Dourdan, France
Du 6 au 10 juin 2005

Tome 2

Ateliers

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

TALN 2005

12ème conférence annuelle
sur le
Traitement Automatique des Langues Naturelles

ATELIERS

Sommaire

Atelier EASy

Atelier EQueR

Atelier Fouille de Textes DEFT

Atelier Traitement des Langues Peu Dotées

Atelier Traitement Automatique de la Langue des Signes

TALN 2005

12ème conférence annuelle
sur le
Traitement Automatique des Langues Naturelles

POSTERS INVITÉS

CAMPAGNE D'ÉVALUATION ANALYSE SYNTAXIQUE

TECHNOLANGUE-EVALDA-EASY

EASy : Campagne d'évaluation des analyseurs syntaxiques

Patrick Paroubek (1), Louis-Gabriel Pouillot (2),
Isabelle Robba (1), Anne Vilnat (1)
(1)LIMSI - CNRS
Université d'Orsay, BP 133
91403 Orsay CEDEX
{prénom.nom}@limsi.fr
(2)Evaluations and Language resources Distribution Agency (ELDA)
55-57, rue Brillat Savarin 75013 Paris
pouillot@elda.org

Mots-clefs : Évaluation, analyseurs syntaxiques, annotation en syntaxe

Keywords: Evaluation, syntactic parsers, syntactic annotation

Résumé

Cet article présente la campagne d'évaluation des analyseurs syntaxiques de français EASy. Nous détaillons la mise en place de la campagne, ainsi que son déroulement jusqu'à ce jour.

Abstract

This paper presents EASy, the evaluation campaign of syntactic parsers of French. We detail the ongoing of the campaign, and its current state.

1 Introduction

Le projet EASY, fait partie de la campagne d'évaluation EVALDA du programme Technolangue[1], il a débuté en janvier 2003. Son ambition est de mettre au point un protocole complet d'évaluation des analyseurs syntaxiques, allant de la constitution d'un large corpus hétérogène et de l'annotation d'une partie de ce corpus, à l'évaluation des analyseurs participants et à la publication d'un large corpus entièrement annoté et validé(grâce à la fusion des sorties des analyseurs participants). Les réflexions autour du choix de ce qui serait annoté ont été menées avec

[1]Le programme Technolangue est à l'initiative du ministère délégué à la Recherche et aux Nouvelles Technologies

les participants et les fournisseurs de corpus, elles nous ont conduit à écrire un guide complet d'annotation disponible en ligne à l'adresse : http://www.limsi.fr/Recherche/CORVAL/easy.

Un des aspects importants pour nous en tant qu'organisateurs étaient de n'écarter la participation d'aucun analyseur : tout participant prêt à transcrire les sorties de son analyseur dans le formalisme retenu dans EASy devait avoir la possiblité de soumettre son analyseur pour l'évaluer.

Dans la campagne EASy, cinq fournisseurs se sont engagés à annoter en constituants et en relations 60 000 mots au total. Quatre d'entre eux sont des organismes de recherche : l'ATILF (Université Nancy 2), le DELIC (Université de Provence), le LLF (Université Paris 7) et le STIM-APHP (Paris). Le cinquième est ELDA (Agence pour l'évaluation et la distribution des ressources linguistiques) qui est co-organisateur avec le LIMSI (groupe LIR) de la campagne EASy.

Le DELIC est le seul fournisseur de corpus qui présente un article dans cet atelier. Ils ont eu la difficile tâche d'annoter des retranscriptions de l'oral, qui posent comme on le sait des problèmes spécifiques et délicats à résoudre, en particulier pour ce qui concerne la segmentation en mots et en énoncés.

Au moment de l'évaluation, treize équipes étaient à même de participer, certaines soumettant plusieurs jeux de sorties. Au final, nous avons recueilli seize ensembles de résultats. Parmi ces équipes on compte neuf laboratoires ou organismes de recherche et développement et quatre entreprises privées. Dans la liste qui suit nous avons marqué d'une étoile les équipes ayant soumis un article dans les actes de cet atelier : le CEA-LIST*, l'ERSS*, France Telecom R&D*, le GREYC*, l'INRIA-Rocquencourt (équipe ATOLL)*, le LATL (Université de Genève)*, le LIRMM, le LORIA*, le LPL*, PERTIMM, Synapse Dévelopement, Tagmatica*, Xerox Research Centre Europe.

La suite de l'article expose en détail la campagne EASy étape par étape et se termine par une présentation de quelques travaux en annotation de corpus et en évaluation d'analyseurs syntaxiques.

2 Présentation de EASy

La campagne EASy a fait l'objet d'une étude préalable (Gendner, 2003) au sein du groupe LIR du LIMSI, elle s'est déroulée sur une période de 2 ans et a permis de proposer une infrastucture contenant les éléments suivants :

1. La constitution d'un corpus d'1 million de mots, composé de textes hétérogènes en genre : des articles de journaux, des extraits de romans, des recueils de questions, des transcriptions de l'oral, des extraits de sites Internet. Ce point est détaillé section 3.
2. Le formatage du corpus complet : normalisation, tokenisation et découpage en phrases.
3. L'annotation manuelle d'un sous-ensemble de 60 000 mots pour servir de référence. Elle est effectuée à l'aide d'un éditeur HTML, et les résultats sont transcripts dans un format XML. Les principes de cette annotation sont précisés section 4.
4. L'analyse par les participants du corpus complet et la transcription des sorties de leur analyseur dans un format XML commun.

5. L'évaluation qui consiste essentiellement en un calcul du rappel et de la précision sur les constituants et sur les relations. Cette étape est en cours de réalisation, comme nous le verrons section 5.

Les décisions concernant le formalisme d'annotation ont été l'objet de nombreuses discussions avec les participants et les fournisseurs de corpus (ils annotent de façon semi-automatique ou manuelle les corpus et doivent donc, dans la mesure du possible, être en accord avec les choix portant sur l'annotation).

Par ailleurs, le formalisme doit permettre une couverture la plus large possible des phénomènes syntaxiques de la langue. Nous avons fait le choix dans EASy d'annoter deux types d'information : les constituants, dont on annote la catégorie et l'étendue et les relations syntaxiques ou relations fonctionnelles, pour lesquelles on indique les éléments en relation, éléments qui peuvent être des mots ou des constituants. Dans ce formalisme, les constituants ne peuvent être discontinus ou imbriqués et ils sont de la taille la plus petite possible.

3 La constitution du corpus

Le corpus de test de la campagne est constitué d'un million de mots issus de sources diverses, provenant de cinq fournisseurs. Ceux-ci ont produit le corpus de test ainsi que les annotations d'une partie de ce corpus qui sert de référence pour les évaluations des systèmes. Ce corpus est réparti de la façon suivante:

- ATILF : corpus de textes littéraires - 150 000 mots fournis dont 15 000 mots annotés, avec des textes de Michelet, Coppée, Theuriet et Sandeau ;
- DELIC
  - 10 fragments de dialogues transcrits extraits du Corpus du Français Parlé ; 8 000 mots fournis et annotés ; thèmes divers ;
  - 2000 courriers électroniques personnels anonymisés ; 114 000 mots fournis ; correspondance privée ;
- ELDA
  - corpus de questions : 137 000 mots dont 5 000 annotés provenant de questions de la campagne TREC traduites et de questions extraites des notices bliographiques de la campagne AMARYLLIS ;
  - 250 courriers électroniques anonymisés : 7.000 mots fournis et annotés provenant de correspondances privées ;
  - extraits du journal Le Monde, de rapports du Sénat et de l'assemblée européenne (MLCC) : 235 000 mots dont 9 000 annotés ;
- LLF (Paris 7) : extraits du corpus Le Monde de 1992 - 15 000 mots annotés provenant d'articles généraux et économiques ;
- STIM-APHP : corpus de textes médicaux ; 100 000 mots fournis dont 5 000 annotés ; textes sur les maladie d'Alzeihmer, Parkinson, hépatite C, rapports et comptes rendus de conférences ;

La diversité des sources a pour objectif de tester les systèmes sur des structures syntaxiques très variées et sur des corpus ayant des spécificités propres. L'idée est d'évaluer sur les particularités de chaque type de corpus et non uniquement sur des phénomènes linguistiques spécifiques. Les corpus littéraires et journalistiques comportent des phrases très structurées et des repères tels que la ponctuation ou la typographie peuvent entrer dans le cadre des traitements. En revanche pour le corpus de courriers électroniques ou les transcriptions orales, la ponctuation est souvent, voire toujours, absente, la typographie (emploi de majuscules par exemple) aléatoire, l'orthographe souvent très relâchée et la structure des phrases mise à mal (Adda-Decker et al., 2003).

Afin de permettre à chaque participant de choisir le format de corpus le plus adapté à son système, le corpus a été fourni en plusieurs formats: texte brut normalisé, texte avec la séparation d'énoncé tokenisé et non tokénisé, texte étiquetté avec la version française de l'étiqueteur de Brill (Brunet 2000) et le jeu d'étiquettes utilisé dans Grace (Adda et al., 1999). Afin de permettre une évaluation cohérente, le découpage en énoncés fourni devait être pris pour référence. Ce découpage en énoncés a été réalisé automatiquement avec des outils développés pour la campagne (et précisé ci-dessous) à l'exception des transcriptions de l'oral. En effet ce corpus ne dispose pas de repères suffisants pour un découpage cohérent. Ce découpage a donc été réalisé à la main, une fois les anotations syntaxiques effectuées.

4 L'annotation de la référence

4.1 Le découpage en phrases et en mots

La première étape de l'annotation a consisté à découper le corpus de référence en phrases et en mots. Ces deux découpages sont souvent considérés comme des problèmes simples et quasiment résolus. Pourtant, nous avons pu constater que lorsque l'on s'intéresse à des textes réels, ce n'était pas toujours le cas ! Nous ne détaillerons pas cet aspect ici, mais nous donnerons quelques exemples. Pour le découpage en phrases, le problème est évidemment rendu plus difficile quand les textes analysés sont des transcriptions d'oral, spontané ou même lu. La présence de ponctuations fortes telles que le point est soit restituée lors de la transcription, soit complètement absente. Il faut alors se fonder sur des critères tels que la durée des pauses pour essayer de réintroduire la coupure en phrases. Les textes écrits posent également un certain nombre de cas difficiles à résoudre : par exemple lorsque le texte présente de nombreuses énumérations, sous forme de "listes à puces", ou du discours rapporté, comme illustré dans les deux exemples suivants (comme l'a également montré Núria Gala Pavia (Gala Pavia, 2003) dans sa thèse).

1. *Pour brancher l' appareil, vous devez :*
   - *vérifier votre installation électrique. Si vous ne respectez pas les normes, votre garantie ne fera plus effet.*
   - *relier le cordon d'alimentation à votre appareil...*
2. *Le directeur affirma :"Je ne peux pas accepter une telle situation.", devant le conseil d'administration de son établissement.*

Nous avons choisi de considérer la phrase la plus longue possible, pour éviter des découpages qui risqueraient de séparer des constituants, ou d'établir des relations en franchissant la frontière

de phrase, comme dans l'exemple 2 où la fin de la phrase se rapporte à la toute première partie, avant le discours rapporté.

Concernant le découpage en mots, nous avons choisi de constituer une liste de mots composés ou de locutions qui ne formeront qu'un seul mot lors de l'annotation en constituants. Exemple :

Dès que le soleil se lève, les coqs chantent.

Dès que ne forme qu'un seul mot. Ce découpage s'est avéré peu satisfaisant, les choix étant difficiles à faire de façon cohérente : ce découpage nécessitera un réalignement pour traiter les données retournées par les participants (voir sectio.

4.2 L'annotation en constituants

Le principe général que nous avons adopté consiste à annoter des constituants minimaux et non récursifs. Ce choix est dicté par le fait que nous souhaitons proposer un cadre qui permette d'évaluer des analyseurs ayant des caractéristiques diverses, en essayant d'être le plus équitable possible envers chacun d'eux. Prenons un exemple simple :

Le chat de la voisine

On peut analyser cette phrase comme un groupe nominal complexe, constitué d'un groupe nominal (*le chat*) et d'un groupe et d'un groupe prépositionnel (*de la voisine*), ce dernier étant une extension (de type complément de nom) du premier. On peut aussi n'annoter que les deux constituants simples, le rattachement du groupe prépositionnel au groupe nominal étant alors noté par l'intermédiaire d'une relation de type *modifieur du nom*. C'est cette dernière solution que nous avons adoptée. Elle nous permet de ne pas rejeter les *chunkers* qui ne relèvent que les constituants simples et de mieux noter les analyseurs plus précis où ces deux informations seront retrouvées. Ce parti pris nous permet de n'avoir ni constituant récursif, ni constituant discontinu, ce qui simplifie la tâche lors de l'annotation. Partant d'exemples réels, et de la littérature sur le domaine, nous avons déterminé une liste de six constituants. Nous en donnerons une première définition, que nous illustrerons sur des exemples simples[2].

- le groupe nominal (GN) : il est constitué d'un nom éventuellement précédé d'un déterminant (*<GN> la porte </GN>*) et/ou d'un adjectif antéposé accompagné de ses modifieurs (*<GN> la très grande porte </GN>*), d'un nom propre (*<GN> Rouletabille </GN>*) ou d'un pronom non clitique (*<GN> eux </GN>, <GN> qui </GN>*).

- le groupe prépositionnel (GP): il est constitué d'une préposition et du GN qu'elle introduit (*<GP> de la chambre </GP>*) ou d'un déterminant et d'une préposition contractés (du, aux, des) avec le GN introduit (*<GP> du pavillon </GP>*) ou d'une préposition suivie d'un adverbe (*<GP> de là </GP>*), ou de pronoms qui remplacent des GP (*<GP> dont </GP>, <GP> où </GP>*),...

- le noyau verbal (NV) : il regroupe un verbe, les pronoms clitiques plus éventuellement les particules euphoniques (*-t-* et *l'*) et la particule *ne* qui lui sont rattachés (*<NV> j'entendais </NV>, <NV> on ne l'entendait </NV> plus*). Un noyau verbal peut être à différents modes : temps conjugués mais aussi participe présent (*<NV> désobéissant </NV> à leurs parents*), participe passé (*<NV> fermée </NV> à clef*) et infinitif (*<NV> ne veut*

[2]Nous indiquerons alors la délimitation d'un constituant de type GX en adoptant une notation à la XML :*<GX> le constituant de type GX </GX>*.

</NV> pas <NV> venir </NV>). En cas de temps composés, nous identifions un NV distinct pour chaque verbe (*<NV> ils n'étaient </NV> pas <NV> fermés </NV>*).

- le groupe adjectival (GA) : il contient un adjectif lorsqu'il n'est pas épithète antéposé au nom (*les barreaux <GA> intacts </GA>*) ou un participe passé (*la solution <GA> retenue </GA> fut...*) ou présent (*les enfants <GA> désobéissants </GA>*) employé comme adjectif.

- le groupe adverbial (GR) : il contient un adverbe, à l'exception du *ne* qui fait partie du NV (*<GR> aussi </GR>, <GR> encore </GR> vous n'auriez <GR> pas </GR>*).

- le groupe verbal introduit par une préposition (PV) : il correspond à un noyau verbal dont le verbe n'est pas conjugué (infinitif, participe présent, ...) et introduit par une préposition (*<PV> d'ébranler </PV>*). Il peut contenir aussi des modifieurs de ce verbe, comme des adverbes (*<PV> de vraiment bouger </PV>*).

Nous pouvons résumer ces différents exemples sur cet extrait du *Mystère de la chambre jaune* :

<GN> la porte </GN> <GP> de la chambre </GP> <NV> fermée </NV> <GP> à clef </GP> <GP> à l' intérieur </GP> , <GN> les volets </GN> <GP> de l' unique fenêtre </GP> <NV> fermés </NV> , <GN> eux </GN> <GR> aussi </GR> , <GP> à l' intérieur </GP> , et , <GP> par-dessus les volets </GP> , <GN> les barreaux </GN> <GA> intacts </GA> ... et <GN> mademoiselle </GN> <GN> qui </GN> <NV> appelait </NV> <GP> au secours </GP> ! ... ou <GR> plutôt </GR> <GR> non </GR> , <NV> on ne l' entendait </NV> <GR> plus </GR> ... <NV> elle était </NV> <GR> peut-être </GR> <GA> morte </GA> ... mais <NV> j' entendais </NV> <GR> encore </GR> , <GP> au fond </GP> <GP> du pavillon </GP> , <GN> monsieur </GN> <GN> qui </GN> <NV> essayait </NV> <PV> d'ébranler </PV> <GN> la porte </GN>

Ce découpage semble assez simple à respecter, mais lors de l'annotation, nous avons pu constater que même pour les constituants qui paraissent simples comme le groupe nominal, la délimitation n'est pas toujours facile à affectuer.

4.3 Annotation en relation

L'annotation en relation va permettre d'établir tous les liens entre les constituants minimaux décrits ci-dessus. Après de nombreuses concertations avec les participants et les fournisseurs de corpus, nous avons retenu une liste de 14 relations, que nous détaillons ci-dessous, en commençant par les relations les plus simples.

- sujet-verbe (SUJ_V) : entre *elle* et *était* dans : *<NV> elle était </NV>*, ou entre *mademoiselle* et *appelait* dans *<GN> mademoiselle </GN> <NV> appelait </NV>* ;

- auxiliaire-verbe (AUX_V) : entre *a* et *construit* dans : *<NV> on a </NV><NV> construit </NV>* ;

- complément d'objet direct-verbe (COD_V) : la relation se note entre le verbe principal et le groupe nominal, comme par exemple entre *construit* et *la première automobile* dans : *<NV> on a</NV><NV> construit </NV> <GN> la première automobile </GN>*;

- complément-verbe (CPL_V) : pour lier au verbe les autres compléments exprimés sous forme de GP ou de PV, que ce soit les circonstants ou les compléments indirects, comme par exemple entre *en quelle année* et *construit* dans *<GP>En quelle année</GP><NV>a-t on</NV> <NV>construit </NV> <GN> la première automobile </GN>* ;

- modifieur-verbe (MOD_V) : cette relation concerne tous les constituants dont on peut affirmer qu'ils sont modifieurs et non compléments du verbe, comme les adverbes ou les propositions circonstancielles, comme entre *profondément* ou *quand la nuit tombe* et *dort* dans *<GN> Jean </GP><NV> dort </NV> <GR> profondément </GR> quand <GN> la nuit </GN><NV> tombe </NV>* ;

- complémenteur (COMP) : pour lier le complémenteur et le noyau verbal de la proposition subordonnée, comme entre *qu'* et *viendra* dans *<NV> Je pense </NV> qu'<NV> il viendra </NV>* ;

- attribut-sujet/objet (ATB_SO) : entre l'attribut et le noyau verbal, en indiquant si l'attribut se rapporte au sujet (comme entre *grand* et *est* dans *<NV> il est </NV><GA> grand </GA>*) ou à l'objet (comme entre *étrange* et *trouve* dans *<NV> il trouve </NV> <GN> cette explication </GN> <GA> étrange </GA>*) ;

- modifieur-nom (MOD_N) : on retrouve ici le lien entre les différents éléments qui composeraient un "gros" GN, comme le lien entre *unique* et *fenêtre* dans *<GN> l'unique fenêtre </GN>* ou entre *de la chambre* et *la porte* dans *<GN> la porte </GN> <GP> de la chambre </GP>* ;

- modifieur-adjectif (MOD_A) : pour indiquer ce qui se rapporte à un adjectif comme entre *très* et *belle* dans *<GN> la très belle collection </GN>* ou entre *de son fils* et *fière* dans *<NV> elle est </NV> <GA> fière </GA> <GP> de son fils </GP>* ;

- modifieur-adverbe (MOD_R) : même type de relation pour les adverbes comme entre *très* et *gentiment* dans *<NV> elle vient </NV> <GR> très </GR> <GR> gentiment </GR>* ;

- modifieur-préposition (MOD_P) : pour annoter ce qui se rapporte à une préposition comme entre *peu* et *avant* dans *<NV> elle vient </NV> <GR> peu </GR> <GP> avant lui </GP>* ;

- coordination (COORD) : pour lier les trois éléments que sont la coordination et ses coordonnés, comme entre *Pierre*, *Paul* et *et* dans *<GN> Pierre </GN> et <GN> Paul </GN> <NV> arrivent </NV>* ;

- apposition (APP) : la relation d'apposition lie l'élément apposé et celui auquel il s'appose en marquant l'identité entre les référents, comme entre *le député* et *Yves Tavernier* dans *<GN> Le député </GN> <GN> Yves Tavernier </GN>...* ;

- juxtaposition (JUXT) : elle est utilisée pour les constituants qui ne sont ni subordonnés ni coordonnés ni apposés, comme dans les cas d'énumération. Ainsi, elle lie *on ne l'entendait* et *elle était* dans *<NV> on ne l' entendait </NV> <GR> plus </GR> ... <NV> elle était </NV> <GR> peut-être </GR> <GA> morte </GA>* ;

Pour plus de détails sur l'ensemble de l'annotation, on peut se reporter au guide d'annotation qui a été rédigé pour servir à la fois aux annotateurs et aux participants, et qui est disponible à l'adresse suivante : http://www.limsi.fr/Recherche/CORVAL/easy

5 Évaluations prévues

Dans le protocole d'évaluation, nous allons effectuer plusieurs mesures, tenant compte à la fois de la diversité des corpus à annoter et de la variété des annotations. Pour les corpus, nous donnerons une évaluation par type de données (journalistique, littéraire, oral, écrit "relâché").

Ensuite nous donnerons les résultats des participants sur la reconnaissance des constituants et de leurs frontières, mais en distinguant les types de constituants. En effet, tous les constituants ne posent pas des problèmes de même niveau. Les mesures mises en œuvre sont de classiques mesures de rappel et de précision. Ce type d'évaluation a déjà été pratiqué dans une pré-version de l'évaluation EASy, permettant déjà de montrer des variations entre deux analyseurs utilisés pour ces tests (Vilnat et al., 2004).

Pour évaluer la reconnaissance des relations, nous distinguerons également des sous-ensembles de relations : d'une part les relations entre constituants minimaux telles que les modifieurs du nom, de l'adjectif et de la préposition ou auxiliaire-verbe. Ensuite nous nous intéresserons au relations correspondant à des relations syntaxiques "de base" telles que Sujet-Verbe ou COD-Verbe. Nous évaluerons dans un autre ensemble les relations réputées plus complexes, telles que l'apposition ou la juxtaposition. Notre but est de fournir des résultats détaillés, permettant d'avoir une évaluation précise des analyseurs, déclinée selon les différents phénomènes et les types de textes, plutôt qu'une évaluation trop globale. Nous espérons ainsi être en mesure de donner à des utilisateurs éventuels les moyens de choisir l'outil le mieux adapté à leurs besoins, et fournir un retour utile aux développeurs de systèmes.

