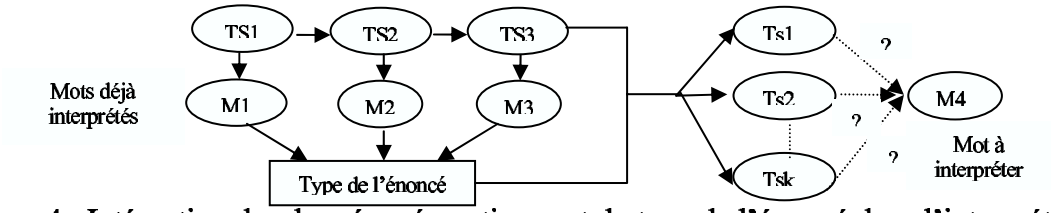


Figure 4 : Intégration des données sémantiques et du type de l'énoncé dans l'interprétation.

2.3 Description du modèle

Les systèmes à base de modèles de langage probabilistes tentent de déterminer le score d'une séquence de mots $S = m_1, m_2, ..., m_i$, dont la formule générale est la suivante:

$P(S) = P(m_1).P(m_2/m_1) ... P(m_i/m_1, m_2, ..., m_{i-1})$ (1)

Dans le cas de l'étiquetage I d'une séquence de mots significatifs $M_1 \ldots M_n$, par $Ts_1 \ldots Ts_n$, le modèle tente de déterminer le score d'interprétation de chacun de ces mots, par chacun de ces traits. Soit $I=Ts_1 \rightarrow M_1 \ldots Ts_n \rightarrow M_n$, la vraisemblance de I est alors définie comme suit:

$P(I) = P(Ts_1...Ts_n|M_1... M_n) = P(Ts_1/M_1...M_k).\ P(Ts_2/Ts_1, M_1...M_k)...\ P(Ts_n/Ts_1...\ Ts_{n-1}, M_1...\ M_n) = P(Ts_1/M_1)\ .\ P(Ts_2/\ Ts_1,\ M_2)...\ .P(Ts_n\ /\ Ts_1...\ Ts_{n-1},\ M_n)$ (2)

Nous signalons que le passage de la deuxième à la troisième ligne correspond à une approximation du modèle, qui considère que la probabilité d'un Tsi ne dépend, conditionnellement à la séquence complète des traits, qu'au mot courant Mi. En fixant à l'avance le domaine de l'application, chaque mot significatif M_i peut être interprété via $Ts_i = (C_i, TM_i)$, où Ci indique la classe à laquelle appartient le mot M_i, et TM_i le trait micro sémantique qui lui correspond. L'équation (2) devient:

$P(I)=P((C_1,TM_1)/M_1).P((C_2,TM_2)/(C_1,TM_1),M_2)..P((C_n,TM_n)/(C_1,TM_1)...(C_{n-1},TM_{n-1}),M_n)$ (3)

Nous avons intégré dans l'équation (4) d'autres sources d'informations afin d'améliorer la qualité du décodeur sémantique. Ceci, en tenant compte du type de l'énoncé noté par NT_j (avec P ($NT_j/M_1 \ldots M_n$) est la probabilité conditionnelle d'avoir un énoncé de type NT_j).
$P(I) = P(Ts_1 \ldots Ts_n | NT_j, M_1 \ldots M_n) = P(NT_j/M_1 \ldots M_n).\ P((C_1, TM_1)/NT_j, M_1) \ldots P((C_n, TM_n)/NT_j, (C_1, TM_1) \ldots (C_{n-1}, TM_{n-1}), M_n)$ *(4)*