L'état actuel de l'évaluation est de rendre possible les mesures prévues pour comparer les résultats des analyseurs et la référence. Pour cela, la concordance de la segmentation en énoncés et en formes de la référence et des sorties des analyseurs doit être vérifiée, car les analyseurs syntaxiques n'effectuent pas nécessairement des segmentations identiques à celles de la référence. Même si le corpus a été fourni sous forme brute et sous forme segmentée en énoncés et en formes (section 3), tous les participants n'ont pas choisi de travailler sur les données segmentées. Une étape de réalignement est donc en cours.

6 Travaux connexes en annotation de corpus et évaluation

Tandis que la construction du Penn Treebank, (Marcus et al., 1993), se poursuit pour la langue anglaise, les travaux en annotation de corpus se développent depuis quelques années pour une grande majorité des langues européennes. Les 3 ateliers intitulés "Workshop on Treebanks and Linguistic Theories" (TLT 2002, 2003, 2004) nous informent sur les langues étudiées (le suédois, le bulgare, ou le basque par exemple) et sur les méthodes employées, de même le recueil d'Abeillé (Abeillé, 2003), qui présente la construction de treebanks dans plusieurs langues et aussi leur utilisation à des fins d'évaluation. Généralement les corpus constitués sont de grande taille, leur annotation concerne des aspects morhologiques, syntaxiques et plus rarement des aspects plus sémantiques (comme le traitement de la co-référence).

Une autre approche, toute récente, de l'annotation de corpus mérite aussi d'être citée, cette fois l'annotation ne porte que sur peu de phrases (très peu même pour l'instant) mais elle vise des aspects plus sémantiques allant jusqu'à la pragmatique. Un premier atelier a eu lieu l'an dernier et le suivant se tiendra pendant la conférence ACL fin juin 2005 (Frontiers in Corpus Annotation

2004, 2005).

En ce qui concerne l'évaluation des analyseurs, la méthode la plus largement utilisée a sans doute été la méthode Parseval (Black et al., 1991). Cette méthode est fondée uniquement sur la comparaison des frontières des constituants, celle-ci étant effectuée entre les constituants du corpus arboré de référence et ceux de l'analyseur se soumettant à l'évaluation. Les mesures effectuées sont les habituelles mesures de précision et de rappel. La méthode Parseval a fait l'objet de plusieurs critiques qui sont détaillées dans (Carroll et al., 2003) et auparavant dans (Carroll, 2002) ou (Briscoe et al., 2002). Parmi ses détracteurs, Lin (Lin, 1998) propose de fonder l'évaluation sur la structure en dépendances et non sur les frontières des constituants. Plus récemment, Carroll et al. (Carroll et al., 2003) ont mis en œuvre une méthode d'évaluation elle aussi fondée sur les relations grammaticales : ils ont annoté 10 000 mots à l'aide d'une douzaine de relations et calculé le rappel et la précision entre cette référence et les résultats d'un analyseur. Dans un esprit proche, Lin (Lin, 2003) a évalué son analyseur en dépendances Minipar en utilisant un corpus de plus grande taille, mais contenant moins de relations. Ces différents auteurs concluent de la même façon : trouver un ensemble commun d'éléments sur lequel mener l'évaluation n'est pas toujours immédiat. Mais l'approche fondée sur les relations leur paraît plus pertinente, car un analyseur qui aura construit les bons constituants d'une phrase pourra malgré tout être très loin d'en avoir une bonne représentation sémantique.

7 Conclusion

Pour la campagne EASY (Vilnat et al., 2003), (Vilnat et al., 2004) l'annotation du corpus s'est voulue exhaustive, les constituants et les relations grammaticales ont tout deux été annotés et peu de phénomènes ont été écartés. Au moment de l'évaluation, treize participants étaient finalement prêts à soumettre leur analyseur. Ce qui constitue sans aucun doute une bonne représentation des équipes travaillant sur les analyseurs du français. Parmi les systèmes présentés, certains sont le fruit de plusieurs années de travail, tandis que d'autres sont trés récents (parfois même encore en développement). Qu'ils soient récents ou anciens, tous ne calculent pas constituants et relations. Pour ces raisons, il devenait naturel de calculer rappel et précision sur les constituants et sur les relations. En outre, tous les résultats seront différenciés : par type de corpus, par type de relations et de constituants ; nous permettant ainsi de dessiner une image plus précise et juste des performances de chacun.

Références

Treebanks : Building and Using Parsed Corpora, 2003, Dordrecht Kluwer, 406 p.

Adda G., Mariani J., Paroubek P., Rajman M., Lecomte J., 1999, L'action GRACE d'évaluation de l'assignation de parties du discours pour le français, revue *Langues*, Vol. 2, No. 2, pp 119-129.

Adda-Decker M., Habert B., Barras C., Adda G., Boula de Mareuil P. et Paroubek P., 2003, A disfluency study for cleaning spontaneous speech automatic transcripts and improving speech language models, *Proceedings of the International Conference on Disfluency in Spontaneous Speech (DISS)*, Septembre 2003, Göteborg, pp 67-70.

Aït-Mokthar S., Chanod J., Roux C., 2002, Robustness beyond shallowness : incremental deep parsing, *Journal of Natural Language Engineering*, Vol. 8, No. 3-2.

Black E., Abney S., Flickenger D., Gdaniec C., Grishman R., Harison P., Hindle D., Ingria R., Jelineck F., Klavans J., Liberman M., Marcus M., Roukos S., Santorini B. et Strzalkozski T., 1991, A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars, *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, Morgan Kaufman, Pacific Grove California, pp. 306-311, Février 1991.

Briscoe E, Carroll J., Graham J. et Copestake A., 2002, Relational evaluation schemes, *Proceedings of the Workshop Beyond PARSEVAL - Toward improved evaluation measures for parsing systems*, Third International Conference on Language Resources and Evaluation (LREC), Las Palmas, Gran Canaria, May 2nd, 2002.

Brunet E., 2000, Qui lemmatise dilemme attise, *Lexicometra*, No 2.

Carroll J., Briscoe T. et Sanfilipo A., 1998, Parser Evaluation: a Survey and a New Proposal, actes de First International Conference on Language Resources and Evaluation (LREC), Granada, May 1998, vol. 1 pp. 447-454 .

Beyond Parseval - Towards improved evaluation measures for parsing systems. *Atelier de la conférence LREC*, Las Palmas Espagne, John Carroll editor, 2002

Carroll J., Minnen G. et Briscoe E., 2003, Parser evaluation using a grammatical relation annotation scheme, in A. Abeillé (ed.), Treebanks: Building and Using Parsed Corpora, Dordrecht Kluwer, pp. 299-316, 2003.

Gala Pavia, Núria, 2003, Un Modèle d'analyseur syntaxique robuste fondé sur la modularité et la lexicalisation de ses grammaires, Thèse Université Paris-Sud, LIMSI-CNRS.

Gendner V., Illouz G., Jardino M., Paroubek P., Monceaux L., Robba I. et Vilnat A., 2003, Proposition de protocole d'évaluation des analyseurs syntaxiques, *Atelier sur l'évaluation des analyseurs syntaxiques de la conférence TALN*, pp 87-94

Lin D., 1998, Dependency-Based Evaluation of MINIPAR *Proceedings of the Workshop on Evaluation of Parsing Systems*, First International Conference on Language Resources and Evaluation (LREC), Granada, May 1998,

Lin D., 2003, Dependancy-based Evaluation of Minipar, in A. Abeillé (ed.), Treebanks: Building and Using Parsed Corpora, Dordrecht Kluwer, pp. 299-316, 2003.

Marcus M., Santorini B. et Marcinkiewicz M., 1993, Building a large annotated corpus of English : The Penn treebank, *Computational Linguistics, 19:313-330*

Monceaux L., 2002, *Adaptation du niveau d'analyse des interventions dans un dialogue. Application à un système de question-réponse.* Thèse de l'univeristé Paris 11, Décembre 2002

Oepen S., Netter N. et Klein J., 1996, Test Suites for Natural Language Processing, *Linguistic Databases, Nerbonne J. editor*, Center for the Study of Language and Information, CSLI Lecture Notes, 1996.

Srivinas B., Doran C., Hockey B.A., Joshi A.K., " An approach to Robust Partial Parsing and Evaluation Metrics ", in Proceedings of the Workshop on Robust Parsing, ESSLI, Prague, August 1996.

Vilnat A., Paroubek P., Monceaux L., Robba I. Gendner V., Illouz G. et Jardino M., 2003, EASY or How Difficult Can It Be to Define a Reference Treebank for French, *Proceedings of the 2nd Workshop on Treebanks and Linguistic Theories (TLT)* , Växjö, Sweden, pp. , November 14th-15th, 2003.

Vilnat A., Monceaux L., Paroubek P., Robba I. Gendner V., Illouz G. et Jardino M., 2004, Annoter en constituants pour évaluer des analyseurs syntaxiques. Actes de TALN 2004, Fés, Maroc, pp. 467-476.

Workshop "Frontiers in Corpus Annotation", http://nlp.cs.nyu.edu/meyers/frontiers/2005.html http://nlp.cs.nyu.edu/meyers/frontiers/2004.html

Workshop on Treebanks and Linguistic Theories, http://www.sfs.uni-tuebingen.de/tlt04/, http://w3.msi.vxu.se/ rics/TLT2003/, http://www.bultreebank.org/TLT2002.html

Problèmes d'annotation d'un corpus oral dans le cadre de la campagne EASY

Christophe Benzitoun, Jean Véronis

Equipe DELIC – Université de Provence
29, Av. Robert Schuman 13100 Aix-en-Provence
{Christophe.Benzitoun, Jean.Veronis}@up.univ-aix.fr

Mots-clés : corpus oral, annotation syntaxique
Keywords : spoken corpus, syntactic annotation

Résumé

Nous présentons, dans cet article, les problèmes que nous avons rencontrés et les solutions que nous avons adoptées pour l'élaboration du corpus oral de référence dans le cadre de la campagne EASY.

Abstract

In this paper, we present some problems and their solutions to annotate the gold standard spoken corpus of the EASY project.

1 Introduction

A la suite de plusieurs mois de réflexion et d'expérimentation liées à la constitution du corpus oral de référence pour le projet d'évaluation des analyseurs syntaxiques EASY (cf. aussi Benzitoun et al. 2004), il a fallu trouver un formalisme permettant de coder l'intégralité des données transcrites, « spécifiques » à l'oral (pauses, intonation, répétitions, amorces, inachèvements…), qui soit en adéquation avec celui proposé dans le cadre du projet. Cette contrainte répond à l'objectif que nous nous sommes fixés de reproduire le plus fidèlement possible les énoncés produits et ainsi de garder toutes les informations, quelle qu'en soit l'origine (prosodie, travail de formulation…), à travers les divers niveaux de l'annotation. En effet, celles-ci sont potentiellement utiles pour l'analyse automatique ultérieure des corpus oraux, et notamment leur utilisation pour l'amélioration des technologies vocales.

Pour cela, nous avons élaboré deux versions du corpus. Une première version, de travail, contient toutes les informations spécifiques à l'oral, sous forme de balises supplémentaires. La seconde, qui respecte scrupuleusement le guide d'annotation fourni par les organisateurs de la campagne EASY, est générée automatiquement à partir de la première (par suppression ou transformation d'informations). Le corpus de travail pourra être utile non seulement aux participants dont l'analyseur utilise des informations autres que textuelles ou qui veulent supprimer automatiquement les « disfluences », mais aussi pour évaluer les programmes détectant les disfluences, les pauses ou l'intonation ou, plus généralement, pour l'évolution

des technologies de la parole, dont les « modèles de langage » sont souvent mis au point à partir de textes écrits reflétant assez mal le langage parlé (par exemple le journal *Le Monde*).

Le corpus lui-même est composé de dix extraits d'environ cinq minutes chacun du *Corpus de Référence du Français Parlé* (DELIC, 2004) spécialement choisis pour leur caractère monologique et leur hétérogénéité situationnelle. La transcription orthographique a été effectuée entièrement à la main par des experts avec un cycle de réécoute/validation extrêmement strict. Les conventions de transcription adoptées (DELIC, 2004) ne contiennent aucun trucage orthographique (du type *p'tit, y'a,* etc.) ni aucune ponctuation, suivant la tradition de notre équipe, qui a clairement montré que la ponctuation de l'écrit était parfaitement inadéquate à la transcription de l'oral (cf. Blanche-Benveniste & Jeanjean, 1986). Par contre, sont notés avec soin les répétitions, les amorces, les *euh* d'hésitation, les allongements, les pauses (avec leur durée exacte) et les accents et mouvements intonatifs majeurs. Ceux-ci ont tous fait l'objet d'une balise particulière (voir la transcription du corpus dans Campione (2001, vol 2.)).

2 L'unité maximale (UM)

Les organisateurs de la campagne d'évaluation devaient fournir aux participants un texte segmenté en « phrases ». Or, il a été largement montré que cette notion est une notion purement graphique et qu'elle ne se retrouve pas, ni de loin ni de près, dans les productions orales (cf. Berrendonner, 2002 ; Blanche-Benveniste, 2002). On note d'ailleurs que même à l'écrit, phrases et unités linguistiques sont souvent non concordantes (Benzitoun, 2004).

Il a donc fallu avoir recours à une méthode de segmentation spécifique à l'oral. Dans une approche analogue à celle Blanche-Benveniste (2002), nous avons donc choisi comme unité de segmentation une *unité maximale* (UM), composée d'un *constructeur* (le plus souvent verbal, mais éventuellement aussi nominal, adjectival ou adverbial), et de tous ses *dépendants* et *associés* (au sens de Blanche-Benveniste et al. (1990)). Pour ce faire, il faut absolument se défaire des présupposés théoriques entourant le marquage des relations syntaxiques. Dans l'extrait suivant, le *parce que* ouvre une nouvelle UM qui n'entretient aucun rapport syntaxique avec la précédente.

1) *enfin dans une zo*ne touristique j'ai été remarque hein↗ ++ donc c'est pas bien difficile euh de bosser* ***parce qu'****en fait bon: effectivement mes économies arrivaient à: leurs fins↗ ++*

En outre, les cas d'UM parenthétiques venant couper une autre UM sont fréquents. Le guide ne disposant pas d'une étiquette pour les énoncés parenthétiques (bien qu'ils soient loin d'être inexistants à l'écrit), ces unités restent « flottantes » dans la version sous-spécifiée du corpus, et seules les relations internes sont marquées.

3 Les constituants

Les marqueurs *euh, hein, bon, ben, quoi, disons, je veux dire*, etc., extrêmement fréquents à l'oral, mais qui n'ont pas de catégorie grammaticale traditionnelle claire (ce ne sont ni des interjections, ni des adverbes) ont été annotés avec la balise « insert » (cf. Biber et al., 1999). Ont aussi été annotés les répétitions, les mots ou constituants inachevés et les segments inaudibles. Les éléments répétés sont inclus dans un groupe, quand ils en font partie, ou sont laissés à l'extérieur du groupe. Les constituants inachevés sont marqués avec la catégorie

qu'ils auraient s'ils étaient achevés et sont reliés à leur constructeur lorsqu'un autre élément n'occupe pas déjà la position syntaxique. Dans l'exemple suivant, *plusieurs* a été annoté GN et il a été relié au verbe *rester* par l'intermédiaire de la relation « modifieur de verbe ».

2) *je suis restée euh je sais pas qu- qu-* ***plusieurs*** *euh*

Seuls les inachèvements au niveau du mot ou du constituant sont notés et pas les inachèvements relationnels. Par exemple, dans 3), *une fille qui part seule en stop* semble en attente d'un verbe constructeur. Malgré cela, nous ne marquons pas d'inachèvement car la question est souvent complexe, l'inachèvement apparent pouvant parfois être complété par des phénomènes extra-linguistiques (geste, mimique, soupir…).

3) *enfin bon là euh ça a été dur dans la famille quand même hein parce que: ++* ***une fille qui part seule en stop*** *euh en Espagne c'était pas à côté*

Les relations entre inachèvement et répétition peuvent parfois être problématiques. Nous avons donc pris la décision de ne marquer les répétitions que dans les cas où celles-ci sont contiguës et où le mot est répété de manière exacte. Dans l'exemple suivant, le premier *de* sera marqué comme étant une répétition alors que le second sera marqué comme faisant partie d'un constituant inachevé car il est suivi par *des* (Adda et al. (2005) suivant les recommandations du LDC parlent de « révision »).

4) *j'ai fait l'expérience* ***de de: des*** *métiers de la restauration*

De même, *en hiver euh il y a plus* dans 5) ne sera pas marqué répétition car il n'est pas contigu.

5) *et puis* ***en hiver euh il y a plus*** *euh bon comme en France sans doute dans les coins touristiques hein / ++ en hiver il y a plus le boulot*

Les mots inachevés sont transcrits comme ils ont été prononcés. Certaines formes n'apparaissent donc pas dans les ressources dictionnairiques d'un analyseur. Dans l'exemple suivant, l'ensemble forme un GN et à l'intérieur on met une balise « fragment » pour signaler que *bou-* est un mot inachevé.

6) *des petites bou- + fioles*

4 Les relations

Il n'a pas été utile de créer des balises spécifiques pour les relations ce qui étaye, une fois de plus, l'hypothèse de notre équipe selon laquelle il n'y a pas de relations syntaxiques propres à l'oral (mais seulement des différences de fréquences). Comme à l'écrit, les relations peuvent être à une distance tout à fait remarquable, ce qui permet de faire des hypothèses concernant nos capacités d'encodage et de décodage (voir à ce sujet l'exemple présenté dans Benzitoun et al. (2004)). Dans ce cas, un élément est généralement répété pour permettre d'effectuer le raccordement.

7) *ce qui fait que j'ai amené des affaires d'hiver des affaires euh d'été* ***plus*** *euh à cette époque j'avais j'étais en maîtrise il me restait le mémoire à faire* ***plus*** *euh donc euh les livres tout ce qu'il me fallait pour faire mon mémoire là-bas*

Un autre phénomène remarquable est celui du double marquage (Blasco-Dulbecco, 1999).

Celui-ci n'a pas fait l'objet d'une mention particulière dans le guide d'annotation ce qui nous a obligé à proposer un traitement spécial en accord avec les organisateurs. Nous avons donc opté pour le marquage de deux relations identiques. Dans l'exemple 8), il y aura donc deux relations sujet malgré l'absence d'accord morphologique du verbe avec *nous*.

8) ***nous on*** *est là*

A propos de l'absence d'accord morphologique, on peut signaler ce cas très intéressant dans lequel *l'enfant, le bébé, les vieux* ont dû être reliés par la relation de « juxtaposition » sur le modèle *l'enfant, le bébé, les vieux peuvent halluciner*. L'accord morphologique ne se ferait que dans le cas où les éléments juxtaposés se trouvent avant le verbe.

9) *l'enfant peut halluciner le bébé↗ +++ ben écoutez* ***les vieux aussi*** ↘

Le dernier point abordé est celui des éléments non dépendants qui sont néanmoins enchâssés dans un énoncé. Vu qu'il n'y avait pas d'étiquette pour les non dépendants, nous avons hésité entre « modifieur » et « complément » pour ce type d'éléments.

10) ***chaque voyage*** *↗ il y a: il y a une re*mise en question au niveau euh++ physique*

5 Bilan & conclusion

Les phénomènes « spécifiques » de l'oral (en particulier les « disfluences ») sont extrêmement fréquents (de l'ordre de 10% des corpus). Leur repérage et leur traitement sont un enjeu très important pour le traitement automatique de la parole (dialogue homme-machine, reconnaissance vocale). Les phénomènes qui sont décrits ici, qui n'apparaissaient pas dans le guide d'annotation de la campagne EASY ou dont le traitement a demandé une interprétation particulière des consignes, nous amène aussi à reconsidérer notre regard sur l'écrit.

Références

ADDA G. et al. (2005), Disfluences et traitement automatique : l'heure de vérité, Journée d'étude de l'ATALA, 2 avril 2005.
BENZITOUN C. (2004), L'annotation syntaxique de corpus oraux constitue-t-elle un problème spécifique ?, Actes de *RECITAL*, pp. 13-22.
BENZITOUN C. et al. (2004), L'analyse syntaxique de l'oral : problèmes et méthode, Journée d'étude de l'ATALA, 15 mai 2004.
BERRENDONNER A. (2002), Les deux syntaxes, *Verbum*, Vol. XXIV, n° 1-2, pp.23-35.
BIBER D. et al. (1999), *Longman grammar of spoken and written English*, Essex, Longman.
BLANCHE-BENVENISTE CL. (2002), Phrase et construction verbale, *Verbum*, XXIV, n°1-2, pp. 7-22.
BLANCHE-BENVENISTE CL. ET AL. (1990), *Le français parlé. Etudes grammaticales*, Paris, CNRS Editions.
BLANCHE-BENVENISTE CL., JEANJEAN C. (1986), *Le français parlé. Edition et transcription*, Paris, Didier-Erudition.
BLASCO-DULBECCO M. (1999), *Les dislocations en français contemporain. Etude syntaxique*, Paris, Champion.
CAMPIONE E. (2001), *Etiquetage prosodique semi-automatique de corpus oraux : algorithmes et méthodologie*, Thèse de doctorat, Aix-en-Provence: Université de Provence.
DELIC (2004), Présentation du *Corpus de référence du français parlé*, *Recherches sur le français parlé*, 18, pp. 11-42.

Syntex, analyseur syntaxique de corpus

Didier Bourigault, Cécile Fabre, Cécile Frérot, Marie-Paule Jacques & Sylwia Ozdowska

ERSS – CNRS & Université Toulouse le Mirail
5, allées Antonio Machado, 31 058 Toulouse Cedex 9
{didier.bourigault,cfabre,frerot,mpjacques,ozdowska}@univ-tlse2.fr

Résumé

Cet article est un document de présentation de l'analyseur syntaxique de corpus Syntex, dans lequel nous décrivons les principes à la base du développement de l'analyseur et son architecture informatique. Une bibliographie du projet SYNTEX est donnée à la fin du document.

1 Analyseur de corpus

L'analyseur SYNTEX a été développé à l'origine (Bourigault, Fabre, 2000) pour remplacer le logiciel *LEXTER*[1], un analyseur syntaxique robuste dédié au repérage des syntagmes nominaux dans les corpus spécialisés et utilisé dans des applications de construction de terminologies ou d'ontologies spécialisées. Les diverses expérimentations réalisées avec *LEXTER* avaient mis en évidence la nécessité d'étendre la couverture du logiciel à l'extraction des syntagmes verbaux. A partir de ce constat, nous avons décidé d'entreprendre à l'ERSS la réalisation d'un nouvel analyseur, avec l'objectif d'en faire un outil opérationnel d'analyse syntaxique de corpus, utilisable dans différents contextes applicatifs, dont la construction de ressources lexicales spécialisées pour des systèmes de traitement de l'information (Bourigault *et al.*, 2004 ; Ozdowska *et al.*, 2005). SYNTEX doit traiter des corpus de phrases réelles, de taille importante (de quelques centaines de milliers à plusieurs millions de mots). Ceci impose des contraintes d'efficacité (temps de traitement), de robustesse (tolérance aux malformations syntaxiques et aux mots ou structures inconnues, possibilité de rendre des analyses partielles et incomplètes) et d'adaptabilité (prise en compte de certaines propriétés syntaxiques particulières des mots dans des corpus spécialisés). Les principes de base de l'analyseur sont les suivants : Syntex analyse des corpus préalablement étiquetés (section 2), il effectue une analyse syntaxique en dépendance (section 3), il est organisé sous la forme d'un enchaînement de modules de reconnaissance de relations

[1] BOURIGAULT D. (1994), Lexter, un logiciel d'extraction de terminologie. Application à l'acquisition des connaissances à partir de textes, Thèse de l'Ecole des Hautes Etudes en Sciences Sociales, Paris.

syntaxiques (section 4) et il exploite de façon combinée des procédures d'apprentissage endogène et des ressources lexico-syntaxiques de sous-catégorisation (section 5).

2 Etiquetage préalable

L'organisation du partage des tâches entre étiquetage morphosyntaxique (attribution d'une étiquette morphosyntaxique aux mots de la phrase) et analyse syntaxique (identification de constituants syntaxiques ou de relation de dépendance syntaxique) est un problème délicat. Disposer des étiquettes morphosyntaxiques des mots pour identifier les relations syntaxiques est extrêmement pratique. Mais, dans certains cas, la levée d'ambiguïtés catégorielles exige une analyse syntaxique partielle du contexte large. Le problème reste ouvert. Notre choix a été de séparer nettement les deux tâches et de confier la tâche préalable d'étiquetage à un outil extérieur. Même s'il y a interdépendance forte entre étiquetage et analyse, quantitativement l'analyse syntaxique a beaucoup plus à profiter de l'étiquetage que l' inverse. Des outils d'étiquetage de bonne qualité sont disponibles pour le français. SYNTEX prend en entrée les résultats du Treetagger[2], développé à l'Université de Stuttgart. Treetagger est un étiqueteur efficace et robuste. Il présente l'intérêt fondamental d'être ouvert, en ce sens qu'il est possible de faire en amont, à sa place, une partie du travail de tokénisation et d'étiquetage. Nous avons développé des procédures (lexiques, règles) de reconnaissance d'unités syntaxiques complexes qui viennent poser sur le corpus des étiquettes sur lesquelles le Treetagger s'appuie pour étiqueter les mots environnants. Nous avons aussi introduit dans la chaîne de traitement la possibilité d'intégrer un fichier de règles de tokénisation et de pré-étiquetage, données sous la forme d'expressions régulières, spécifiques au corpus à analyser. Cette fonctionnalité est essentielle quand il s'agit de traiter des corpus comportant des mots inconnus ou des structures « bizarres » (codes de produits, nomenclature d'éléments chimiques, etc.). Enfin, la frontière entre étiquetage et analyse n'est pas étanche. Dans certains contextes syntaxiques, l' analyseur effectue des retours en arrière sur l'étiquetage en venant modifier des étiquettes attribuées par le Treetagger (Jacques, 2005).

3 Analyse en dépendance

SYNTEX effectue une analyse en dépendance. Nous ne nous basons sur aucune théorique syntaxique particulière et nous n'avons pas élaboré une grammaire de dépendance spécifique pour ce projet. Notre base d'appui est la grammaire traditionnelle, tant au niveau des catégories morphosyntaxiques que des relations syntaxiques. Les principales relations syntaxiques actuellement reconnues sont les suivantes : sujet, objet direct, complément prépositionnel (de nom, de verbe et d'adjectif), antécédence relative, modification adjectivale (épithète, attribut), subordination. Les théories syntaxiques ou les descriptions linguistiques sont utiles pour définir des modes de représentation des relations (pour telle structure complexe, quel est le recteur, quel est le régi et dans quel sens s'établit la relation de dépendance, comment représenter les dépendances dans le cas des complexes verbaux et dans des structures discontinues, comme les structures comparatives, etc.). En revanche, pour faire de la syntaxe opérationnelle, c'est-à-dire pour écrire des règles de repérage de relations syntaxiques dans une chaîne étiquetée, le recours

[2] http://www.ims.uni-stuttgart.de

aux théories et descriptions syntaxiques est moins nécessaire. En particulier, le traitement des coordonnants et des virgules (apposition, incise, coordination), qui foisonnent dans les textes réels, exigent le développement de procédures d'analyse complexes, qui empruntent peu aux descriptions linguistiques classiques.

4 Architecture modulaire séquentielle

Nous décomposons le problème de l'analyse syntaxique d'une phrase en sous-problèmes élémentaires du type : soit *m* un mot de catégorie *C* dans la phrase étiquetée *S*, quel est le recteur syntaxique de *m* dans *S* ? De façon simplifiée, la résolution de ce problème s'effectue par un enchaînement en cascade d'une suite de modules qui prennent en charge chacun une relation syntaxique. Chaque module prend en entrée les sorties du module précédent. Cette organisation séquentielle des traitements impose de choisir un ordre dans l'analyse. On est face à un dilemme du type de celui du partage entre étiquetage et analyse. Par exemple, faut-il reconnaître les relations sujet avant de chercher à identifier les relations de coordination, ou faire l'inverse, ou répartir le traitement à deux moments de la chaîne ? Le choix de l'ordre est un choix difficile qui a un impact fort sur la programmation des différents modules, et sur lequel il est de plus en plus difficile de revenir au fur et à mesure que l'analyseur s'enrichit et se complexifie. A l'intérieur même de la chaîne d'analyse, les retours en arrière sont là aussi possibles, certains modules venant détruire et remplacer des relations syntaxiques posées par des modules antérieurs. Les modules sont constitués d'un ensemble d'heuristiques de parcours de la chaîne étiquetée et partiellement analysée, qui partent d'un régi (resp. recteur) potentiel pour aboutir à son recteur (resp. régi). Ils sont développés « à la main » par des linguistes informaticiens (dans le langage Perl), selon une méthode qui met en œuvre le recours à la connaissance grammaticale et à des tests nombreux et variés sur des corpus diversifiés.

5 Ressources lexicales

L'analyseur SYNTEX est peu (mais de plus en plus) lexicalisé. Nous avions fait le choix initial de la table rase. Contrairement aux approches qui choisissent, pour réaliser un analyseur syntaxique, de développer au préalable un lexique syntaxique très riche recensant les propriétés syntaxiques des mots de la langue, nous avons commencé sans aucune information de ce type. Cette approche est possible à partir du moment où l'on a choisi de s'appuyer sur les résultats d'un étiqueteur (on bénéficie indirectement des ressources lexicales éventuellement exploitées par celui-ci). Des informations lexicales sont intégrées dans l'analyseur au fur et à mesure des besoins : liste de locutions prépositionnelles, liste de verbes transitifs, liste de verbes se construisant avec des compléments en *que*, en *de*, etc. Pour résoudre les ambiguïtés de rattachement prépositionnel, l'analyseur exploite des informations de sous-catégorisation associées aux couples (mot, 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. Les expériences menées sur de nombreux corpus spécialisés 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-même, par analyse du corpus à traiter, des informations de sous-catégorisation spécifiques à ce corpus. Devant les limites inhérentes à l'exploitation exclusive d'informations de sous-

catégorisation endogènes, 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 la main à partir des tables du Lexique-Grammaire (Frérot, à paraître), puis de lexiques construits automatiquement à partir de corpus. Dans son état actuel, l'analyseur exploite un lexique de probabilités de sous-catégorisation construit à partir d'un corpus de 200 millions de mots (Bourigault, Frérot, 2005).

Bibliographie du projet Syntex

BOURIGAULT D., AUSSENAC-GILLES N., CHARLET J. (2004), Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas, *Revue d'Intelligence Artificielle (RIA)* , « Techniques Informatiques et structuration de terminologies », PIERREL J.-M. et SLODZIAN M. (Ed.) , Paris : Hermès. Vol. 18, n°1/2004, pp. 87-110

BOURIGAULT D., FABRE C. (2000), Approche linguistique pour l'analyse syntaxique de corpus, *Cahiers de Grammaire,* 25, Université Toulouse le Mirail, pp. 131-151.

BOURIGAULT D., FRÉROT C. (2005), Acquisition et évaluation sur corpus de propriétés de sous-catégorisation syntaxique *Actes de la conférence Traitement automatique des langues naturelles, TALN'05*, Dourdan, France, juin 2005

FRÉROT C. (2005), *Etude en corpus variés de 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 *TAL*, 44-3

JACQUES M.-P. (2005), Que : la valse des étiquettes. *Actes de la conférence Traitement automatique des langues naturelles, TALN'05*, Dourdan, France,

OZDOWSKA S., BOURIGAULT D. (2004) Détection de relations bilingues entre termes à partir d' une analyse syntaxique de corpus*Actes du 14ème Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle*, *RFIA'04*, Toulouse, France, janvier 2004

OZDOWSKA S., CLAVEAU V. (2005) Alignement de mots par apprentissage artificiel de règles de propagation syntaxique en corpus. *Actes de la conférence Traitement automatique des langues naturelles, TALN'05*, Dourdan, France

OZDOWSKA S., NÉVÉOL A., THIRION B.. Traduction automatique compositionnelle de bitermes dans des corpus alignés anglais/français. *Actes de la Conférence Terminologie et Intelligence Artificielle, TIA'05*, Rouen, France, avril 2005

L'analyseur syntaxique de LIMA dans la campagne d'évaluation EASY

Romaric Besançon et Gaël de Chalendar
CEA/LIST/LIC2M
BP 6 92265 Fontenay-aux-Roses Cedex France
{Romaric.Besancon,Gael.de-Chalendar}@cea.fr

Mots-clefs : analyse syntaxique, campagne d'évaluation, EASY, grammaires de dépendances, automates

Keywords: syntactic analysis, evaluation campaign, EASY, dependency grammars, automatons

Résumé Le LIC2M, laboratoire du CEA/LIST, a participé à la campagne d'évaluation EASY avec l'analyseur syntaxique de son système LIMA, un analyseur syntaxique robuste qui implémente une grammaire de dépendance. Les résultats obtenus sur le corpus d'exemples sont encourageants et permettent de valider les techniques utilisées. En revanche, le traitement de corpus plus généraux couvrant des phénomènes syntaxiques plus variés nécessiteront sûrement le développement de ressources supplémentaires ou la mise en place de traitements particuliers.

Abstract The LIC2M, a CEA/LIST laboratory, participated in the EASY campaign to test the syntactic parser of the NLP system LIMA, a robust syntactic parser that implements a dependency grammar. The development of this parser is an on-going work, but the results on the test set are promising. Nevertheless, the parsing of more general corpora, containing more varied syntactic phenomena should require additional work on the development of resources.

1 Introduction

Le LIC2M, laboratoire du CEA/LIST, développe depuis trois ans un ensemble d'outils de traitement automatique des langues en vue de leur utilisation dans diverses applications (recherche d'information, question-réponse, résumé automatique, etc.). L'ensemble de ces outils forme le système LIMA[1] (pour "LIc2m Multilingual Analyzer"), un système d'analyse linguistique avancée pensé dans une optique multilingue. Le LIC2M a participé à la campagne d'évaluation EASY (EASY, 2004) pour valider le module d'analyse syntaxique de LIMA, qui était en cours de développement lors du lancement de la campagne. Nous présentons dans la section 2 le module d'analyse syntaxique de l'analyseur multilingue LIMA et les adaptations nécessaires pour la participation à EASY. Puis nous présentons dans la section 3 une évaluation du système sur le corpus d'exemples annoté de la campagne EASY.

[1] LIMA est un travail collégial du LIC2M. En dehors des auteurs du présent article, ont participé à sa réalisation: O. Ferret, C. Fluhr, G. Grefenstette, M. Laib-Boukari, Y. Li, B. Mathieu, O. Mesnard, H. Naets et N. Semmar.

2 L'analyse syntaxique du système LIMA

Le système LIMA a été réalisé dans la lignée des travaux de Christian Fluhr et ses collègues au CEA (Fluhr et al., 1997), et reprend, en les enrichissant, les principes proposés à cette époque : dictionnaires *full-form*, catégories morphosyntaxiques positionnelles, dictionnaires bilingues etc. L'analyse linguistique est réalisée par une chaîne configurable de modules indépendants appliqués successivement sur un texte : segmentation, traitement des expressions figées, analyse morphologique, désambiguïsation syntaxique, reconnaissance des entités nommées, analyse syntaxique, création de termes composés. Actuellement, l'analyseur LIMA fonctionne sur six langues: allemand, anglais, arabe, chinois, espagnol et français. Son extension est en cours pour l'italien et le russe.

L'analyseur syntaxique de LIMA implémente une grammaire de dépendance (Kahane, 2000) en ce sens que les analyses produites sont exclusivement représentées par des relations de dépendance entre deux mots, un recteur et un régi. Nous nous inscrivons aussi dans le cadre des analyseurs robustes (Aït-Mokthar and Chanod, 1997; Grefenstette, 1998) puisque l'analyse est effectuée à l'aide d'automates à états finis dont la stratégie d'application permet d'obtenir une analyse pour toute phrase, même agrammaticale, du moment qu'un sous-ensemble de la phrase est reconnu par un des automates de la grammaire. L'analyse est séparée en deux étapes : la recherche des chaînes nominales et verbales et la recherche des relations de dépendance.

2.1 Chaînes nominales et verbales

Les chaînes nominales et verbales ne représentent pas un objet linguistique standard. En effet, il s'agit surtout d'une aide pour la recherche ultérieure des relations de dépendance et la désambiguisation de l'analyse syntaxique. Les chaînes nominales et verbales peuvent être décrites comme des syntagmes maximaux reliant entre eux l'ensemble des syntagmes minimaux non-récursifs (tels que définis dans le cadre de EASY (EASY, 2004)) susceptibles d'être liés, respectivement, à un même nom ou un même verbe.

Pour l'idenfitication de ces chaînes, une matrice définissant les successions autorisées de catégories et de mots est utilisée ainsi que des listes de catégories et de mots pouvant débuter ou terminer une chaîne, sachant qu'une même catégorie ne peut pas débuter à la fois une chaîne verbale et une chaîne nominale. Toutes les configurations possibles de chaînes sont cherchées pour chaque phrase.

2.2 Recherche des relations de dépendance

La recherche des relations de dépendance utilise des ensembles de règles représentant des grammaires locales et implémentées sous formes d'automates à états finis. Un phénonème syntaxique particulier peut être traité par une ou plusieurs règles (par exemple, la relation adverbe-adjectif est traitée par une seule règle alors que la relation sujet-verbe nécessite 12 règles).

Les règles sont définies par un élément *déclencheur* (qui peut être un mot ou une catégorie morphosyntaxique), et ses *contextes* gauche et droit, représentés par des expressions régulières. Des *contraintes* peuvent être spécifiées sur le déclencheur et/ou les éléments dans ses contextes gauche et droit, permettant de vérifier par exemples des accords en genre et en nombre, ou la

présence de relations existantes entre deux éléments. Des *actions* sont enfin attachées à chaque règle, par exemple pour créer des relations de dépendance supplémentaires. Pour chaque phrase, chaque mot est testé pour savoir s'il est déclencheur d'une règle : si c'est le cas, ses contextes droit et gauche sont testés. Si la règle s'applique, les actions sont effectuées.

Les règles sont regroupées en plusieurs ensembles qui sont appliqués successivement, pour permettre de traiter incrémentalement les relations cherchées en fonction de leur priorité et parce que certaines règles doivent s'appliquer avant d'autres alors que leurs déclencheurs sont situés plus avant dans la phrase. En particulier, les règles des relations homosyntagmatiques sont traitées avant les règles des relations hétérosyntagmatiques : les relations homosyntagmatiques sont les relations internes à une chaîne, donc les relations locales dans les syntagmes non interrompus par des incises, des parenthèses, etc. Elles comptent les relations entre les noms et les adjectifs, celles entre les verbes et les auxiliaires, mais aussi les relations entre les noms ou verbes et les mots dits grammaticaux comme les articles ou les adverbes. Les relations homosyntagmatiques sont traitées pour le français par trois ensembles de règles qui représentent un total de 56 règles. Les relations hétérosyntagmatiques sont les relations entre syntagmes et à longue distance, comme les relations entre le verbe et ses actants (sujet, compléments divers) ou les relations de conjonction ou de subordination. Ces relations sont recherchées après les relations homosyntagmatiques à l'aide d'un seul ensemble contenant 62 règles.

2.3 Adaptations pour EASY

Le corpus EASY étant déjà segmenté, les modules de LIMA concernant la segmentation (avec traitement des mots à tirets et des expressions figées) ont été remplacés par un simple découpage sur les espaces. La liste des formes composées fournie par les organisateurs, avec leurs catégories morphosyntaxiques a été intégrée dans le dictionnaire.

Concernant l'analyse syntaxique, l'analyse en dépendance produite par le système LIMA est fondamentalement équivalente à une analyse en constituants et dépendances. Pour trouver un constituant tel que le GN défini dans EASY, il suffit, à partir d'un noeud donné, par exemple un article, de suivre récursivement certains types de relations (ici, entre autres, les relations déterminant-substantif et adjectif prénominal-substantif) et de collecter les noeuds. L'ensemble de noeuds collectés forme le GN. En ordonnant les tests sur les noeuds et les ensembles de relations à suivre, on parvient à obtenir des syntagmes cohérents avec ceux définis dans le guide d'annotation EASY. Pour ce qui est des relations définies dans EASY, à un renommage près, il s'agit des relations extraites par notre système qui ne sont pas utilisées dans la construction des groupes. Leur extraction ne présente donc pas de difficulté particulière.

3 Evaluation

Les résultats officiels de la campagne EASY n'étant pas disponibles lors de l'écriture de cet article, l'évaluation proposée dans cette section a été faite sur le corpus d'exemples annoté fourni lors de la campagne. Les résultats obtenus sur ce corpus (en précision et rappel, pour chaque type de groupe et de relation) sont présentés dans la table 1. Ces résultats sont assez bons, mais il faut noter que ce corpus est essentiellement un corpus d'illustration des phénomènes syntaxiques de la langue française et n'est pas représentatif des phrases rencontrées dans un corpus réel. De plus, ce corpus ayant servi de corpus de référence lors de l'écriture des règles, ces

groupes	prec	rappel	*relations*	prec	rappel		prec	rappel
GA	81.2	97.5	ATB-SO	100.0	30.2	MOD-A	60.0	13.0
GN	85.4	95.0	AUX-V	94.4	87.2	MOD-N	61.4	50.0
GP	85.9	93.4	COD-V	60.2	60.2	MOD-R	87.5	87.5
GR	83.9	100.0	COMP	72.7	34.8	MOD-V	86.7	35.1
NV	96.5	100.0	COORD	71.4	20.8	SUJ-V	93.0	81.2
PV	81.8	90.0	CPL-V	52.9	42.2			
TOT	88.9	90.6				TOT	75.8	54.9

Table 1: Résultats obtenus sur le corpus d'entraînement EASY

résultats sont biaisés, et les résultats attendus sur les corpus de test plus généraux seront certainement moins bons. En particulier, il est difficile d'écrire des règles génériques (qui restent par essence assez locales) pour capter des relations lointaines dans des phrases complexes.

4 Conclusion et perspectives

L'analyseur syntaxique robuste du système LIMA donne des résultats encourageants sur le corpus d'exemples annoté fourni dans la campagne EASY. Néanmoins, les résultats attendus sur les corpus de test seront sans doute moins bons. Il est en effet difficile de développer manuellement (mais aussi d'apprendre automatiquement) des ensembles de règles complets et cohérents permettant d'analyser correctement du texte tout venant, pouvant contenir des structures arbitrairement complexes (relatives, incises, etc). Le type d'analyse que nous effectuons fonctionne bien sur des phrases simples. Par conséquent, nous développons actuellement des algorithmes et des ressources permettant de détecter les éléments complexes dans une phrase, de les supprimer temporairement, de les remettre en place après analyse de la phrase simple obtenue et enfin de les rattacher aux restes de l'analyse, ce qui accentuera l'aspect incrémental de l'analyse.

Par ailleurs, une des utilisations de l'analyse syntaxique dans LIMA est la construction de termes composés (utilisés pour la recherche d'information). Les utilisateurs du moteur de recherche se disent très satisfaits de ces mots composés mais une évaluation plus quantitative reste à faire. Une partie de cette évaluation pourra être faite dans le cadre de la campagne CESART, cousine de EASY dédiée à l'extraction de ressources terminologiques.

Références

Aït-Mokthar, S. and Chanod, J.-P. (1997). Incremental finite-state parsing. In *Proceedings of the 8th conference on Applied Natural Language Processing ANLP-97*, pages 72–79, Washington.

EASY (2004). Campagne d'évaluation des analyseurs syntaxiques. http://www.technolangue.net/article64.html http://www.elda.org/easy http://www.limsi.fr/corval/easy.

Fluhr, C., Schmit, D., Ortet, P., Elkateb, F., and Gurtner, K. (1997). Spirit-w3, a distributed crosslingual indexing and retrieval engine. In *INET'97*.

Grefenstette, G. (1998). Light parsing as finite-state filtering. In Kornai, A., editor, *Extended Finite State Models of Language*. Cambridge University Press.

Kahane, S., editor (2000). *Les grammaires de dépendance*, volume 41 of *Traitement automatique des langues*. Hermès.

Analyse syntaxique en dépendances et Evaluation

Christine Chardenon
France Télécom Division R&D, TECH/EASY/LN
Christine.Chardenon@francetelecom.com

Mots-clés : analyse syntaxique, dépendances, évaluation
Keywords: syntactical analyzer, dependency grammar, evaluation

Résumé Nous décrivons un analyseur syntaxique et commentons brièvement notre participation à la campagne d'évaluation EASY.

Abstract we describe a syntactical analyzer and we briefly comment our participation to the EASY evaluation campaign.

Introduction

L'équipe Langues Naturelles de France Télécom Division R&D a développé ces dernières années une chaîne de traitement linguistique constituée de modules réalisant des tâches de différents niveaux (lexical, syntaxique, sémantique, ...). Nous avons utilisé cette chaîne pour l'action d'évaluation des analyseurs syntaxiques TECHNOLANGUE/EASY. Nous ferons dans une première partie une brève description des modules nécessaires en amont du module d'analyse syntaxique, qui sera présenté dans une seconde partie. Nous complèterons la présentation de chaque module par une description des adaptations que nous avons faites sur les données utilisées par ce module pendant la phase d'annotation automatique des corpus.