2.4 Lissage du modèle

La première approximation appliquée à ce modèle consiste à ne considérer pour la détermination du type de l'énoncé, que certains mots appelés *mots de référence* notés Mr_k. Les mots de référence sont des mots dont leurs occurrences dans un énoncé permettent de déterminer son type. Ces mots sont en fait des uni-grammes, ou des bi-grammes, et dans certains cas des tri-grammes, dont la probabilité est égale à un. Par exemple le bi-gramme القطار (le train) أريد (je veux) constitue un mot de référence, permettant d'identifier les énoncés de type réservation. Ce bi-gramme ne peut être rencontré que dans les énoncés du corpus d'apprentissage qui sont étiquetés par l'étiquette <Réservation> (on a: P (القطار (le train)/أريد (je veux))=1). On obtient ainsi la substitution suivante:
$P(NT_j / M_1 \ldots M_n) = P(NT_j / Mr_k)$ *(5)*
La deuxième hypothèse de modélisation porte sur les relations d'indépendance conditionnelle dans le modèle et concerne la probabilité jointe $P((C_i,TM_i)/NT_j,(C_1,TM_1)\ldots(C_{i-1},TM_{i-1}), M_i)$.
$P((C_i,TM_i)/NT_j,(C_1,TM_1)\ldots(C_{i-1},TM_{i-1}),M_i) =$
$P(C_i / NT_j, (C_1, TM_1) \ldots (C_{i-1}, TM_{i-1}), M_i).\ P(TM_i / Ci, NT_j, (C_1,TM_1)\ldots(C_{i-1}, TM_{i-1}), M_i)$ *(6)*
Afin de simplifier ce modèle, nous avons considéré seulement les Ts jugés pertinents TsP (CP, TMP) à la prédiction du Ts correspondant au mot M_i noté par $Ts(M_i)$. Un Ts n'est considéré pertinent, que lorsqu'il est suivi par un nombre k minime de Ts (k tend vers 1). Nous avons fixé k = 3, car nous pensons que pour k= 1, la grammaire devient très rigide et ça revient à considérer dans l'historique du mot M_i que les mots jouant le rôle de marqueurs (Fillmore, 1968). Par exemple, l'ensemble de traits Ts = (مؤشر_حركة (Indice_mouvement), مؤشر_وجهة (Indice_destination)) est un Ts pertinent car cet ensemble est toujours succédé dans le corpus d'apprentissage par Ts = (مدينة (ville), وجهة (destination)). k correspondant à Ts=(Indice_mouvement, Indice_destination) est ainsi égal à 1. La deuxième approximation considérée est que C_i à un instant t, ne dépend que des classes pertinentes précédentes CP et du type de l'énoncé, on a ainsi:
$P(C_i / NT_j, (C_1, TM_1) \ldots (C_{i-1}, TM_{i-1}), M_i) = P(C_i/NT_j, CP_{i-1}, \ldots, CP_{i-l})$ *(7)*
Une autre approximation considérée, est que TM_i du $Ts(M_i)$, à un instant t, ne dépend que de la classe C_i affectée à M_i et du trait pertinent précédent $TsP_{i-1}(CP_{i-1}, TMP_{i-1})$. Ainsi on a:
$P(TM_i / Ci, NT_j, (C_1, TM_1) \ldots (C_{i-1}, TM_{i-1}), M_i) = P(TM_i / Ci, CP_{i-1}, TMP_{i-1})$ *(8)*
A partir de ces deux approximations (7) et (8), on déduit de l'équation (6) que:
$P((C_i,TM_i)/NT_j,(C_1,TM1)\ldots(C_{i-1},TM_{i-1}),M_i) = P(C_i/NT_j,CP_{i-1}\ldots CP_{i-l}).P(TM_i/Ci,CP_{i-1},TMP_{i-1})$ *(9)*
Et enfin à partir des équations (5) et (9), on déduit à partir de l'équation (4) que:
$P((C_i,TM_i) \rightarrow M_i/NT_j) = P((C_i,TM_i)/M_i,NT_j)$
$= P(N_j / Mr_k).\ P(C_i/NT_j, CP_{i-1}, \ldots, CP_{i-l}) \ldots P(TM_i / Ci, CP_{i-1}, TMP_{i-1})$ *(10)*

2.5 Applicabilité du modèle à la langue arabe

Comme signalé ci-dessus, nous avons considéré une approche sélective (voir paragraphe 2.2).

Dû aux spécificités de la langue arabe, on peut s'interroger sur l'adéquation de cette approche au traitement de cette langue. L'absence de la voyellation est l'une des sources d'ambiguïtés majeure de la compréhension de cette langue. Pour mieux comprendre, le mot non voyellé ذهب (thalhab) par exemple peut avoir deux significations différentes selon la manière de sa voyellation. Il a le sens du verbe partir en le prononçant (thahaba), et de l'or lorsqu'il est prononcé (thahabon). Ce mot peut être ainsi interprété par Ts= (حركة (Mouvement), وجهة (destination)), ou par Ts=((métal) معدن , ثمين (cher)). Or la détermination de la voyellation correspondante à un mot (et par conséquent son sens), nécessite plusieurs niveaux de connaissances: morphologiques, syntaxiques, ... (Debili et al., 2002). Cette nécessité est surmontée dans notre cas par la nature du domaine restreint de l'application. Nous prospectons d'améliorer la performance de l'étiqueteur, en lui intégrant des données syntaxiques.

3 Application du modèle

Nous avons utilisé une centaine d'énoncés (différents du ceux du corpus d'apprentissage), portant tous sur des demandes d'horaires pour le test. Le corpus d'apprentissage a été étiqueté avec 37 Ts. Pour juger de la qualité de notre étiqueteur, nous avons calculé le pourcentage d'étiquettes sémantiques qui sont incorrectement attribuées, à partir de la formule suivante: $Taux\_erreur=N_{inc}/N\times100$. Où, N_{inc} est le nombre de Ts incorrectement attribués, et N est le nombre total des Ts attribués par un expert au corpus de test. N est égal à 500 dans ce test. La table ci-dessous montre les Taux_erreur des étiqueteurs sémantiques obtenus en considérant des modèles bi-classes et tri-classes ainsi que le modèle hybride défini. La longueur de l'historique est fixée à 3 pour la détermination des C_i et à 2 pour TM_i.