1 Description de la chaîne

Nous distinguons trois étapes principales dans la chaîne de traitement produisant l'analyse syntaxique d'un texte : segmentation du texte, analyse minimale et analyse syntaxique. Nous allons nous intéresser dans cette partie aux deux premières.

Le module de *segmentation* découpe un texte en paragraphes, phrases et segments. Il exploite des données qui déterminent les types de segments et leur associe une description : un segment de type MOT correspond à un ensemble de lettres accentuées ou non, sans espace ni ponctuation, ni chiffres. Une adresse mail est reconnue comme un segment unique. Les segments obtenus sont regroupés en phrases, elles-mêmes regroupées en paragraphes. L'analyse syntaxique peut se faire au niveau phrase ou paragraphe. Durant la phase d'annotation de corpus, les données de segmentation ont été très légèrement adaptées, car

certains segments étaient inutilement découpés (general_elda , adresses de site internet). Nous avons conservé la segmentation en phrases fournie avec les corpus bruts, mais nous avons ensuite appliqué notre segmentation pour retrouver nos types de segments. Certains corpus (general_lemonde) présentait une segmentation en phrase peu pertinente (coupure de phrase après un "M."), le choix de conserver la segmentation en phrases aurait pu être remis en cause.

L'étape d'*analyse minimale* effectue des actions sur chaque segment obtenu lors de l'étape précédente, et ce en fonction de son type. Pour un segment de type MOT, l'action privilégiée est l'analyse lexicale, qui retrouve dans un lexique toutes les interprétations lexicales possibles du texte associé au segment. Le lexique utilisé pour l'action EASY est d'environ 200 000 formes fléchies. Chaque interprétation lexicale permet de créer un ou plusieurs objets appelés *terminaux*. Chaque terminal porte d'une part une catégorie syntaxique principale (*Catep*), d'autre part des informations morpho-syntaxiques, codées sous forme de traits. Par exemple, l'analyse lexicale du mot "livres" donne des terminaux correspondant à des interprétations nominales et verbales. En cas d'échec de l'analyse lexicale, le module contrôle l'application de stratégies de correction. Les types de correction (phonétique, ré-accentuation, analyse morphologique, etc) activées dépendent du corpus traité.

Pour les segments de type autre que MOT, comme les ponctuations, il est possible de créer des terminaux également associés à une description morpho-syntaxique. Par exemple, la virgule génère deux terminaux, l'un de catégorie PONC_GAUCHE et l'autre COORD.

Durant la phase d'annotation, les stratégies de corrections appliquées ont bien été différentes suivant les corpus : ré-accentuation, correction phonétique pour le corpus littéraire par exemple, le faible nombre de fautes d'orthographe dans ce corpus ne justifiant pas l'application de correction typographique. Pour les corpus de type mail, la correction phonétique était efficace, ainsi que la correction morpho-prédictive (prédiction de la catégorie d'un mot en fonction de sa terminaison).

L'étape d'analyse minimale se charge enfin de la reconnaissance des mots composés ou *locutions*. Une locution reconnue donne lieu à la production de terminaux, comme pour les mots simples. Une liste de locutions avait été fournie par les organisateurs de la campagne. Nous avons fait en sorte de compléter notre lexique de locutions quand cela nous a paru nécessaire. Cependant, nous n'avons pas gardé celles qui remettaient en cause les choix linguistiques de notre grammaire (nous traitons "l'un et l'autre" comme une coordination et pas une locution). Nous avons par ailleurs conservé nos locutions pendant l'analyse, il est donc certain que nous avons perdu des relations par rapport aux corpus annoté manuellement (exemple elda/general_elda : "sous traitants" est pour nous une locution).

2 Un analyseur syntaxique basé sur le formalisme des grammaires de dépendance

L'analyseur syntaxique commence par regrouper les terminaux dans des groupes syntaxiques de premier niveau (GS1). Un GS1 rassemble les terminaux issus d'un même segment qui ont la même catégorie principale, celle-ci devenant celle du GS1. Ces terminaux peuvent différer par leurs informations codées sous forme de traits (transitif/intransitif,…). Les terminaux issus de "livres" seraient ainsi répartis entre deux GS1, un de catégorie GN-NC

pour les interprétations nominales, l'autre de catégorie GV-PT pour les interprétations verbales.

L'analyse syntaxique d'une phrase est construite par créations successives de relations entre les GS1. Le processus est bottom-up et se fait par ilôts, générant des analyses partielles de la phrase analysée. Une analyse partielle est associée à un GS1 de tête. En début d'analyse, pour tout GS1, on crée une analyse partielle dont il est la tête. La construction d'une nouvelle analyse se fait par application d'une règle dite de dépendance entre deux analyses partielles déjà construites, ou plus exactement entre les deux GS1 tête de ces analyses. Si une règle s'applique, la nouvelle analyse contient toutes les relations des deux analyses partielles, plus la nouvelle relation créée entre les deux GS1 tête de ces analyses. La règle détermine quel est le GS1 tête de la nouvelle analyse, ainsi que le nom de la relation créée. Dans une analyse partielle donnée, tout GS1 ne peut avoir qu'un père, chaque analyse partielle est donc un arbre.

Les règles de dépendance sont décrites dans des fichiers de grammaire externalisés du module. Leur format est le suivant :

RègleAtt IdentifiantRegle NomRelation **Schéma** (CATEP)* Sens CATEP **CondsPrinc** (Traits*) **CondsDép** (Traits*) **AutresCondConcs** ((Trait*))	**RègleAtt** SUJ-PRN SUJ **Schéma** GV-PT >> PRN-S **CondPrinc** (SY_SUJ/!) **AutresCondConcs** ((P += SY_SUJ) (P/NOMBRE U D/NOMBRE) (P/PERS U D/PERS))

Le premier élément du schéma représente l'ensemble des choix possibles pour la catégorie du GS1 qui sera la tête ou *Principal* de la nouvelle analyse. Le troisième élément représente l'ensemble des choix possibles pour la catégorie du GS1 qui sera le fils ou *Dépendant* de la nouvelle relation. L'élément Sens détermine les positions relatives des deux GS1 (Principal devant ou derrière le Dépendant) dans la phrase. L'ensemble de traits correspondant à l'élément CondsPrinc constitue un filtre de sélection des GS1 candidats à être la tête d'une nouvelle analyse, de même, l'ensemble de traits correspondant à l'élément CondsDép constitue un filtre de sélection des GS1 pouvant être le dépendant d'une nouvelle analyse. Le dernier item est une série de conditions qui doivent être vérifiées par unification entre les ensembles de traits des deux GS1 Principal et Dépendant. Il contient aussi des conclusions à ajouter/retirer aux ensembles de traits résultant de ces unifications. L'application de ces conditions/conclusions se traduit donc par l'affectation d'un nouvel ensemble de traits au GS1 Principal (tête de la nouvelle analyse) et au GS1 Dépendant.

Voici l'analyse de la phrase : "tu livres la porte", dans le format de sortie simplifié utilisé pour générer ensuite la solution au format EASY. A un GS1 sont associés un identifiant (ID), un nom de relation (FONC) s'il n'est pas la tête de l'arbre, et l'identifiant de son Principal (PI).

<GS1 MOT="tu" LEM="tu" CAT="PRN-S" GRA="" ID="0" FONC="SUJ" PI="7" ></GS1>
<GS1 MOT="livres" LEM="livrer" CAT="GV-PT" GRA="" ID="7" ></GS1>
<GS1 MOT="la" LEM="le" CAT="GN-D" GRA="" ID="13" FONC="DET" PI="19" ></GS1>
<GS1 MOT="porte" LEM="Porte" CAT="GN-NC" GRA="" ID="19" FONC="OBJD" PI="7" ></GS1>

Le problème majeur à gérer dans ce type d'analyseurs est l'explosion du nombre d'analyses partielles. En effet, l'attachement de certains éléments de la phrase peut être ambigu en l'absence d'information sémantique levant cette ambiguïté. C'est le cas entre autre pour les attachements de groupes nominaux prépositionnels. Pour des applications à vocabulaire limité (quelques centaines de mots), il est possible d'introduire un contrôle sémantique des attachements. Pour un vocabulaire large, nous ne disposons pas de données sémantiques suffisamment riches pour permettre ce contrôle. L'explosion combinatoire peut être alors contrôlée par applications de diverses stratégies : application prioritaire de certaines règles par rapport à d'autres (par exemple, règles de construction de syntagmes minimaux), priorisation de la construction de relations sous-catégorisées (sujet, objet, etc), limitation de la distance entre un dépendant et sa tête, limitation du nombre d'analyses partielles concurrentes pour un tronçon de phrase, etc. Dans certains cas, l'analyseur produit un ou plusieurs arbres couvrant l'intégralité de la phrase. Parfois, il n'arrive pas à atteindre ce résultat, par manque de couverture de la grammaire pour les énoncés long (corpus du monde), ou parce que les énoncés sont agrammaticaux (corpus de mail). Dans ce cas, il est possible de sélectionner plusieurs arbres syntaxiques successifs qui couvrent la totalité de la phrase, de manière à identifier un maximum de relations syntaxiques (solution en morceaux).
Notre grammaire de dépendance ayant été développée en priorité pour des applications de type requête à un service, elle ne couvrait pas tous les phénomènes de la langue au lancement de la campagne EASY. Nous avons donc fourni un effort pour l'enrichir. Les phénomènes de coordination sont cependant incomplètement gérés, et ce d'autant plus que leur résolution satisfaisante nécessiterait dans certains cas l'exploitation de connaissances sémantiques. Pour des phrases complexes, nous avons donc assez souvent obtenu une solution en morceaux.

3 Conclusion

Pour l'action EASY, nous avons choisi de ne produire que les relations. Pour cela, nous avons fourni des relations entre formes de la phrase car il avait été décidé qu'une relation serait considérée comme valide si les deux formes sur lesquelles elle portait étaient contenues dans les constituants figurant pour cette même relation dans le corpus annoté manuellement. Nous avons par ailleurs choisi de ne pas transformer la sortie actuelle de notre analyseur pour générer des résultats conformes au format requis, mais plutôt d'ajouter un module de transformation de ces sorties. Nous avons développé un programme d'alignement de nos sorties avec les fichiers segmentés fournis par les organisateurs, en tenant compte des divergences portant sur les locutions. En assurant l'alignement, nous avons pu sortir nos relations en les faisant porter sur le numéro de forme correspondant au texte segmenté fourni. En même temps, nous avons renommé nos relations selon les recommandations EASY. Le travail de mise en correspondance de nos 200 noms de relations avec la quinzaine de noms de relations EASY a été assez rapide, mais nous n'avons pas essayé de résoudre finement certains cas de distinction entre compléments du verbe et modifieur du verbe, qui pouvaient être liés en plus à des différences de lexique.

Références

KAHANE S. (2000), Les grammaires de dépendance, *TAL 2000*, Vol. 41 no1, Paris, Hermès.

TagParser et Technolangue-Easy

Gil Francopoulo
Tagmatica
126 rue de Picpus 75012 Paris
gil.francopoulo@wanadoo.fr

Mots-clés : chunking, analyse syntaxique à large couverture
Keywords: chunking, large scale coverage syntactic parsing

Résumé TagParser est une chaîne d'analyse à stratégie montante dont les ressources ont été conçues afin de permettre un développement incrémental de la grammaire. TagParser combine TagChunker avec un module de calcul de relations syntaxiques.

Abstract TagParser is a 'bottom up' parser whose linguistic resources have been created in order to allow an incremental development. TapParser combines TagChunker with a module for syntactic relations computation.

1 Introduction

Nous présentons ici l'analyseur que nous avons utilisé pour participer au programme d'évaluation Technolangue-Easy organisé par le CNRS-LIMSI et ELDA sous l'égide du Ministère de la Recherche. TagParser est un analyseur de français[1] robuste, rapide et à large couverture. Il appartient à la famille des analyseurs dits « montants » dans le sens où les éléments constitutifs de l'analyse sont agrégés de proche en proche et par strates successives afin de former un résultat syntaxique.

2 Démarche

Notre démarche n'est pas purement scientifique mais plutôt technologique. Il ne s'agit pas d'établir ou de tester une théorie. Il s'agit d'élaborer le meilleur système en tenant compte des paramètres suivants :

- Qualité du résultat

[1] La version anglaise est en cours de développement selon les mêmes principes.

- Rapidité d'analyse
- Coût de développement
- Coût de la maintenance
- Couverture et robustesse

Notons que la couverture et la robustesse (bien que quelque fois confondues) sont, d'après nous, des critères légèrement différents même s'ils sont liés. La couverture, c'est le nombre de phénomènes linguistiques que le programme est capable d'analyser. Dans un système qui applique plusieurs stratégies d'analyse, on distingue la couverture globale de la couverture de chacune des stratégies d'analyse. La robustesse, c'est la faculté du système de pouvoir choisir une stratégie de remplacement quand un échec avec une certaine stratégie a été détectée.

De plus, TagParser n'est pas conçu avec des « bouts de ficelle » mais selon la conception objet et le développement itératif, avec un emploi généralisé d'UML et de Java. Tout en étant innovant, ce n'est pas un prototype de laboratoire.

3 Evaluation

Il est tentant de « sur-simplifier » l'évaluation d'un système au couple précision / rappel. En fait, ce couple ne fournit qu'une indication quantitative qu'il faut prendre pour ce qu'elle est : c'est-à-dire un couple d'indicateurs, rien de plus.

Loin de nous l'idée de dénigrer la campagne d'évaluation Technolangue-Easy. Elle a le mérite pour la première fois en France, en plus du calcul de la précision-rappel de 18 analyseurs :

- D'avoir permis la confrontation intellectuelle de toutes les équipes françaises du domaine lors des réunions de la mise en place de la campagne d'évaluation[2] ;
- D'avoir motivé de nombreux développements ;
- De fournir un guide d'annotation ;
- De fournir la première version d'un corpus annoté, bien qu'à ce propos, il semble assez dangereux de le figer une fois pour toute car cela aurait pour biais que les développeurs d'analyseurs se focalisent sur le corpus annoté plutôt que sur la tâche prise dans son ensemble. Dans cette optique, le corpus annoté devrait plutôt être considéré comme un ensemble de départ qui serait augmenté (ou modifié) périodiquement, plutôt que comme un corpus de référence immuable.

Mais d'autres critères sont tout aussi intéressants pour évaluer un système. L'analyseur n'est pas un but en soi : **il doit rendre des services à des utilisateurs et pouvoir évoluer dans un**

[2] Personnellement, nous avons beaucoup appris lors de ces réunions.

contexte industriel. Que vaut le couple précision-rappel si l'analyse d'une phrase prend plus de 5 minutes ? Que vaut ce couple si le coût d'évolution et de maintenance est prohibitif ?

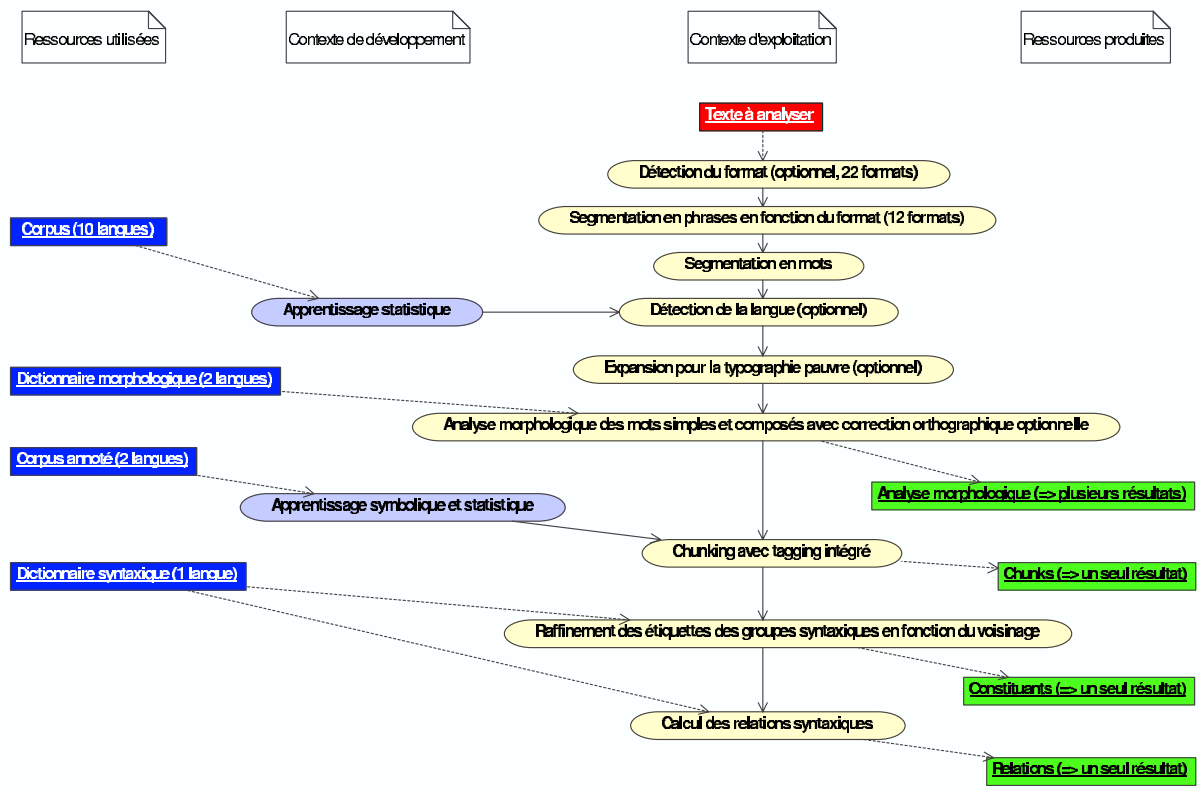
4 Chaîne d'analyse

4.1 Présentation

TagParser est une chaîne d'analyse complète associant segmenteur, analyseur morphologique, chunker et calculateur de relations syntaxiques. Un certain nombre d'outils auxiliaires peuvent lui être associés selon les circonstances comme la détection automatique des formats (est-ce un fichier Word ou HTML ?), la détermination de la langue, la correction orthographique ou le traitement de la typographie pauvre (absence d'accent et de casse).

La chaîne d'analyse prend un texte en entrée et produit un résultat syntaxique au format PEAS [Gendner]. Les deux parties les plus importantes sont le chunker [Francopoulo] et le calcul des relations syntaxiques.

4.2 Diagramme d'activité UML



4.3 Chunking

4.3.1 Objectif

4.3.1.1 étiquetage des parties du discours (i.e. Part of Speech Tagging)

Par exemple, dans les deux phrases suivantes : "Le comité d'experts table sur une croissance de 1%." et "La table est à 3 mètres du mur." Il s'agit, pour le mot "table", de le catégoriser en tant que verbe dans la première phrase et en tant que nom dans la seconde.

4.3.1.2 délimitation des groupes syntaxiques

Il s'agit de fixer les frontières des groupes syntaxiques. Ainsi, dans le premier exemple, le découpage suivant est construit : [Le comité][d'experts][table][sur une croissance][de 1%].

4.3.1.3 étiquetage des groupes

Les groupes constitués sont étiquetés et produisent le résultat suivant : GN GP GV GP GP.

4.3.2 Stratégie d'analyse

Le chunker appartient à la famille des analyseurs dits "hybrides" qui combinent des informations symboliques (via un dictionnaire et un automate) et un modèle probabiliste (via une matrice de pondération). La stratégie d'analyse repose sur trois principes simples : a) une analyse par automate est tentée, b) si elle échoue, des fragments d'automate sont combinés et tentés, c) quand des solutions multiples sont trouvées, la "meilleure" solution est choisie en fonction de la matrice de pondération.

Au lieu d'opposer les méthodes symboliques aux méthodes statistiques, nous les combinons pour en tirer le meilleur de chacune d'elle. Autrement dit : quand la situation est prévue lors du développement, on l'applique sans autre forme de procès, car elle correspond à une situation totalement maîtrisée par le développeur de l'analyseur. En revanche, si la situation n'est pas prévue, il faut faire appel à des variables cachées que seul un corpus est capable de procurer. On observe d'ailleurs que la frontière entre les deux méthodes varie en fonction de l'état de développement du système. Ainsi, pour un système peu développé, une méthode probabiliste est parfaitement adéquate. Mais les performances en terme de qualité plafonnent assez vite, même en raffinant le modèle mathématique. Si la part des connaissances symboliques s'accroît, plus la méthode symbolique est applicable car plus les situations sont prévues à l'avance, et de ce fait, moins les pondérations sont nécessaires.

Les informations linguistiques qui pilotent l'analyse sont issues d'un dictionnaire et d'un entraînement sur un corpus. **Aucune règle de grammaire n'est jamais écrite**. Si une phrase pose problème, il suffit de l'ajouter avec son annotation dans le corpus d'apprentissage, puis de relancer le programme d'apprentissage. Ce dernier vérifie d'une part que la phrase annotée ne contredit pas les annotations pré-existantes et d'autre part produit un nouvel automate et une nouvelle matrice de pondération. Son développement est de ce fait complètement incrémental. Il est de plus facilement adaptable à un certain type de texte en fonction d'une application particulière. Vous trouverez les détails dans l'article publié au congrès TALN-2003. Le chunker a été écrit en 2002 (et employé dans diverses applications) et les principes

de base exposés en 2003 sont toujours valables, mais depuis lors, le corpus d'apprentissage est passé de 18 000 à 34 000 mots et sa couverture améliorée d'autant.

4.4 Raffinement des étiquettes syntaxiques

Il reste quelques problèmes localisés à résoudre. Ainsi, par exemple, un certain type de groupe qui commence par "du" en français est ambigu entre GN ou un GP. Il n'est pas possible en se fondant uniquement sur les constituants de déterminer son étiquette. Prenons par exemple les phrases: "Il fait du ski" comparativement à "il arrive du ski". Dans ce cas précis, le chunking marque le groupe comme étant "indéterminé entre GN et GP". L'indétermination est levée en croisant deux types d'information : le voisinage du groupe en question et un dictionnaire syntaxique. Une petite grammaire désambiguïsation a été écrite explicitement dans ce but. Dans toutes les autres circonstances où il possible de déterminer si l'on a affaire à un GN ou un GP, l'étiquette est fixée dans la phase de chunking.

4.5 Calcul des relations syntaxiques

Le chunking produit un seul résultat constitué d'une liste de groupes syntaxiques (i.e. liste de constituants). Il s'agit maintenant de calculer les relations qui font que ces groupes occupent certaines fonctions dans la phrase. Au contraire du chunker, qui est dérivé d'un corpus annoté, les règles de grammaire sont écrites explicitement. En effet, l'apprentissage à partir d'un corpus de taille moyenne (i.e. de l'ordre de quelques milliers de mots) se prête mal aux relations syntaxiques. Il faudrait un corpus de l'ordre du million de mots pour en capter la généralité. Les relations sont catégorisées en 14 types comme Sujet-Verbe, Modifieur-Nom, Coordination décrites dans [Gendner]. Ces relations sont subdivisés en relations externes quand elles connectent des groupes entre eux et en relations internes quand elles connectent des mots à l'intérieur d'un même groupe. Le calcul des relations est en réalité l'application en séquence de 14 grammaires locales spécialisées pour tel ou tel type de relation. Ces grammaires exploitent deux critères différents: d'une part, une description (sommaire, il est vrai actuellement) du régime verbal représenté dans le dictionnaire syntaxique et d'autre part la position linéaire du groupe relativement aux autres groupes dans la phrase.

5 Conclusion

TagParser est un logiciel qui a été construit par étapes. Une fois le chunker élaboré et testé, avec l'objectif d'extraire des termes techniques, le besoin s'est fait sentir de distinguer les actants du verbe des modifieurs du verbe, dans cette optique le calcul des relations syntaxiques a été ajouté.

Références

FRANCOPOULO G. (2003) TagChunker : mécanisme de construction et évaluation, Actes TALN.

GENDNER V., VILNAT A. (2004) Les annotations syntaxiques de référence PEAS V-1.6. www.limsi.fr/Recherche/CORVAL/easy/

TALN 2005, Dourdan, 6-10 juin 2005

L'analyseur syntaxique multilingue FiPS dans la campagne EASy

Jean-Philippe Goldman, Christopher Laenzlinger, Gabriela Soare, Eric Wehrli

(1)Laboratoire d'Analyse et de Traitement du Langage
Département de Linguistique
Faculté des Lettres
Université de Genève
Rue de Candolle, 2
CH-1211 Genève 4 , Suisse
{goldman,laenzlinger,soare,wehrli}@lettres.unige.ch

Mots-clés : analyse syntaxique, grammaire générative, évaluation
Keywords: chart parser, generative grammar, assessment

Résumé

L'analyseur FiPS permet de transformer une phrase en une structure syntaxique accompagnée d'informations lexicales, grammaticales et thématiques. La présente communication décrit l'adaptation des structures en constituants de FiPS aux annotations syntagmatiques et relationnelles choisies dans le cadre de la campagne d'évaluation EASy.

Abstract

The FiPS parser analyzes a sentence into a syntactic structure reflecting lexical, grammatical and thematic information. The present paper offers a description of the adaptation of the structures in terms of constituents as existent in FiPS to the syntagmatic and relational annotations required by the assessement procedure EASy.

1 L'analyseur syntaxique FiPS

L'analyseur syntaxique FiPS (Laenzlinger et Wehrli 1991, Wehrli 1997), développé depuis plusieurs années au LATL, est un outil linguistique capable d'associer à chaque phrase d'un texte une structure syntaxique accompagnée d'informations lexicales, grammaticales et

sémantiques (« thématiques »).[1] Les applications de l'analyseur sont multiples : traduction automatique (ou aide à la traduction) (Wehrli 2003), synthèse et reconnaissance de la parole (Gaudinat et al. 1999 et Goldman et al. 2001), indexation et recherche 'intelligente' d'informations, extraction terminologique (Seretan et al. 2004) et apprentissage des langues (L'Haire & Vandeventer-Faltin 2003).

FiPS a été développé sur la base de la théorie *Principes & Paramètres* de la Grammaire Générative (Chomsky 1995, Haegeman 1994, Laenzlinger 2003). La structure en constituants assignée aux phrases repose sur un schéma X-barre réduit à deux niveaux : $[_{XP}$ L X R]. XP est une projection maximale de la tête X, alors que L (Spécifieurs) et R (Compléments) sont des listes (éventuellement vides) de projections maximales correspondant respectivement aux sous-constituants gauches et droits de la tête X. X est une variable correspondant aux catégories Adv (adverbe), A (adjectif), N (nom), D (déterminant), V (verbe), P (préposition), C (conjonction), T (temps). Une phrase complète a donc la structure suivante :[2]

$[_{TP}$ $[_{DP}$ *le* $[_{NP}$ *garçon*]] *a* $[_{VP}$ *recueilli* $[_{DP}$ *un* $[_{NP}$ $[_{AP}$ *petit*] *chat* $[_{AP}$ *noir*] $[_{AP}$ *affamé*]]]]]

La **stratégie d'analyse** est de type gauche à droite avec traitement parallèle des alternatives, combinant une approche incrémentale, essentiellement ascendante avec un filtre descendant. Selon cette stratégie dite du 'coin droit', l'algorithme est dirigé par les données (*data-driven*), c'est-à-dire on cherche à attacher un nouvel élément au coin droit d'un constituant dans le contexte gauche déjà existant. Ce dernier spécifie un ensemble de noeuds actifs auxquels le nouvel élément est susceptible de s'attacher. Les trois mécanismes fondamentaux utilisés par l'analyseur sont (i) la **projection,** (ii) la **combinaison** et (iii) le **déplacement**. Le mécanisme de **projection** (project) crée une structure syntaxique complète sur la base d'un élément lexical. Il permet aussi de créer des projections syntaxiques à partir d'autres structures syntaxiques (p.ex. un NP qui devient DP). L'opération de **combinaison** (merge) regroupe les constituants entre eux sur la base de règles de grammaire spécifiques à une langue particulière. Le **déplacement** (move) sert à établir une relation de chaîne entre un élément antéposé et la position où il est interprété thématiquement.

L'**implémentation objet** de cet analyseur (Wehrli 2004) tire parti des avantages de la programmation par objet (*object-oriented*), que sont l'extensibilité et la réutilisabilité des logiciels. Combinés aux propriétés d'*héritage* et de *liage dynamique de procédures*, ils facilitent une implémentation entièrement multilingue. L'idée de base dans notre modélisation 'objet' consiste à concevoir les objets linguistiques, telles que les structures lexicales et les projections syntaxiques, comme des structures abstraites dont l'implémentation peut varier d'une langue à l'autre. Ces variations sont traitées par l'extension de type en ce qui concerne les structures de données et par la redéfinition des méthodes pour ce qui est des processus de traitement de ces données. Le niveau le plus abstrait dans la hiérarchie des objets décrit les propriétés fondamentales qui sont vérifiées dans toutes les langues.[3] Les familles de langues et

[1] Par contraste, les analyseurs 'superficiels' ne cherchent pas à construire une représentation globale, ni a fortiori une forme logique, mais restent à un niveau de représentation morpho-syntaxique, avec un regroupement des constituants minimaux (groupes nominaux, groupes prépositionnels, etc.).

[2] Le groupe nominal est analysé comme un syntagme déterminant (DP) contenant un syntagme nominal (NP).

[3] Ceci s'apparente d'une certaine manière au concept chomskyen de 'grammaire universelle'.

les langues particulières étendent ce type en ajoutant des propriétés de plus en plus spécifiques, comme par exemple les pronoms clitiques au sein des langues romanes.

Cette approche syntaxique formelle et les avantages décrits ci-dessus de l'implémentation objet permettent à la fois un temps de traitement rapide (de l'ordre de 200 à 300 mots par seconde, autrement dit un million de mots par heure) et une souplesse et une facilité de développement, tant du point de vue de l'ajout d'une nouvelle langue que de la maintenance des ressources lexicales et grammaticales (règles morphologiques et syntaxiques).

2 La campagne d'évaluation EASy : la mise en correspondance des étiquettes syntagmatiques et des relations dans FiPS

La mise en relation des annotations syntagmatiques EASy avec les structures en constituants de FiPS a nécessité quelques adaptations. FiPS analyse jusqu'au niveau de la phrase, celle-ci étant formée d'un TP (tense phrase) et d'un CP (complementizer phrase). Ces constituants ont été ignorés de même que d'autres catégories fonctionnelles. Les structures syntagmatiques de FiPS étant construites en profondeur, il était parfois compliqué d'établir des correspondances avec le découpage linéaire des étiquettes EASy. Le noyau verbal (NV) de EASy correspond au verbe sous T (verbe conjugué, auxiliaire) ou V (participe, verbe infinitif) avec éventuellement les clitiques sujet et objet qui s'y attachent. Le groupe prépositionnel (GP) a pour correspondant direct le syntagme prépositionnel (PP) de FiPS. Le groupe adjectival (GA) est identifié comme un adjectif postnominal ou prédicatif (AdjP). L'étiquette GR de EASy correspond au syntagme adverbial noté AdvP dans FiPS. Quant à l'étiquette PV, elle correspond dans notre analyseur au complémenteur (C) prépositionnel suivi du TP qu'il introduit. Enfin, le groupe nominal (GN) a nécessité un découpage plus subtil pour nous, puisque FiPS crée un DP (déterminer phrase) qui peut contenir d'autres DP. Le GN est délimité par les têtes D et N avec les éventuels adjectifs prénominaux. Les correspondances EASy/FiPS sont résumées dans le tableau ci-dessous.

Noyau verbal **NV**	V ¦ V en T ¦ +Clitiques suj/obj¦ *ne* (+adv) V
Groupe nominal **GN**	$DP_{nom\ commun}$¦ $DP_{nom\ propre}$¦ $DP_{pron\ fort}$ ¦ D+Adj
Groupe prépositionnel **GP**	PP ¦ + *dont* ¦ +*où*
Groupe adjectival **GA**	$AdjP_{postposé}$ ¦ $AdjP_{prédicatif}$
Groupe adverbial **GR**	AdvP
Groupe verbal introduit par une préposition **PV**	P en C + TP et VP infinitif

Quant aux relations, certaines ont été facilement identifiées, comme les relations sujet et objet entre le verbe et ses arguments. La relation SUJ-V est établie entre un DP en Spec de TP et le verbe en T, entre un clitique sujet et le verbe en T, entre un sujet inversé (postverbal) et le verbe et enfin entre un sujet contrôleur ou monté et un verbe infinitif. La relation AUX-V est établie par la relation de sélection entre un auxiliaire et une forme (auxiliaire ou verbale) participiale. La relation COD-V est obtenue entre le verbe (en V ou T) et son complément direct en Compl de V, entre un clitique accusatif et le verbe, entre un élément-wh direct antéposé et le verbe gouverneur et enfin entre un verbe enchâssé sélectionné et le verbe sélectionneur. La relation CPL-V est établie entre un complément prépositionnel et le verbe, entre un clitique datif/génitif/oblique et le verbe, et entre un constituant prépositionnel antéposé et le verbe sélectionneur et enfin entre un ajout adverbial (GN=DP/GP=PP) et le

verbe. Quant à la relation MOD-V, elle concerne un adverbe (GR=AdvP) et le verbe modifié, et entre le verbe d'une phrase ajout, et le verbe principal. La relation COMP concerne le complémenteur (en C) d'une phrase conjuguée et le verbe de la phrase. La relation ATB-SO porte sur trois arguments dans une relation prédicative : (i) un constituant adjectival (AdjP), (ii) le verbe qui le sélectionne et (iii) l'argument (sujet ou objet) du prédicat. Cette relation est identifiée grâce aux propriétés de sélection spécifiées dans l'entrée lexicale du verbe. La relation MOD-N concerne un adjectif prénominal en Spec de N et le nom (N), entre un adjectif postnominal ou un groupe prépositionnel et le nom, entre le verbe d'une phrase relative et le nom que celle-ci modifie, et enfin entre deux noms dont l'un modifie l'autre (sans inversion possible). La relation MOD-A porte sur un adverbe et l'adjectif qu'il modifie, ainsi que sur un complément prépositionnel ou phrastique et l'adjectif sélectionneur. La relation MOD-R vaut pour un adverbe modifiant un autre adverbe ou les rares cas où un syntagme prépositionnel est complément de l'adverbe. La relation MOD-P est limitée aux adverbes modifiant une préposition (AdvP en Spec de P). La relation COORD est plus complexe, impliquant la conjonction et ses arguments. Les arguments en question peuvent être deux ou plusieurs syntagmes nominaux (DP), syntagmes prépositionnels (PP), syntagmes adjectivaux (AdjP), syntagmes adverbiaux (AdvP), syntagmes verbaux (VP). En cas de coordination de phrases, la relation est établie entre les verbes des phrases en questions. La relation APP vaut pour deux groupes nominaux (DP) dont l'inversion est possible. L'un se trouve attaché à l'autre. Enfin, la relation JUXT est établie grâce au repérage d'incise de phrase, annotée par les signes {…}* dans FiPS. Dans ce cas précis, la relation est effectuée entre le verbe de la phrase matrice et le verbe de la phrase en incise.[4]

Quelques remarques sur la campagne d'évaluation EASy doivent être apportées. Il a fallu adapter l'analyseur FiPS tant au niveau de l'entrée que de la sortie. A l'entrée, il s'agissait de pouvoir exploiter le corpus-test dont le format était connu au préalable. Parmi les 3 niveaux de représentations disponibles, nous avons choisi d'analyser celui qui se rapprochait le plus d'un texte brut, c'est-à-dire une segmentation en énoncé.[5] Ce choix nous permet d'évaluer le système complet en incluant l'analyseur lexical qui comprend un mécanisme complexe de segmentation lexicale et d'analyse morphologique. Par ailleurs, il aurait été difficile d'imposer à l'analyseur la segmentation lexicale présente dans les autres formats du corpus-test. Toutefois, l'analyse d'énoncés bruts impliquait de procéder, à la suite de l'analyse syntaxique, à un réalignement lexical pour que les données analysées par FiPS soient conformes avec le corpus de référence et donc évaluables. Pour ce qui concerne le format de sortie, l'analyseur a été adapté afin de générer des arborescences syntaxiques conformes aux recommandations précisées. En pratique, nous nous sommes heurtés aux aléas du traitement de gros corpus et avons dû réajuster certains mécanismes pour prendre en compte deux types majeurs de problèmes : 1) des cas extrêmes dans le corpus test ou des inconsistances[6] et 2) des problèmes

---

[4] Nous avons buté sur un bon nombre de problèmes particuliers, notamment dans le repérage des quantifieurs flottants, des constituants discontinus (p.ex. *combien….de*), des incises multiples, des ellipses (notamment dans la coordination), des relatives sans antécédent, des déterminants complexes.

[5] La segmentation lexicale et les informations grammaticales fournies par les deux autres niveaux de représentation ont donc été ignorées.

[6] En voici quelques exemples : (i) la segmentation lexicale de certains mots était incorrecte (ii) certains mots présents dans la liste des mots composés apparaissent redécomposés (iii) les corpus prétendument réalistes sont

de réalignement lexical post-analyse avec le corpus de référence, étape obligatoire pour une évaluation comparative de plusieurs systèmes. Ces problèmes nous ont fait prendre conscience de la nécessité d'augmenter la robustesse de l'analyseur et d'en améliorer la fiabilité.

Remerciements

Cette recherche est financièrement soutenue par le Fonds National Suisse de la Recherche Scientifique (projets FN n°101412-103999 et n°101511-101943).

Références

CHOMSKY, N. 1995. *The Minimalist Program*, Cambridge, Mass., MIT Press.

GAUDINAT, A., GOLDMAN J-P. & WEHRLI E. 1999. "Syntax-Based Speech Recognition: How a Syntactic Parser Can Help a Recognition System". *EuroSpeech Conference*, Budapest, Hungary, 1999, vol.4, p.1587-1590

GOLDMAN J.-P, GAUDINAT A,. NERIMA N., WEHRLI E. 2001. "FipsVox : a French TTS based on a syntactic parser". *4th Speech Synthesis Workshop*. Edinburgh, 2001

HAEGEMAN, L. 1994. *Introduction to Government and Binding Theory*, Oxford, Blackwell.

LAENZLINGER, C. 2003. *Initiation à la Grammaire Formelle du Français : Le Modèle Principes & Paramètres de la Grammaire Générative Transformationnelle.* Peter Lang, Berne/Berlin.

LAENZLINGER, C. ET E. WEHRLI, 1991. "FIPS : Un Analyseur interactif pour le français". *TA Informations*, 32:2, 35-49.

L'HAIRE, S. & VANDEVENTER-FALTIN, A. 2003. "Error diagnosis in the FreeText project". CALICO 20(3), T. Heift & M. Schulze (éds.). *Special Issue Error Analysis and Error Correction in Computer-Assisted Language Learning*

SERETAN, VIOLETA, LUKA NERIMA & ERIC WEHRLI. 2004. "Multi-word collocation extraction by syntactic composition of collocation bigrams". Dans Nicolas Nicolov et al (éds.) *Recent Advances in Natural Language Processing III: Selected Papers from RANLP 2003.*, 91-100. Amsterdam & Philadelphia: John Benjamins.

WEHRLI, E. 1997. *L'analyse syntaxique des langues naturelles: Problèmes et méthodes*, Paris, Masson

WEHRLI, E. 2003. "Translation of Words in Context". *IXth MT Summit*, New Olreans,

WEHRLI, E 2004. "Un modèle multilingue d'analyse syntaxique". Dans A. Auchlin et al. (éds.) *Structures et discours. Mélanges offerts à Eddy Roulet.* Pp 311-29. Nota bene, Québec.

---

en fait truffés d'espaces après les apostrophes et autour des ponctuations (iv) certains énoncés faisaient plus de 5000 caractères. L'énoncé est considéré ici comme une unité linguistique dont la longueur est de l'ordre de celle de la phrase (v) les corpus de messages électroniques connus pour leur agrammaticalité élevée nécessitent une grande robustesse de la part des analyseurs.

Comparaison de trois analyseurs symboliques pour une tâche d'annotation syntaxique

Jean-Marie Balfourier, Philippe Blache, Marie-Laure Guénot, Tristan Vanrullen
Laboratoire Parole et Langage – CNRS / Université de Provence
{prenom.nom}@lpl.univ-aix.fr

Mots-clefs : Analyse superficielle, Analyse profonde, Analyseur symbolique, Campagne d'Evaluation des Analyseurs SYntaxiques (EASY), Grammaires de Propriétés (GP).

Keywords: *Shallow parsing, Deep parsing, Symbolic parser, Property Grammars (PG).*

Résumé Nous présentons quelques réflexions concernant les différentes capacités des trois parseurs symboliques engagés dans la campagne d'évaluation EASY, face à une tâche d'annotation syntaxique.

Abstract *We present some reflections about the different abilities of three symbolic parsers, evaluated in the EASY campaign, in a specific syntactic annotation task.*

Introduction

Nous présentons dans cet article trois analyseurs évalués dans le cadre de la campagne EASY reposant sur le formalisme des *Grammaires de Propriétés* (ci-après GP). Les approches utilisées sont toutes les trois symboliques, mais utilisent des techniques différentes. Le premier est un analyseur superficiel, utilisant une grammaire simplifiée adaptée à la campagne d'évaluation dans laquelle un sous ensemble des propriétés de GP est exploité. Les second et troisième analyseurs utilisent des techniques d'analyse profonde. Ils emploient tous deux une représentation de l'information syntaxique sous la forme d'une grammaire complète et contrôlent le processus d'analyse et de déterminisation des résultats grâce à des algorithmes différents qu'il nous a intéressé de comparer.

Ces trois approches présentent bien entendu des résultats différents en termes d'efficacité et de couverture. Cette expérience permet de comparer, au sein d'une même approche symbolique et à partir de ressources identiques, les approches superficielles et profondes, ainsi que différents algorithmes d'analyse.

Nous présentons dans un premier temps les ressources utilisées comme base pour l'analyse, puis nous exposons les trois analyseurs syntaxiques. Enfin, nous abordons la question de l'évaluation des résultats en tentant de différencier les origines des erreurs dans les processus d'analyse.

1 Les ressources utilisées

Le lexique et l'étiquetage. — Les trois analyseurs prennent une entrée identique se présentant sous la forme d'un texte tokenisé et étiqueté. Chaque token est imposé dans le corpus à analyser, de sorte que tous les participants soient évalués sur les mêmes données. Chacun de ces tokens est accompagné d'un ensemble d'étiquettes possibles fournies de façon automatisée grâce à l'étiqueteur WinBrill. Nous avons réétiqueté ces tokens à l'aide de notre propre étiqueteur (celui du LPL), afin de faire correspondre le jeu de traits morphosyntaxiques de ces étiquettes avec celui qu'emploie notre grammaire, ainsi que dans le but d'affiner la qualité de l'étiquetage. De fait, l'étiqueteur du LPL présente une bonne efficacité et se base sur un ensemble de traits plus fins que ceux proposés par WinBrill. Le lexique de 430000 formes utilisé par notre étiqueteur est une variante de celui du LPL (présenté dans Vanrullen *et al.* (2005)), adaptée aux entrées du corpus d'évaluation EASY. Il reste cependant un certain nombre d'erreurs qui se répercuteront bien entendu sur le résultat de l'analyse. Ce point est discuté plus loin.

La grammaire. — D'un point de vue linguistique, la participation à la campagne Easy a consisté à concevoir une grammaire formelle dont les résultats visent à coïncider avec les indications fournies dans le Protocole d'Evaluation des Analyseurs Syntaxiques (ci-après PEAS, Gendner & Vilnat (2004)), qui servait de référence à la fois pour les annotateurs (référence pour les résultats) et pour les développeurs de grammaire (production des résultats). Le formalisme que nous utilisons pour développer des grammaires au LPL est celui des Grammaires de Propriétés (ci-après GP, cf. par exemple Blache (2005)).

Le guide PEAS a été conçu dans un souci de consensus entre les différentes théories linguistiques prises en considération lors de son élaboration. De fait, les choix effectués à la base du guide étant plus ou moins éloignés (suivant les cas) des thèses couramment soutenues dans le paradigme des GP, la conséquence de cela est que nous avons développé une grammaire spécifique à la tâche demandée pour Easy. Cette campagne nous a donc permis de tester, dans des circonstances concrètes, la flexibilité du modèle et de vérifier sa capacité d'expression pour une théorie significativement différente de celles habituellement adoptées.

Concrètement, le développement de la GP a consisté en les étapes suivantes :

1. suite à une première lecture du guide, un choix des propriétés utilisées, et la définition de leur sémantique (leur mode de fonctionnement, variable à volonté, cf. par exemple Vanrullen *et al.* (2003)),
2. interprétation approfondie des indications données dans PEAS et transcription de celles-ci en propriétés,
3. vérification de l'adéquation des résultats fournis par les trois parseurs sur un corpus-test,
4. estimation des causes des erreurs produites : méthodes de parsing, descriptions grammaticales, interférences entre propriétés, etc.,
5. ajustements de la grammaire pour les cas qui la concernent (modifications de la sémantique des propriétés, et/ou des propriétés elles-mêmes et/ou des ensembles de propriétés),
6. reprise des étapes 3, 4 et 5 jusqu'à ce que les résultats soient satisfaisants.