Etiqueteurs sémantiques considérés	Taux_erreur
bi-classes: (1) $P((C_i, TM_i) / M_i) = P(C_i / C_{i-1})) \times P(TM_i / Ci, Ts_{i-1})$	57%
avec considération lexicale: (2) $P((C_i, TM_i) / M_i) = P(C_i / M_{i-1}, C_{i-1})) \times P(TM_i / M_i, Ci, Ts_{i-1})$	45%
tri-classes: (1) $P((C_i, TM_i) / M_i) = P(C_i / C_{i-1}, C_{i-2})) \times P(TM_i / Ci, Ts_{i-1})$	48,6%
(2) $P((C_i, TM_i) / M_i) = P(C_i / M_{i-1}, C_{i-1}, C_{i-2})) \times P(TM_i / M_i, Ci, Ts_{i-1})$	41,2%
hybride k=2: (1) $P((C_i, TM_i)/M_i, NT_j) = P(NT_j) \times P(C_i/NT_j, CP_{i-1}, CP_{i-2})) \times P(TM_i / Ci, TsP_{i-1},)$	50%
(2) $P((C_i, TM_i)/M_i, NT_j) = P(NT_j) \times P(C_i/NT_j, M_{i-1}, CP_{i-1}, CP_{i-2})) \times P(TM_i / M_i, Ci, TsP_{i-1})$	39,4
hybride k=3: (1) $P((C_i, TM_i) / M_i, NT_j) = P(NT_j) \times P(C_i / NT_j, CP_{i-1}, CP_{i-2})) \times P(TM_i / Ci, TsP_{i-1})$	46,8
(2) $P((C_i, TM_i)/ M_i, NT_j) = P(NT_j) \times P(C_i / NT_j, M_{i-1}, CP_{i-1}, CP_{i-2})) \times P(TM_i / M_i, Ci, TsP_{i-1})$	37%

Figure 5 : Taux d'erreur des étiqueteurs sémantiques considérés.

4 Interprétation des résultats

D'après la table ci-dessus, chaque fois que l'on intègre des données lexicales dans un modèle, le résultat s'améliore. Nous avons utilisé l'approche de (Katz, 1987) pour l'estimation des données manquantes. L'amélioration est encore meilleure, en considérant en même temps le type de l'énoncé et les Ts pertinents, pour la prédiction du Ts suivant. Nous remarquons que malgré l'amélioration de la qualité de l'étiqueteur sémantique, le taux d'erreur (qui atteint 37%) reste comme même un peu élevé. Ceci est dû au fait, que certains énoncés du corpus de test ont une structure syntaxique très complexe. Afin de remédier ce problème, certains

systèmes combinent une analyse syntaxique profonde avec une analyse sélective tel que le système TINA de (Seneff, 1992). D'autres systèmes utilisent les stratégies d'analyses du TAL robuste (Antoine et al., 2003). Ces systèmes sont performants dans des applications ouvertes.

5 Conclusion

Nous avons présenté dans cet article un étiqueteur sémantique basé sur un modèle de langage hybride. Ce modèle permet d'intégrer des données contextuelles lexicales, sémantiques ainsi qu'illocutoire en même temps. Il permet en plus de ne tenir compte que des traits sémantiques pertinents dans l'historique du mot à interpréter. Afin de montrer l'avantage de ce modèle, nous l'avons évalué et comparé par rapport aux modèles n-classes classiques, qui ne tiennent pas compte de la nature et du type de l'énoncé dans le calcul de la probabilité d'interprétation d'un mot par un Ts donné.

Références

Antoine J-Y., Goulian J., Villaneau J. (2003), Quand le TAL robuste s'attaque au langage parlé: analyse incrémentale pour la compréhension de la parole spontanée, Actes de *TALN*.

Bennacef S., Bonneau-Maynard H., Gauvain J-L., Lamel L., Minker W. (1994), A spoken language for information retrieval, Actes de *ICSLP*, 1271-1274.

Bousquet-Vernhettes C. (2002), *Compréhension robuste de la parole spontanée dans le dialogue oral homme-machine – Décodage conceptuel stochastique*, Thèse de l'université de Toulouse III, 84-85.

Débili F., Achour H., Souici E. (2002), La langue arabe et l'ordinateur: de l'étiquetage grammatical à la voyellation automatique, *Correspondances de l'IRMC,* N° 71, 10-28.

Fillmore C. J. (1968), *The case for case,* Hollt and Rinehart and Winston Inc.

Katz S.M. (Katz, 1987), Estimation of probabilities from sparse data for the language model component of a speech recognizer, *IEEE Transactions on Acoustics, Speech and Signal Processing*, 400-401.

Lefèvre F. (2000), *Estimation de probabilité non paramétrique pour la reconnaissance markovienne de la parole,* Thèse de l'Université Pierre et Marie Curie.

Minker W. (1999), *Compréhension automatique de la parole spontanée*, Paris, L'Harmattan.

Seneff S. (1992), Robust parsing for spoken language systems, Actes de *ICASSP*, 189-192.

Zouaghi A., Zrigui M., Ben Ahmed M. (2004), Une structure sémantique pour l'interprétation des énoncés en langue arabe, Actes de *JEP-TALN-ARABIC*.

Zouaghi A., Zrigui M., Ben Ahmed M. (2005), A statistical model for semantic decoding of Arabic language statements, Actes de *NODALIDA*.

TALN

RECITAL