Ce va-et-vient entre ajustements de grammaire et tests sur corpus représentatif nous a permis d'adapter la grammaire, à la fois aux indications de PEAS en amont, et aussi aux différents traitements qui peuvent en être faits par chacun parseurs, en aval.

2 Les parseurs

Analyseur superficiel. — L'analyseur superficiel prend en entrée un texte étiqueté et désambiguisé. Il construit dans une première passe l'ensemble des groupes, puis les relations. La construction des groupes repose sur des informations syntaxiques partielles. Plus précisément, seules les propriétés de constituance et de linéarité ont été utilisées.

La stratégie repose sur une technique d'analyse coin-gauche. Il s'agit de repérer pour chaque token, grâce aux deux propriétés citées, sa faculté d'être coin gauche d'un groupe. Dans de nombreux cas, les informations citées sont suffisantes et l'initialisation du groupe correspondant est systématique. En revanche, certaines situations nécessitent la vérification du contexte immédiat. C'est par exemple le cas des groupes PV, qui débutent par une préposition. Cette dernière initialise habituellement un GP, mais dans un contexte verbal à droite, la préposition devient coin gauche du PV. Chaque initialisation de groupe entraîne la fermeture du groupe précédent.

Le mécanisme est donc en une passe unique et consiste à analyser successivement toutes les suites de trois tokens (le token candidat coin gauche, son contexte gauche et son contexte droit). La connaissance du type du groupe en cours permet de compléter la décision d'initialisation.

Les relations sont calculées dans une seconde passe sur la base des groupes construits. Chaque relation correspond à un traitement spécifique. Un premier traitement consiste à construire différentes tables regroupant les groupes et formes susceptibles d'être source ou cible d'une relation. Chaque relation consiste ensuite à parcourir ces tables et vérifier, en fonction des positions des candidats, leur appartenance à une relation. Dans certains cas (par exemple la relation complément-verbe) les candidats sont des ressources uniques. En d'autres termes, un candidat ne peut être complément d'un seul verbe. Cette information, correspondant à une consommation de ressource, est ajoutée dans les tables pour les items concernés. La détection des relations repose donc globalement sur des critères topologiques. Il s'agit d'une approximation qui ne permet pas d'assurer un bon contrôle ce qui entraîne une surgénération.

Les techniques utilisées par l'analyseur superficiel sont donc très simples, ce qui a bien entendu des conséquences sur le résultat obtenu (en particulier pour ce qui concerne les relations). L'avantage majeur de cette technique est sa robustesse et son efficacité : le corpus total est analysé en 4 minutes environ.

Premier analyseur profond. — Le premier analyseur profond utilise une version XML de la grammaire comme ressource permettant l'analyse. Deux algorithmes spécifiques sont mis en oeuvre : celui lié à l'analyse et celui destiné à déterminiser la sortie. La technique d'analyse repose sur une transformation de la grammaire en un graphe de contraintes. Ce graphe, ainsi que la sémantique des contraintes sont confrontés aux corpus à analyser. L'analyse consiste à produire l'ensemble des contraintes satisfaites et enfreintes par chaque succession de tokens constituant un énoncé phrastique. En cours d'analyse, les catégories de la grammaire EASY sont construites en fonction du nombre de contraintes qui les satisfont. Cette technique utilise la mesure de *densité de satisfaction* présentée dans Vanrullen (2004). Si une construction est suffisamment satisfaite (ceci en fonction d'un seuil préétabli pour chaque corpus), elle sera conservée dans le résultat. Une fois l'analyse achevée, une déterminisation des résultats est réalisée en utilisant à nouveau la mesure de densité de satisfaction : les constructions définitivement choisies pour le résultat final sont celles qui maximisent la somme des densités de satisfaction pour

chaque énoncé. Cette maximisation est réalisée grâce à un calcul sur des cliques (au sens de la théorie des graphes), où chaque clique présente un conflit entre plusieurs hypothèses d'analyse. Cet analyseur présente l'avantage de séparer les algorithmes et les données (le programme est indépendant de la grammaire et de sa sémantique, contrairement au troisième analyseur qui doit les inclure au sein même du programme). Son inconvénient réside par contre dans sa lenteur. La complexité moyenne reste polynomiale, mais la durée du traitement est de plusieurs jours sur plusieurs machines pour le million de mots que constitue le corpus. Pour cette raison, bien qu'il était possible de calculer aussi bien les constituants que les relations à l'aide de ce parseur, nous n'avons pu effectuer que la première tâche dans les délais impartis. Ceci confronte les contingences de la campagne d'évaluation aux possibilités réelles des analyseurs, lorsque ceux-ci sont programmés dans le cadre de la recherche expérimentale en laboratoire.

Second analyseur profond. — Ce second analyseur profond, dans son principe, permet la production de tous les arcs possibles conformes à une grammaire de propriétés donnée. Cette analyse se fait en différentes passes, chaque passe engendrant un niveau hiérarchique supplémentaire dans les syntagmes produits. Les constituants EASY n'ayant qu'un niveau de hiérarchie, une seule passe est nécessaire pour les produire. Mais l'analyse a été complétée par des passes supplémentaires afin de produire les relations.

Pour cela, la grammaire EASY a été étendue par l'ajout de catégories syntagmatiques décrivant l'intégralité d'une phrase. Le choix a été fait, au sein de cette grammaire étendue, de présenter chaque relation comme une contrainte de dépendance particulière entre 2 (ou 3 pour certaines relations) constituants d'un syntagme supérieur. Ainsi, la relation <sujet-verbe> est une contrainte de dépendance entre le groupe nominal et le groupe verbal au sein de la catégorie phrase.

Une fois produit par l'analyseur l'ensemble des constituants possibles d'une phrase, on recherche sa meilleure couverture possible (le syntagme le plus englobant, si possible du genre phrase) ainsi que l'ensemble de ses constituants jusqu'aux catégories lexicales. Les constituants EASY correspondent alors au premier niveau de cet assemblage et les relations, à toutes les dépendances nécessaires à sa construction.

3 Interprétation des résultats

Les résultats de la campagne ne sont pas encore disponibles ; il serait donc prématuré de parler d'une évaluation complète de nos résultats. Cependant à la lumière des travaux de développement effectués au cours de la participation à la campagne et des comparaisons entre les différentes sorties obtenues en fonction des techniques employées, on a pu mettre en évidence un certain nombre de caractéristiques (avantages et limites) propres à chacune des approches adoptées, pour le traitement d'un même corpus à l'aide d'une même grammaire.

Il est bien évident que, nos parseurs se basant sur des entrées étiquetées et cette étape préalable n'étant pas fiable à 100%[1], les erreurs provenant de la phase d'étiquetage sont autant de causes d'erreurs systématiques de parsing (bien que dans ce cas ce ne soit pas le parsing lui-même qui soit à mettre en cause). Cela étant dit, même à partir d'entrées correctement étiquetées,

[1] Pas plus que la transcription elle-même : on peut trouver autant de coquilles dans les textes de sources écrites que dans les transcriptions de corpus oraux, coquilles qui augmentent d'autant la probabilité d'erreur d'étiquetage.

nous avons pu lors de nos étapes de test successives mettre en évidence un certain nombre de différences caractéristiques de traitements entre les parseurs. Compte-tenu du fait que les indications d'annotation ne sont pas censées laisser place à l'ambiguïté, cela signifie que dans ces cas seule l'une des réponses est à considérer comme étant celle attendue. Nous allons donc maintenant présenter quelques-unes de ces différences de traitements, qui peuvent provenir soit des techniques de parsing (méthodes d'introduction des "groupes" ou des "relations"), soit de la grammaire elle-même.

Constituants. — Comme on l'a vu précédemment, le fichier que les parseurs prennent en input est un texte étiqueté. A chaque token correspond une liste d'étiquettes possibles, et parmi elles une (sous-)liste des propositions retenues par le désambiguïseur. Pour des raisons de simplicité (et de probabilité), les deux premiers parseurs ont retenu comme technique de ne considérer que le premier élément de cette liste pour leurs analyses, alors que le troisième prend en compte toutes les possibilités proposées. Il s'est avéré que dans certains cas cette dernière technique ait permis de "rattrapper" une imprécision récurrente du désambiguïseur, et ait ainsi permis de produire une analyse en constituants juste là où les deux autres étaient nécessairement erronées, par exemple dans les cas fréquents d'ambiguïté entre un déterminant et un amalgame préposition + déterminant qui ont la même forme (*des, du*,...), et pour lesquels le désambiguïseur ne choisissait pas toujours la meilleure possibilité. Cela avait pour conséquence que les deux parseurs se contentant de la première possibilité retenue par le désambiguïseur introduisaient systématiquement, en cas d'erreur, des GN à la place de GP et *vice versa*, alors que le troisième pouvait vérifier la cohérence de son choix et opter pour l'étiquette qu'il évaluait comme fournissant l'analyse la plus satisfaisante.

Relations. — Seuls deux des trois parseurs ont produit des relations, le premier et le troisième. Les deux techniques d'introduction des relations relèvent de deux approches totalement différentes :

- Comme il le fait pour les groupes, le premier parseur établit des relations en fonction de leur *constituance* (construction des tables regroupant les candidats possibles) et de leur *linéarité* (introduction des relations pour tous les cas où l'ordre des éléments est celui recherché).
- Le troisième parseur a intégré les relations comme étant des contraintes de *dépendance* caractéristiques, mettant en relation les deux ou trois éléments concernés au sein de syntagmes de niveaux supérieurs aux groupes Easy.

Il en résulte que là où le premier parseur génère non seulement toutes les relations attendues mais aussi un nombre conséquent de relations superflues, le troisième introduit souvent moins de relations, cependant chacune de celles-ci dépend directement de l'exactitude des groupes qui les contiennent et a par conséquent une plus forte probabilité d'être juste. Prenons l'exemple du traitement de l'énoncé suivant :

(1) Tout en adoptant le principe de l'adhésion de ces pays, le Conseil européen a précisé que ceux-ci devraient répondre à certains critères et que la capacité de l'Union à accueillir de nouveaux membres devrait également être prise en compte.

Pour cet énoncé, notre premier parseur a introduit 20 relations, dont 7 justes. Notre troisième parseur a introduit 22 relations, dont 11 justes. Le détail est donné en figure 1[2].

[2]Bien évidemment ces pourcentages ne sont calculés que sur l'énoncé donné en exemple et ne sauraient en aucun cas être représentatifs des résultats généraux des parseurs ; il s'agit juste ici d'illustrer les différences de traitements et non d'évaluer les résultats.

	Parseur 1		Parseur 2	
Relation	Nombre	dont justes (précision)	Nombre	dont justes (précision)
Mod-N	9	4 (44 %)	4	3 (75 %)
Suj-V	3	1 (33 %)	1	1 (100 %)
Cod-V	2	0 (0 %)	2	2 (100 %)
Cpl-V	1	0 (0 %)	4	2 (50 %)
Mod-V	1	0 (0 %)	1	1 (100 %)
Aux-V	1	1 (100 %)	2	2 (100 %)
Comp	2	0 (0 %)	8	0 (0 %)
Coord	1	1 (100 %)	0	0
Total	20	7 (35 %)	22	11 (50 %)

FIG. 1 – Détail des relations pour les parseurs 1 et 3 sur l'énoncé de l'exemple (1).

On voit que même si le nombre total de relations introduites n'est pas très différent (20 dans un cas, 22 dans l'autre), par contre la pertinence de ces relations diffère d'un parseur à l'autre, puisque dans la moitié des cas le parseur 3 ne fournit que des bonnes relations (pour Suj-V, Cod-V, Mod-V et Aux-V), et au moins 50 % de justes dans la moitié des cas restants (Pour Cpl-V et Mod-N). Par contre toutes ses propositions de Comp sont erronées, et il n'a pas trouvé la relation de Coord que le premier parseur a su construire.

Grammaire. — Le Protocole d'Evaluation propose une description des Groupes Nominaux indiquant, globalement, que fait partie du GN tout ce qui est compris entre le déterminant et le nom (ou l'objet qui occupe sa place). Cela ne comprend pas, donc, tous les possibles constituants du syntagme nominal (au sens classique) qui figurent après le nom. Le traitement de ce point a été facilement représentable dans notre GP. Cependant, on peut lire plus loin dans le guide PEAS que cette description a une limite : elle n'est plus valable pour les "*éléments en langue étrangère, (l)es formules, (l)es équations mathématiques ou chimiques*", ni pour les "*références bibliographiques au sein de textes (comme dans les articles)*" (Gendner & Vilnat (2004), section D.1.X). Pour ces cas précis, il est dit qu' "*ils peuvent être regroupés dans des constituants*" (*ibid.*). Les trois exemples présentés (et leurs analyses respectives) dans PEAS sont les suivants :

(2) a. La patiente présentait **<GN> un placenta prævia </GN>.**

b. L' amour est plein de quiétude et gardé de sentinelles <GP> à toutes les portes des sens </GP>, et **<GP> in cunclis sensibus custoditus </GP>**.

c. Le cas souvent étudié (**<GN> Hamburger 99 </GN>**) est revu dans cet article.

L'exemple (2a) montre un cas où un mot étranger (*prævia*) a reçu un traitement différent de la norme, du fait de sa nature de "mot étranger" : s'il avait été considéré comme un mot "normal" il n'aurait pas été intégré au GN *un placenta*, mais aurait été l'objet d'un GA unaire (puisque postposé au nom auquel il se rapporte). L'exemple (2b) montre le parallèle fait entre le GP *à toutes les portes des sens* et le groupe suivant, qui est une expression latine, *in cunclis sensibus custoditus*, qui se voit affecter l'étiquette de GP parce que coordonné au GP qui le précède[3]. Enfin, l'exemple (2c) montre le cas exceptionnel de la citation (référence) où un objet qui non

[3]Ou alors du fait de l'analyse syntaxique du groupe latin, mais le parsing du latin n'étant pas l'objet de la grammaire ni de la campagne, et le latin n'étant pas la seule langue possible dans ce cas, nous avons par principe abandonné cette possibilité.

seulement n'est pas un Nom propre (une date en l'occurrence) mais qui est également post-posé, peut faire partie d'un GN lui-même constitué d'un Nom propre (ce qui théoriquement est impossible, en vertu de la description du GN donnée en section B.2).

Les cas tels que celui de l'exemple (2c) ont pu être décrits sans problème dans la grammaire. En revanche, les exceptions illustrées par les exemples (2a) et (2b) ont été impossibles à exprimer. En effet, les groupes que la grammaire permet d'introduire ne contiennent pas d'information qui permette de savoir si un groupe donné est constitué d'éléments de langue étrangère, de formules ou d'équations mathématiques ou chimiques. Cette information, si elle pouvait figurer, proviendrait de l'étiquette des constituants et non de l'analyse, puisqu'en termes strictement syntaxiques, ces "natures" de groupes ne sont pas pertinentes en soi : leur analyse demeure la même que pour tous les autres groupes. Or ni l'étiquetage fourni pour les besoins de la campagne, ni notre couple étiqueteur-désambiguïseur, ne permet cela : soit le lexique contient le mot à étiqueter (cas des expressions mathématiques) et nos étiquettes ne donnent pas d'information de ce type, soit le lexique ne contient pas le mot à étiqueter (cas des mots de langue étrangère[4]), et dans ce cas la tâche consistera à évaluer quelle est la catégorie la plus probable du mot inconnu en fonction de son contexte, mais rien ne pourra nous permettre d'affirmer qu'il s'agit d'un mot de langue étrangère.

Dans ces cas donc, nous avons fait le choix de nous référer uniquement aux étiquettes fournies aux analyseurs, et aux propriétés de la grammaire. Tous les mots donc, qu'ils soient d'origine étrangère ou non, qu'ils soient des formules mathématiques ou non, on été traités selon les définitions des groupes fournies en section B, même si cela constitue une limite de l'adéquation des résultats de nos parseurs avec l'annotation de référence.

Protocole. — A l'étude détaillée du corpus test, nous avons pu mettre en évidence certains cas dont le traitement demandait un choix, lequel n'était pas spécifié dans PEAS. C'est le cas notamment du traitement des bribes et des amorces : il n'est pas spécifié dans le guide si dans un cas de disfluence, l'on doit considérer chaque occurrence de la répétition (du *reparandum* au *repair*) comme faisant partie du groupe (ce qui donne l'annotation de (3a)), ou alors si l'on ne doit faire figurer dans le groupe que l'occurrence du *repair* (exemple (3b)) :

(3) a. **<NV> il il se tachait </NV>** sa sa **<NV> il ne ne buvait </NV>** que des Blancs

b. il **<NV> il se tachait </NV>** sa sa il ne **<NV> ne buvait </NV>** que des Blancs

Dans un cas comme celui-ci, nous avons donc du faire un choix parmi les possibilités, ne sachant pas lequel de ces choix avait été fait par les annotateurs lors de l'établissement de la référence. En l'occurrence, nous avons choisi de faire figurer toutes les occurrences des répétitions dans les groupes, comme dans l'exemple (3a). Mais si la décision des annotateurs a été différente, alors dans tous ces cas notre annotation sera considérée comme erronée alors qu'il ne s'agit précisément que d'une question de convention et non de justesse d'analyse.

Conclusion

L'utilisation d'une approche symbolique dans une tâche d'annotation de corpus n'est sans doute pas la plus naturelle. Les techniques stochastiques sont en effet parfaitement adaptées à un trai-

[4]Sauf certains latinismes et anglicismes courants. Il serait d'ailleurs intéressant de définir précisément ce que l'on entend par "élément de langue étrangère" dans le cadre du Protocole, pour savoir si des entrées telles que *a priori*, *cool* ou bien *(e)mail*, *erzatz* en font partie ou non.

tement de ce type qui s'appuie sur un style d'annotation fermé. Cependant, les analyseurs symboliques offrent d'autres avantages, en particulier si le formalisme qu'elles utilisent permet une flexibilité de traitement ; c'est le cas des Grammaires de Propriétés dans lesquelles la granularité d'analyse peut être choisie. Ce réglage s'effectue en choisissant le type et le nombre de contraintes à satisfaire. Nous sommes donc en mesure, à partir d'une même grammaire et d'une même stratégie d'analyse, de proposer plusieurs types de traitement offrant des résultats plus ou moins détaillés en fonction des besoins. Là où les approches stochastiques nécessitent un réglage particulier en fonction de chaque tâche d'annotation demandée, une approche symbolique du type de celle décrite ici permet au contraire d'envisager une réutilisabilité à la fois des ressources exploitées (lexique, grammaire), mais également des moteurs utilisés.

Références

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.

Véronique Gendner & Anne Vilnat. Les annotations syntaxiques de référence peas, version 1.6. Révisions par : Laura Monceaux, Patrick Paroubek, Isabelle Robba, 2004.

T. Vanrullen, P. Blache, C. Portes, S. Rauzy, J.F. Maeyhieux, J.M. Balfourier, M.L. Guénot, & E. Bellengier. Une plateforme pour l'acquisition, la maintenance et la validation de ressources lexicales. In *Actes de TALN 2005*, 2005.

Tristan Vanrullen. Analyse syntaxique et granularité variable. In *Actes de RECITAL 2004*, 2004.

Tristan Vanrullen, Marie-Laure Guénot, & Emmanuel Bellengier. Formal representation of property grammars. In *Proceedings of ESSLLI Student Session*, 2003.

TALN 2005, Dourdan, 6-10 juin 2005

Premier bilan de la participation du LORIA à la campagne d'évaluation EASY

Azim Roussanaly , Benoît Crabbé , Jérôme Perrin

LORIA
BP 239, 54506 Vandoeuvre-lès-Nancy Cedex
{Azim.Roussanaly}|{Benoit.Crabbe}|{Jerome.Perrin}@loria.fr

Mots-clés : Analyseur syntaxique, évaluation
Keywords: Parser, evaluation

Résumé

Ce papier décrit LLP2, analyseur à base de grammaires d'arbres adjoints lexicalisés (LTAG) que nous avons utilisé pour le projet EASY (Évaluation des analyseurs syntaxiques) ainsi que les ressources linguistiques associées. Quelques commentaires à propos de cette expérience inspirés des premiers résultats obtenus, sont également présentés.

Abstract

This paper describes LLP2 the Lexicalized Tree Adjoining Grammar (LTAG) based parser we have used for the french evaluation project EASY.as well as associated linguistic resources. Moreover, some comments about this experiment based upon the first results are set out.

1 Analyseur LLP2

1.1 Caractéristiques

L'analyseur du LORIA utilisé pour la campagne EASY s'intitule LLP2. Il s'agit d'un analyseur de type *deep parser* qui s'appuie sur une grammaire d'arbres adjoints lexicalisés (LTAG) (Joshi et al 1975). L'algorithme implémenté est celui de l'analyse par connexité décrit dans (Lopez 1999). L'intégration d'un module de traitement de structures de traits et d'unification, permet de prendre en compte les traits *top* et *bottom* aux nœuds des LTAG. En d'autres termes, LLP2 a la capacité de traiter des *Feature-based* TAG (Vijay-Shanker et al.1988)

Cependant, la version actuelle ne permet pas de prendre en compte les arbres auxiliaires décrivant des adjonctions englobantes (*wrapping adjunction*). Par conséquent, formellement, l'analyse est restreinte aux grammaires d'arbres insérés (TIG) (Schabes et al 1995). LLP2 a été est développée en Java et est disponible sous licence GPL.

1.2 Architecture

LLP2 offre une boîte à outils constituée d'une bibliothèque logicielle et de divers utilitaires. Dans le cadre d'un traitement par lots d'un corpus de phrases, l'architecture est illustré à la Figure 1 :

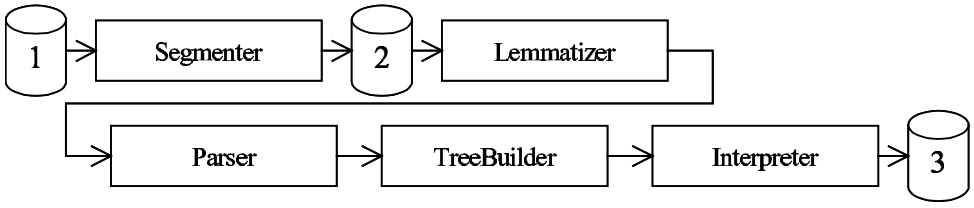


Figure 1 : Architecture LLP2

La liste de phrases à analyser est fournie sous forme d'un fichier texte standard (1). Le pré-processeur *Segmenter* effectue la tokenisation et l'analyse morphologique du texte initial. Le résultat intermédiaire est un liste d'unités lexicales étiquetées (que nous appelons segments) qui peut être stockée dans un fichier intermédiaire selon format XML que nous avons défini (2). Ce fichier est ensuite traité par le processeur *Lemmatizer* qui, en s'appuyant sur les lemmes identifiés lors du pré-traitement, se charge de relier les segments aux arbres élémentaires associés. L'étape suivante consiste à effectuer l'analyse syntaxique. Le résultat de cette étape est un état du *chart* à la fin de l'analyse. L'étape finale consiste à construire les arbres de dérivations et les arbres dérivés pour les analyses complètes et de les stocker dans un fichier résultat (3).

1.3 Adaptations pour la campagne EASY

Une première adaptation a été nécessaire en début de chaîne de traitement (voir Figure 2).

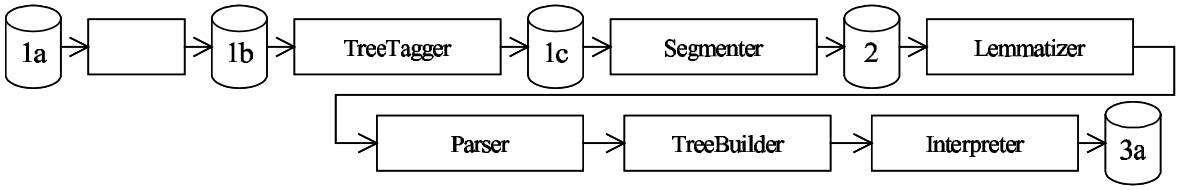


Figure 2 : Adaptation des entrées et des résultats

Ces adaptations sont motivées par :

- la nécessité de prendre en compte le format d'entrée imposé par EASY (textes déjà tokenisés). Il devient alors nécessaire de désactiver la fonction de tokenisation interne du processeur *Segmenter* afin d'éviter d'éventuels conflits avec la segmentation EASY et, de ce fait, de supprimer le risque de produire des résultats difficiles à synchroniser avec ceux attendus par EASY.
- l'usage conjoint d'un lexique morphologique très large et d'un lexique syntaxique peu contraignant (voir à la section suivante) provoquant la multiplication des arbres élémentaires à l'entrée et entraînant en définitive des ambiguïtés multiples et des temps d'analyse rédhibitoires

Ainsi le fichier en entrée devient celui fourni pour la campagne EASY (1a). Ce fichier est traduit dans un format d'entrée compatible avec *TreeTagger*.(Schmid 1994) (1b) dont l'usage permet de réduire les sources d'ambiguïtés. Le résultat de l'étiqueteur (1c) est ensuite

complété par le processeur *Segmenter* qui se contente d'enrichir les traits morphologiques et de remplacer éventuellement les unités inconnues. A ce stade du traitement, on obtient un fichier de segments étiquetés au format XML évoqué précédemment.

LLP2 fournit en résultat des arbres de dérivations et des arbres dérivés tandis que EASY attend un résultat sous la forme de relations de dépendance entre les segments présents dans une phrase. Un processeur (*Interpreter*) a été développé dans le but d'extraire les relations à partir des arbres de dérivations.

Par ailleurs, contrairement à une évaluation effectuée sur la base de TSNLP (Lehmann 1996), l'exploitation des analyses partielles, en cas d'échec d'analyse, permet de proposer certaines relations. Cela nous a conduit à développer un nouveau processeur *TreeBuilder* qui repose sur des heuristiques permettant de sélectionner les analyses partielles pertinentes.

2 Ressources

Du point de vue des ressources, LLP2 s'inspire de l'architecture XTAG (XTAG 1995, Crabbé 2004) qui distingue le lexique morphologique (permettant d'étiqueter les segments et d'identifier les lemmes correspondant), le lexique syntaxique (qui permet la sélection les arbres par filtrage et leur ancrage) et la grammaire (qui contient les arbres TAG).

2.1 Lexique morphologique

Pour la campagne EASY, le lexique morphologique est majoritairement construit à partir de MULTEXT (Ide 1994). Les principales modifications sont le fruit de l'adjonction de traits nécessaires à l'analyseur et à la mise en conformité des mots composés imposés par EASY.

2.2 Lexique syntaxique

C'est dans ce domaine que nous avons constaté les plus grandes lacunes de notre système en raison de l'absence de ressources syntaxiques réellement exploitables. Nous avons tout de même extrait un lexique syntaxique sur la base du lexique fourni par Lionel Clément et utilisé par l'analyseur XLFG (Clément 2001). Malgré quelques aménagements « manuels », cette ressource demeure encore incomplète. Un mécanisme par défaut de sélection des arbres élémentaires sur la base de règles reposant sur les traits morphologiques a dû être mis en place pour pallier les insuffisances du lexique syntaxique.

2.3 Grammaire

La grammaire que nous avons utilisée, a été engendrée à l'aide d'une méta-grammaire conçue par Benoît Crabbé (Crabbé 2005).et « compilé » avec à l'outil XMG développé au LORIA (Duchier et al. 2005). Une méta-grammaire peut être vue comme un moyen compacte d'exprimer une grammaire LTAG. Actuellement, la grammaire traite de manière satisfaisante les verbes et les adjectifs. Mais ce travail est encore en cours et la version utilisée pour la campagne comportait encore de nombreuses imperfections. Très récemment une évaluation avec TSNLP a été effectué avec des résultats encourageants. Des éléments chiffrés seront présentés lors de la session poster. Il aurait été judicieux de refaire les tests pour la campagne EASY avec cette nouvelle version.

3 Conclusion et perspectives

Il est indéniable que notre participation à EASY a été un moteur à nos travaux sur l'analyse syntaxique ; ce qui constitue en soi une expérience positive malgré le fait notre analyseur fournisse encore très peu de relations. Ce qui nous permet d'ores et déjà de penser que l'évaluation EASY sera sans aucun doute très négative. Mais nous pensons que ces résultats peuvent être nettement améliorés à court terme, en utilisant, d'une part, la dernière version de la grammaire qui a été testée sur le TSNLP et, d'autre part, en effectuant des « réglages » sur les stratégies de choix d'analyses partielles afin d'obtenir davantage de relations correctes lorsque l'analyse échoue. Cependant, il nous paraît impossible de parvenir à des résultats satisfaisants sans un effort de développement significatif au niveau du lexique syntaxique.

Nous considérons notre participation à la campagne EASY comme un point de référence de notre système. Nous espérons pouvoir réitérer régulièrement l'expérience afin de mesurer objectivement les améliorations des performances apportées par les solutions mise en œuvre dans le futur. Cette forme d'évaluation est complémentaire à une évaluation de type TSNLP.

Références

L. CLÉMENT: XLFG : A Parser to Learn LFG Framework, *NAACL 2001*, Pittsburgh

CRABBÉ, B, GAIFFE, B ET ROUSSANALY A. : Représentation et gestion de grammaires TAG *Revue TAL* , 2004

B. CRABBÉ :La représentation du lexique syntaxique, le cas de la grammaire d' arbres adjoints, *Thèse de doctorat Université Nancy2*, 2005 (à paraître)

D. DUCHIER, J. LE ROUX, Y. PARMENTIER : XMG, un compilateur de méta-grammaire extensible, *TALN 05* Dourdan, Juin 2005

N. IDE N., J. VÉRONIS:.MULTEXT (Multilingal Tools and Corpora*) COLING'94* Kyoto Japan 90-96, 1994

A. JOSHI, L. LEVI, M. TAKAHASHI : Tree Adjunct Grammars, *Journal of Compuper and System Sciences*, 1975

S. LEHMANN, D. ESTIVAL, S. OEPEN :N. TSNLP - Des jeux de phrases-test pour l'évaluation d'applications dans le domaine du TALN, *TALN* 96, Mai 1996 Marseille

P. LOPEZ : Analyse d'énoncés oraux pour le dialogue homme-machine à l'aide de grammaires lexicalisés d'arbres, *Thèse de doctorat UHP- Nancy1*, 1999

Y. SCHABES, R.C. WATERS : Tree Insertion Grammar : A Cubic Time Parsable Formalism that Lexicalizes Context-Free Grammar without Changing the Trees Produced, *Computational Intelligence*, vol. 21, 479-514, 1995

H. SCHMID: Probabilistic Part-of-Speech Tagging Using Decision Trees, *International Conference on New Methods in Language Processing,* 1994

K. VIJAY-SHANKER, A. JOSHI: A Feature-based Tree Adjoining Grammar *COLING*'88, Budapest, 1988,

THE XTAG RESEARCH GROUP: A lexicalized tree adjoining grammar for English, *Technical Report IRCS Report 95-03*, The Institute for Research in Cognitive Science, Univ. of Pennsylvania, 1995

L'analyseur syntaxique Vergne-98 présenté aux actions d'évaluation GRACE et EASy

Jacques Vergne, Frédérick Houben

GREYC – Université de Caen
BP 5186 -14032 Caen cedex
{Jacques.Vergne, Frederick.Houben}@info.unicaen.fr
www.info.unicaen.fr/~jvergne

Mots-clés : analyseur syntaxique de phrases en français, analyseur déterministe, moteur à base de règles, constituants non récursifs

Keywords: French sentence syntactic parser, deterministic parser, rule based parser, non recursive constituents

Résumé Nous présentons courtement l'analyseur syntaxique Vergne-98, participant aux actions d'évaluation GRACE et EASy : les principaux concepts mis en œuvre, ainsi que les post-traitements nécessaires à l'action EASy.

Abstract We briefly present the syntactic parser Vergne-98, involved in the evaluation actions GRACE and EASy : main implemented concepts, and post-processing necessary to the EASy evaluation action.

Introduction

L'analyseur Vergne-98 est un analyseur syntaxique de phrases écrites en français; il est déterministe, et utilise des moteurs à base de règles; il est fondé sur une hiérarchie de constituants non récursifs : tokens, chunks et phrases. Il a été conçu et développé de 1985 à 1998. Il a d'abord été combinatoire[1] (démonstration au CoLing 1990 à Helsinki : Vergne, 1990), puis, à partir de 1993, il est devenu déterministe et de complexité pratique linéaire[2].

Cet analyseur est un logiciel d'étude, c'est-à-dire un moyen d'expérimenter et d'évaluer des concepts. Il a obtenu la meilleure évaluation à l'action GRACE (décision : 100%, précision : 94,5%), action d'évaluation des étiqueteurs du français, cette validation opératoire étant aussi une validation des concepts. On trouvera une présentation plus détaillée de cet analyseur dans notre mémoire d'Habilitation à Diriger des Recherches (Vergne, 1999, 3.3).

---

[1] Et donc de complexité théorique exponentielle (processus de parcours de l'arbre des catégories possibles des tokens), avec un nombre imprévisible de solutions (aucune ou beaucoup).

[2] Ce qui a été rendu possible par l'abandon des concepts de syntagme récursif, et de structures syntaxiques attendues représentées par une grammaire formelle (Vergne, 2001).

Il produit en sortie deux fichiers de résultats : les tokens étiquetés (sortie GRACE), et les chunks étiquetés et reliés par des relations de dépendance, de coordination, et d'antécédance des pronoms relatifs (sortie EASy). Ces deux fichiers utilisent les résultats complets de l'analyse (la mise en relation des chunks complète l'étiquetage des tokens), et chaque unité a une catégorie unique (toujours 100 % de décision, propriété d'un analyseur déterministe).

1 Principaux concepts

1.1 Hiérarchie de constituants non récursifs et stratégie d'analyse dans cette hiérarchie

Les constituants sont non récursifs, et forment une hiérarchie dont chaque niveau est typé : constituants physiques : texte, phrases, tokens, et constituants calculés : les chunks. On notera l'absence des concepts de proposition et de fonction dans la proposition (sinon implicitement dans les relations de dépendance sujet-verbe et verbe-objet).

1.1.1 Segmentation descendante dans la hiérarchie des constituants

Le texte est segmenté en phrases, par un traitement contextuel des points (point final de phrase, ou point d'abréviation). Puis les phrases sont segmentées en tokens. Un token peut être un mot, une ponctuation, un groupe de mots (locution, nombre composé), ou une partie de mot (les amalgames sont traités comme deux tokens : préposition+déterminant).

1.1.2 Analyse montante dans la hiérarchie des constituants

Les tokens sont étiquetés avec les ressources lexicales et des règles de déduction contextuelle (350 règles). Les chunks sont délimités et typés. La structure de la phrase est alors exprimée sous la forme de chunks nominaux ou verbaux, et de tokens externes aux chunks : conjonctions, pronoms relatifs, pronoms sujets, prépositions, adverbes de phrase, et ponctuations. C'est sur cette structure qu'est calculée la mise en relation des chunks.

1.2 Catégories de token et catégories de chunk

Le jeu des catégories de token est distributionnel : une catégorie regroupe les tokens d'une classe distributionnelle de tokens[3]. Une catégorie appartient soit au chunk nominal, soit au chunk verbal ou bien est externe au chunk; elle a une position à l'intérieur du chunk nominal ou verbal : en début ou fin de chunk, après ou avant telle autre classe. Le caractère distributionnel du jeu des catégories est la base sur laquelle repose la régularité des déductions contextuelles à l'intérieur d'un chunk nominal ou verbal. Ceci conduit à dissocier des catégories habituellement réunies. Par exemple, les adjectifs sont subdivisés en épithètes antéposées, postposées dans le chunk nominal, attributs dans le chunk verbal ; les «adjectifs» possessifs et démonstratifs sont inclus dans les déterminants, dans le chunk nominal.

1.3 Ressources

1.3.1 Ressources lexicales

Le lexique partiel (190 Ko) contient les mots grammaticaux, des adjectifs le plus souvent antéposés, les adverbes qui ne se terminent pas par *-ement*. Les radicaux verbaux sont dans un

[3] Pour des informations plus précises, cf. : www.info.unicaen.fr/~jvergne/cat_mots_SNR.html

fichier à part (45 Ko). Des règles sur les finales complètent les ressources lexicales (16 Ko) pour les noms, les adverbes en *-ement* et les néologies verbales (Vergne, 1999, 3.3.3).

1.3.2 Ressources syntaxiques

Les règles de syntaxe (120 Ko, environ 550 règles), sont interprétées par deux moteurs : le premier opère sur la structure en tokens, le deuxième sur la structure en chunks + tokens externes aux chunks (cf. 1.5 ci-dessous).

1.3.3 Collaboration entre ressources lexicales et ressources syntaxiques au niveau des tokens

Les ressources lexicales affectent une catégorie, le plus souvent unique par défaut pour chaque token. L'ensemble de catégories d'un token peut ensuite être modifié par des règles de déduction contextuelle. Ces règles sont affirmatives : dans tel contexte, ce token a telle(s) catégorie(s), ce qui permet d'étiqueter un token n'appartenant pas aux ressources lexicales, ou ayant localement une catégorie inhabituelle (Vergne, Giguet, 1988).

Chronologie de l'application des ressources lexicales et des paquets de règles syntaxiques :

- étiquetage par le lexique des mots grammaticaux;

- premier paquet : affectation d'une catégorie générique «token de chunk nominal» ou «token de chunk verbal» aux tokens suivant un mot grammatical étiqueté par le lexique (15 règles) ;

- étiquetage par les radicaux verbaux, et par les règles sur les finales («guesser»), par union avec les catégories posées précédemment ;

- deuxième paquet : suppression des polycatégories éventuelles par intersection des catégories posées par les règles avec les catégories posées précédemment ; propagation du non-désaccord genre-nombre dans le chunk nominal, et pose d'une frontière de chunk en cas de désaccord (environ 300 règles).

1.4 Mise en relation des chunks

La mise en relation des chunks (Vergne, 1999, 2.3.2) modélise une saturation de valence généralisée, en reprenant le concept de Tesnière, mais en l'appliquant non pas aux mots mais aux chunks, et en ne la limitant pas aux valences verbales, par généralisation à toute relation de dépendance ou de coordination entre deux chunks :

- premier temps : valence détectée et à saturer ; un premier chunk d'un type T1 donné est mis en attente d'un deuxième chunk d'un type T2 donné pour le type de relation caractérisé par la valence à saturer ;

- deuxième temps : saturation d'une valence ; arrivée du deuxième chunk de type T2, mise en relation des deux chunks, oubli de la valence maintenant saturée.

Ces deux temps de la mise en relation nécessitent donc deux règles pour chaque mise en relation. Les chunks en attente sont mémorisés dans une pile pour chaque type d'attente : un chunk nominal attend un chunk verbal (relier sujet-verbe), un chunk verbal attend un chunk nominal (relier verbe-sujet postposé ou relier verbe-objet), un chunk nominal attend un chunk nominal coordonné, ... (13 piles, 250 règles). Ce processus permet de relier deux chunks sans aucun attendu de structure sur ce qui les sépare; il fonctionne de la même manière qu'ils soient contigus ou éloignés. Tout processus de mise en relation P1 peut interagir avec un autre processus de mise en relation P2 par une action sur la pile de P2 (Vergne, 1999, 3.3.6) ; par exemple, un chunk sujet n'attend plus de coordonné après sa mise en relation avec son verbe.

1.5 Règles symboliques et moteurs à base de règles

Les règles sont de la forme : conditions => actions, portant des deux côtés sur un même nombre d'unités (tokens ou chunks). Les conditions portent sur les catégories, les attributs, les graphies, les lemmes verbaux, les relations (dépendance ou coordination), la présence dans une pile, un opérateur de non désaccord du token courant avec le token précédent (genre-nombre dans les chunks nominaux, personne-nombre dans les chunks verbaux), un opérateur de non désaccord du chunk courant avec un chunk empilé (personne-nombre dans la relation sujet-verbe), un opérateur d'isomorphisme entre chunks à coordonner. On utilise les opérateurs booléens (*non ou et*). L'utilisation du *non* permet d'écrire des règles exclusives les unes des autres, et donc moins sensibles à leur ordre d'évaluation. Les actions sont : affectation d'une valeur de catégorie, d'attribut, empiler ou dépiler un chunk, relier deux chunks. Pour chaque unité de la phrase (token ou chunk), chaque moteur passe les règles du paquet courant sur l'unité courante et son contexte, d'où une complexité théorique et pratique linéaire selon le nombre d'unités (Vergne, 1999, 3.3.5).

2 Quelques caractéristiques de l'implémentation

Cet analyseur est écrit en Pascal sur Mac OS, et est en cours de réécriture en java pour faciliter sa portabilité. Les sources font 1,4 Mo, et l'exécutable 460 Ko. Les ressources lexicales et syntaxiques font ensemble 370 Ko. L'analyseur complet fait donc 830 Ko au total.

3 Adaptation des sorties de l'analyseur au format d'entrée EASy

La sortie des chunks étiquetés et reliés, avec leurs tokens internes et les tokens externes, est sous la forme d'un fichier texte (cf. : www.info.unicaen.fr/~jvergne/format_prs.html). Ce format est ensuite transcodé en un fichier XML, sans aucune perte d'information (étiquettes, et valeur des attributs), qui, associé à des feuilles de style, permet des regards différents sur les résultats. Ce premier fichier XML constitue un format de description pivot, qui est ensuite projeté dans le format XML EASy. Cette projection n'est pas seulement un changement de notation : la segmentation en chunks est différente ; par exemple, l'adverbe postposé au verbe est inclus dans le chunk verbal dans notre format, alors que, dans le format EASY, il constitue à lui seul un chunk adverbial dépendant du chunk verbal. De plus, certaines informations ne doivent pas figurer dans le format EASY : catégorie et attributs des tokens, attributs de genre, nombre et personne des chunks.

Références

VERGNE J. (1990), A parser without a dictionary as a tool for research into French syntax, Actes de *CoLing 1990*, vol. 1, 70-72. (www.info.unicaen.fr/~jvergne/JVergneColing1990.pdf)

VERGNE J., GIGUET E. (1998), Regards Théoriques sur le "Tagging", Actes de *TALN 1998*, 22-31. (www.info.unicaen.fr/~jvergne/VergneGiguetTaln98.pdf)

VERGNE J. (1999), *Étude et modélisation de la syntaxe des langues à l'aide de l'ordinateur - Analyse syntaxique automatique non combinatoire*, Habilitation à Diriger des Recherches, Université de Caen. (www.info.unicaen.fr/~jvergne/HDR_J.Vergne.pdf)

VERGNE J. (2001), Analyse syntaxique automatique de langues : du combinatoire au calculatoire, Actes de *TALN 2001*, 15-29. (www.info.unicaen.fr/~jvergne/Taln2001FR_JV.pdf)

« Simple comme EASy :-) »

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 présente les deux systèmes d'analyse déployés par le projet ATOLL (INRIA) lors de la campagne EASy d'Évaluation d'Analyseurs Syntaxiques (décembre 2004). Nous donnons quelques résultats quantitatifs en termes de couverture et de temps d'analyse. Cette expérience permettra la comparaison à grande échelle du résultat de différents analyseurs, mais montre aussi que les techniques d'analyse syntaxique profonde sont désormais à même de traiter des corpus volumineux tout en conservant leur puissance d'expression linguistique.

Abstract This paper presents both parsing systems used by project ATOLL (INRIA) for the EASy parsing evaluation campaign (December, 2004). We give a few quantitative results in terms of coverage and parsing time. These experiments will allow the comparison of parsing results on huge corpus, but also show that deep parsing techniques can cope with large corpus while preserving their linguistic expressive power.

1 Introduction

Pour la campagne EASy d'Évaluation des Analyseurs Syntaxiques, l'équipe ATOLL de l'INRIA a déployé deux chaînes de traitement syntaxiques (Boullier *et al.*, 2005a). L'analyse d'environ 35000 phrases fournies par les organisateurs nous permet de présenter ici quelques résultats préliminaires en attendant les résultats définitifs.

Les phrases se répartissaient en un ensemble de corpus non retravaillés couvrant divers styles de langage et présentant toutes sortes de problèmes à régler, bien en amont de l'analyse syntaxique proprement dite. En particulier, la segmentation en phrases et en tokens n'était pas toujours justifiée linguistiquement, et par conséquent pas toujours compatible avec nos outils. Il a fallu adapter nos programmes pour leur permettre de re-segmenter les corpus tout en préservant au mieux la segmentation fournie qui devait être celle des résultats.

En sortie d'analyse syntaxique s'est posé le problème de convertir nos sorties au format attendu par les organisateurs (Gendner & Vilnat, 2004), donnant des informations sur des constituants

simples non récursifs et sur un jeu de dépendances. Malgré le fait que nos analyseurs produisent des analyses ambiguës, nous avons essayé de rendre des résultats non-ambigus sur les constituants et les dépendances en nous appuyant sur quelques heuristiques.

2 Les systèmes FRMG et SxLFG

Nos deux systèmes s'appuient sur deux formalismes syntaxiques « profonds » différents :
- le système FRMG[1] (Thomasset & de la Clergerie, 2005) repose sur une grammaire TAG compacte constituée de quasi-arbres sous-spécifiés générés automatiquement à partir d'une méta-grammaire ; cette grammaire est compilée par le constructeur d'analyseurs syntaxiques DyALog pour produire un analyseur,
- le système SxLFG (Boullier *et al.*, 2005b) repose sur une grammaire LFG qui est une évolution de celle citée par (Clément & Kinyon, 2001) ; cette grammaire est utilisée par le générateur SxLFG d'analyseurs LFG qui est lui-même fondé sur l'environnement SYNTAX de génération d'analyseurs non contextuels.

Les deux systèmes utilisent le même lexique *Lefff* (Sagot *et al.*, 2005) et la même chaîne SxPipe de traitement pré-syntaxique (Sagot & Boullier, 2005), ces deux composants étant d'une importance considérable. En particulier, SxPipe s'est révélé essentiel pour traiter au mieux les corpus fournis, avec leurs fautes, bizarreries typographique et artefacts divers. Notons que SxPipe produit une sortie ambiguë sous forme de treillis de mots.

Enfin, chaque système a dû mettre en place des procédures distinctes de désambiguïsation et de conversion des sorties des analyseurs vers le format demandé par le Guide d'annotation EASy.

3 Résultats et comparaisons préliminaires

Les résultats en temps d'analyse et en couverture[2] pour nos deux systèmes sont détaillés dans les tableaux 1, 2 et 3. Les taux de couverture pour des analyses complètes[3] sont comparables entre les systèmes mais varient significativement selon les types de corpus. Une analyse fine des temps d'analyse permet de montrer que SxLFG est beaucoup plus rapide pour les phrases courtes (moins de 15 tokens environ)[4], mais que la polynomialité du formalisme TAG permet à FRMG d'être plus efficace pour les longues phrases. Ces conclusions sont toutefois à nuancer par le fait que les deux systèmes d'analyse sont très récents, et continuent à être améliorés dans des proportions importantes, permettant d'envisager une forte diminution des temps d'analyse.

Puisque nous disposons de deux jeux de résultats au format EASy, nous avons mené quelques expériences préliminaires pour les comparer, en se focalisant sur les constituants (les dépen-

[1] utilisable en ligne sur `http://atoll.inria.fr/parserdemo`.

[2] Pour la définition précise de ce que signifient les expressions *couverture de la CFG support*, *couverture avec/sans vérification de cohérence*, le lecteur est invité à se reporter à (Boullier *et al.*, 2005a).

[3] Les 2 systèmes sont robustes et fournissent également des résultats partiels quand une analyse complète n'est pas trouvée.

[4] On notera aussi l'extrême efficacité de la partie CFG de l'analyse effectuée par SxLFG. Cette partie, qui repose sur le système SYNTAX, permet par exemple de trouver en moins de 6 secondes les 5.10^{52} analyses CFG d'une des phrases du corpus de courrier électronique, et ne met que 2 minutes environ à construire une forêt partagée représentant 3.10^{73} analyses CFG après rattrapage d'erreur pour une autre phrase du même corpus qui n'a pas d'analyse CFG correcte.

« Simple comme EASy :-) »

Corpus	#phrases	% couv.	temps d'analyse					amb.
			moy.	méd.	$\geq$ 1s	$\geq$ 10s	Timeout	
general	6160	41.01%	10.31s	2.44s	71.19%	18.01%	5.27%	0.65
litteraire	7960	38.93%	5.59s	2.10s	71.75%	9.64%	1.80%	0.53
mail	7962	32.83%	4.37s	1.46s	59.65%	7.08%	1.27%	0.70
medical	2225	44.00%	5.47s	1.46s	63.26%	8.21%	2.40%	0.70
oral	6892	44.39%	5.58s	1.19s	54.58%	6.39%	0.78%	0.53
questions	3509	66.28%	3.47s	1.32s	66.04%	4.53%	1.08%	0.58
Total	**34438**	**42.45%**	**5.55s**	**1.61s**	**64.41%**	**9.32%**	**2.07%**	**0.60**

TAB. 1 – Résultats pour FRMG (*time-out*=100s)

corpus	#phrases	couv. de la CFG support	couverture sans vérif. de cohérence	couverture avec vérif. de cohérence	temps médian	*timeout* (t $\geq$ 15s)
general	6952	89.24%	57.03%	32.42%	0.54s	22.64%
litteraire	11408	89.92%	69.07%	40.52%	0.07s	13.03%
mail	9308	83.02%	66.08%	40.18%	0.01s	9.53%
medical	2553	87.34%	60.95%	39.99%	0.06s	12.10%
oral	7075	85.24%	67.94%	46.47%	0.01s	8.30%
questions	3563	92.73%	80.33%	62.19%	0.01s	5.22%
Total	**40859**	**87.51%**	**66.62%**	**41.95%**	**0.03s**	**12.31%**

TAB. 2 – Résultats pour SXLFG (*time-out*=15s)

dances étant plus délicates à comparer). Nos deux systèmes produisent exactement la même analyse en constituants pour 7714 phrases. La phrase la plus longue sur laquelle nous soyons d'accord a 42 tokens EASy, la moyenne étant d'environ 7 tokens. Le tableau 4 indique les distributions des différents constituants pour chaque système et celle des constituants communs aux deux (même type et même couverture de tokens).

4 Conclusion

Naturellement, cette brève présentation n'est pas complète car il manque les informations sur le taux de précision des analyses fournies. Nous avons déjà repéré de nombreux problèmes, certains liés à la segmentation et à la difficulté des corpus, de nombreux autres liés à la couverture des grammaires et aux processus de désambiguïsation et de conversion au format EASy. Néanmoins, nous tirons déjà trois principaux enseignements de cette campagne :

		Corpus complet	Phrases valides pour la CFG support	
		Analyse CFG	Analyse CFG	Analyse complète
Corpus	#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	med	0.00s	0.00s	0.03s
	$\geq$ 0.1s	1.79%	1.20%	42.2%
	$\geq$ 1s	0.24%	0.09%	29.0%
Nombre d'analyses	med - max	32 028 - 3.10^{73}	29 582 - 5.10^{52}	1 - 1
	$\geq 10^6$	36.13%	35.28%	0%
	$\geq 10^{12}$	8.86%	7.84%	0%

TAB. 3 – Détail des temps d'analyse pour SXLFG (*time-out*=15s)

Groupes	GA	GN	GP	GR	NV	PV
FRMG	27843	96880	71332	25833	90166	6953
SXLFG	28128	118321	61392	34229	85564	6823
communs	16182	56281	42656	15728	57460	1061

TAB. 4 – Comparaison par types de constituants entre FRMG et SXLFG

- elle n'évalue pas uniquement la phase d'analyse syntaxique, mais également le lexique, la chaîne de traitement pré-syntaxique, et la phase d'extraction des annotations EASy à partir des sorties des analyseurs ;
- elle permettra des travaux intéressants de comparaison des résultats d'analyseurs différents, notamment pour la constitution de corpus annotés mais aussi pour la détection d'erreurs et de manques dans les grammaires, les lexiques et la chaîne de traitement pré-syntaxique ;
- elle montre que les technologies d'analyse syntaxique profonde retenues par ATOLL sont suffisamment efficaces pour permettre le traitement de gros corpus.

Les deux derniers points nous confortent dans nos choix de recherche. En effet, la plupart des traitements linguistiques actuels sur gros corpus reposent sur des technologies de surface (probabilités, chunks, automates finis) choisies pour leur efficacité algorithmique, mais ne permettant pas une modélisation linguistiquement satisfaisante des mécanismes de la langue. À l'inverse, les technologies d'analyse profonde permettent une telle modélisation, seule à même de permettre à moyen terme d'effectuer avec un haut degré de précision des tâches linguistiquement complexes telles que la traduction automatique. Nos expériences montrent que leurs points faibles traditionnels sont de moins en moins pertinents : les analyseurs vont de plus en plus vite et les grammaires peuvent être développés de plus en plus rapidement, en particulier grâce aux développements récents autour du concept de méta-grammaires. De plus, le développement des lexiques riches en information qui sont nécessaires peut être facilité par l'examen statistique des sorties d'analyses profondes de corpus suffisamment importants.

Références

BOULLIER P., CLÉMENT L., SAGOT B. & VILLEMONTE DE LA CLERGERIE E. (2005a). Chaînes de traitement syntaxique. In *Proceedings of TALN'05*, Dourdan, France.

BOULLIER P., SAGOT B. & CLÉMENT L. (2005b). Un analyseur LFG efficace : SXLFG. In *Actes de TALN'05*, Dourdan, France.

CLÉMENT L. & KINYON A. (2001). XLFG-an LFG parsing scheme for French. In *Proc. of LFG'01*.

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`.

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.

TALN 2005

12ème conférence annuelle
sur le
Traitement Automatique des Langues Naturelles

POSTERS INVITÉS

CAMPAGNE D'ÉVALUATION QUESTION-RÉPONSE

TECHNOLANGUE-EVALDA-EQUER

Campagne d'évaluation EQueR-EVALDA
Evaluation en question-réponse

Christelle Ayache (1), Brigitte Grau (2), Anne Vilnat (2)

(1) Evaluations and Language resources Distribution Agency (ELDA)
55-57, rue Brillat Savarin 75013 Paris
ayache@elda.org
(2) Groupe LIR - Langues Information et Représentation - LIMSI
BP 133, 91403 Orsay Cedex
{brigitte.grau, anne.vilnat}@limsi.fr

Mots-clés : campagne d'évaluation, système de question-réponse, type de question, type de réponse, fichier-résultats (soumission).

Keywords : evaluation campaign, question-answering system, question type, answer type, run.

Résumé - Abstract

Cet article présente dans son ensemble la campagne d'évaluation EQueR-EVALDA. Cette campagne a bénéficié d'une aide du Ministère délégué à la Recherche dans le cadre de l'action Technolangue[1].

This paper describes the EQueR-EVALDA Evaluation Campaign. This campaign is supported by the French Ministry of Research within Technolangue.

1 Introduction

La campagne EQueR a offert un cadre d'évaluation aux systèmes de question-réponse pour la langue française, avec l'objectif d'alimenter l'activité de recherche dans le domaine en fournissant une photographie de l'état de l'art, notamment en France.

[1] L'action Technolangue est une action interministérielle destinée à mettre en place de manière pérenne une infrastructure de production et diffusion de ressources linguistiques.

EQueR a proposé deux tâches de recherche automatique de réponses : une tâche générique sur une collection hétérogène de textes – en large partie des articles de presse – et une tâche spécifique, liée au domaine médical, sur une collection de textes de cette spécialité.

L'esprit de la campagne EQueR correspondait davantage à une réflexion collective qu'à une véritable compétition ; néanmoins, aucune intervention manuelle n'a été autorisée pour la recherche et l'extraction des réponses.

2 Spécifications de la campagne

2.1 Collections de documents

Les participants ont eu accès aux collections textuelles quelques mois avant le test d'évaluation, après avoir rempli un accord d'utilisation finale des données. Les données ont été fournies sous forme de DVD ainsi que par téléchargement.

Les textes fournis étaient composés d'un balisage simple avec un identifiant de document, de titre et de paragraphe, et codés en ISO-Latin-1 (ISO-8859-1). Voici un exemple extrait du corpus au format EQueR :

```
<DOC>
<DOCID>LEMONDE95-000001</DOCID>
<LEAD1>DIMANCHE 01 JANVIER 1995 : NAISSANCE DE L'OMC,
ORGANISATION MONDIALE DU COMMERCE</LEAD1>
<TITLE>Un commerce mondial mieux réglementé</TITLE>
<P> AVEC l'année 1995, une nouvelle institution voit le
jour, qui devrait être porteuse de plus de justice
économique : l'Organisation mondiale du commerce(OMC).
Aux pays soumis à la dure concurrence internationale et à
ses coups bas, l'OMC apporte l'espoir qu'aux rapports de
force vont se substituer progressivement des
rapports</P></DOC>
```

Les textes originaux, avec leur balisage propre, ont également été mis à la disposition des participants.

Deux collections ont été élaborées : une collection pour la tâche « générale » et une collection pour la tâche « spécialisée ».

La collection générale, d'une taille d'environ 1,5 Go, était composée d'articles de presse de plusieurs années des journaux *Le Monde* et *Le Monde Diplomatique*, de dépêches de presse et de rapports d'information du Sénat français portant sur des sujets très variés.

La collection de textes de spécialité, d'une taille d'environ 140 Mo, était composée principalement d'articles scientifiques et de recommandations de bonne pratique médicale, sélectionnés par le CISMeF (Catalogue et Index des Sites Médicaux Francophones) du Centre Hospitalier Universitaire de Rouen.

2.2 Questions

Cinq types de questions ont été proposés aux systèmes participants : des questions « factuelles simples » (comprenant 7 catégories : personne, organisation, date, lieu, mesure, manière et objet/autre, « Qui est le président du Chili ? », « Quand a eu lieu le festival d'Avignon ? », etc.), des questions de type « définition » (autrement dit dont la réponse attendue est une définition, comprenant deux catégories : personne et organisation, « Qu'est-ce que l'OTAN ? », « Qui est Salvador Dali ? », etc.), des questions de type « liste » (« Quelles sont les 4 religions pratiquées en Hongrie ? »), des questions de type « oui/non » (« Existe-t-il une ligne de TGV Paris-Valencienne ? ») ainsi que des « reformulations » de questions factuelles simples déjà présentes dans le jeu de test.

Des questions sans réponse possible dans les collections de documents ont été introduites au sein du corpus de questions « général ». Dans ce cas, le système devait renvoyer en réponse : la valeur « NIL ».

Le type des questions était indiqué par un codage d'identification attribué à chaque question. Les identifiants de classes étaient : F (factuelle simple), D (définition), L (liste), et B (oui/non). Un « R » (reformulation) a été ajouté à l'identifiant de classe si nécessaire. Ci-après, un exemple de codage d'une question : « GF18 Où est né Jacques Chirac ? ». Ce codage indique que la question n°18 est de type factuel simple (F) et s'applique à la tâche générale (G).

Les sources et les modes de génération des questions ont été diversifiés. Une partie a été dérivée de mots clés qui accompagnaient les articles et les dépêches de presse, une autre partie a été créée par un groupe d'utilisateurs potentiels, dont certains connaissaient le domaine du TAL. La présence d'au moins une bonne réponse a été vérifiée manuellement dans le corpus pour chaque question proposée aux participants (hormis pour les questions « NIL »).

Pour la tâche générale, ELDA a élaboré un corpus de 500 questions réparties comme suit : 407 « factuelles », 32 « définitions », 31 « listes » et 30 « oui/non ».

Pour la tâche spécialisée, l'équipe du CISMeF a élaboré un corpus de 200 questions réparties comme suit : 81 « factuelles », 70 « définitions », 25 « listes » et 24 « oui/non ».

2.3 Réponses

Pour chaque question, les systèmes pouvaient renvoyer soit une réponse courte exacte, un passage (moins de 250 caractères contigus extrait d'un document de la collection) et un identifiant de document justifiant de cette réponse et de ce passage, soit au moins un passage et un identifiant de document justifiant ce passage.

Pour chaque type de questions (sauf pour les questions de type « liste), les systèmes pouvaient renvoyer jusqu'à cinq réponses ordonnées (20 pour les questions de type « liste »). Les réponses (ordonnées) devaient être présentées dans l'ordre des questions. Concernant les réponses de type « oui/non », les systèmes devaient pouvoir justifier du passage que ce soit pour une réponse positive ou négative.

3 Evaluation

La phase d'évaluation des différents systèmes a eu lieu sur chacun des sites des participants, et a duré une semaine, du 16 au 23 juillet 2004.

3.1 Réponses courtes et passages

La majeure partie des systèmes participants ont renvoyé un passage et une réponse courte exacte (un seul groupe a fait le choix de ne pas être évalué sur les réponses courtes). Les deux types de réponses ont été évalués distinctement.

En accord avec les participants lors de l'évaluation, deux sortes de jugements ont été appliqués, l'un porte sur les réponses courtes, l'autre sur les passages.

Concernant les réponses courtes, quatre types de jugements étaient possibles, la réponse était soit « correcte » (réponse juste et la plus précise possible, c'est-à-dire sans information obsolète), soit « inexacte » (réponse juste, mais pas assez précise, soit il manquait de l'information, soit au contraire de l'information avait été ajoutée), soit « incorrecte » (la réponse n'était pas juste, elle ne contenait pas la réponse attendue), soit « non justifiée » (la réponse était juste et exacte mais le document associé à la réponse ne justifiait pas celle-ci) . Pour l'évaluation des passages, seuls deux jugements étaient possibles : le passage était « correct » s'il contenait la réponse à la question et était justifié par le document associé, sinon, il était jugé « incorrect ».

Lorsqu'un système renvoyait « NIL », il s'agissait d'évaluer cette réponse comme si on évaluait un passage. Tout d'abord, vérifier que cette question était bien supposé renvoyer « NIL » ; si c'était le cas, le passage était jugé « CORRECT » ; sinon il était jugé « INCORRECT ».

3.2 Mesures adoptées

Pour les questions de type « factuel », « définition » et « oui/non », la mesure que nous avons adoptée est la Moyenne des Réciproques du Rang (MRR). Ce critère tient compte du rang de la première bonne réponse trouvée (métrique TREC[2]). Si une bonne réponse est trouvée plusieurs fois, elle n'est comptée qu'une seule fois.

$$MRR = \frac{1}{\text{nb questions}} \sum_{i=1}^{\text{nb questions}} \frac{1}{answer_i\ rank}$$

Pour les questions de type « liste », la mesure que nous avons adoptée est la précision moyenne (*non interpolated average precision, NIAP*, métrique TREC). Ce critère tient compte à la fois du rappel (pourcentage de bonnes réponses présentes dans la liste parmi toutes les bonnes réponses à trouver) et de la précision (pourcentage de bonnes réponses trouvées parmi toutes les réponses trouvées) mais aussi de la position des bonnes réponses dans la liste.

[2] TREC, Text REtrieval Conference, http://trec.nist.gov/

$$\text{prec\_moy}(q_i) = \frac{\sum_{j=1}^{j=n} I(rep_j).prec(j)}{R} \leq 1$$

avec :

$$I(rep_j) = \begin{cases} 1 \text{ si } rep_j \text{ est une bonne réponse} \\ 0 \text{ si } rep_j \text{ est une mauvaise réponse ou une réponse déjà proposée} \end{cases}$$

et :

$$\text{prec}(j) = \frac{\sum_{k=1}^{j} I(rep_k)}{j} = \frac{\text{Nombre de bonnes réponses différentes jusqu'au rang } j}{j} \leq 1$$

4 Résultats

4.1 Tâche générale

Sept groupes ont participé à la tâche générale de la campagne EQueR. Quatre laboratoires publics : le LIMSI, l'Université de Neuchâtel, le Laboratoire d'Informatique d'Avignon en collaboration avec iSmart et le CEA-LIST/LIC2M. Ainsi que trois institutions privées : France Télécom R&D, Synapse Développement et Sinequa.

Les travaux de la plupart de ces systèmes sont présentés plus en détail dans les pages suivantes des actes de l'atelier.

Au total, douze runs (ou fichier-résultats) ont été évalués. Deux juges ont évalué les résultats pendant un mois. De nombreuses discussions et mises au point ont permis d'optimiser la cohérence inter-juges.

Parmi les 500 questions du corpus de départ, cinq comportaient des erreurs. Nous avons décidé de supprimer ces cinq questions du corpus ainsi que de l'ensemble des fichier-résultats. Les scores ont donc été calculés sur la base de 495 questions réparties comme suit : 400 « factuelles », 33 « définitions », 31 « oui/non » et 31 « listes ».

Les trois systèmes de question-réponse ayant obtenu les meilleurs résultats pour la tâche générale lors de la campagne EQueR/EVALDA 2004 sont : pour les passages, les systèmes de Synapse Développement (participant 5), de Sinequa (participant 4), et du LIMSI (participant 2) ; Pour les réponses courtes : les systèmes de Synapse Développement, du LIA (participant 6) et du LIMSI.

Les résultats ont été fournis aux participants sous forme de deux tableaux, respectivement pour les réponses courtes et les passages. Le premier (Figure 1) présente pour chaque fichier-résultats, le nombre de questions traitées, le nombre de passages (ou réponses) corrects renvoyés, ainsi que les scores obtenus pour chaque type de question et combinaison. Le second (Figure 2) présente pour chaque fichier-réponse, un détail sur les passages (ou réponses) corrects renvoyés en indiquant le nombre de passages (ou réponses) corrects par type de réponse attendue (personne, temps, lieu, organisation…).

Identifiant du run	Nb de questions répondues [464]	Nb passages corrects	Nb passages incorrects	MRR sur tout les types sauf Listes	MRR sur Factuelles et Définitions	MRR sur Factuelles	MRR sur Définitions	MRR sur Oui/ Non	NIAP sur Listes	Nb de NIL renvoyés en rang 1	NIL	
											Préci-- sion	Rap-- pel
participant 5	464	378	86	0.7	0.71	0.7	0.74	0.67	0.29	4	1	0.8
participant 4	464	237	227	0.37	0.37	0.36	0.55	0.32	0	20	0.05	0.2
participant 2	464	210	254	0.37	0.38	0.37	0.47	0.25	0.09	69	0.01	0.2
participant 6	388	182	206	0.33	0.32	0.31	0.43	0.38	0.08	0	0	0
participant 3	464	184	280	0.31	0.31	0.3	0.43	0.35	0.08	54	0.01	0.2
participant 1	458	126	332	0.22	0.24	0.24	0.23	0.04	0	168	0.01	0.4
participant 7	464	113	351	0.18	0.17	0.17	0.17	0.38	0.13	236	0	0.4

Figure 1 : Résultats de l'évaluation tâche générale pour les passages

Identifiant du run	Passages corrects											
	Nb Définitions [33]		Nb Factuelles [400]							oui non [31]	Total	
	org [19]	pers [14]	lieu [65]	man [26]	mes [77]	org [28]	autre/ objet [67]	pers [95]	date [42]		Total [464]	%
participant 5	16	14	52	20	65	25	57	71	36	22	378	81.46
participant 4	14	13	38	9	32	17	40	41	23	10	237	51.07
participant 2	8	11	34	4	32	20	19	46	28	8	210	45.25
participant 6	10	7	32	0	31	9	7	49	25	12	182	39.22
participant 3	9	10	30	6	26	11	22	38	23	11	186	40.08
participant 1	6	6	20	2	20	7	12	35	14	4	126	27.15
participant 7	10	0	14	8	12	8	33	13	3	12	113	24.35

Figure 2 : Résultats de l'évaluation tâche générale pour les passages selon le type de réponse attendu

lieu = lieu, localisation, mes = mesure, org = organisation, pers = personne. man = manière, autre/objet = objet ou autre, date = date, temps

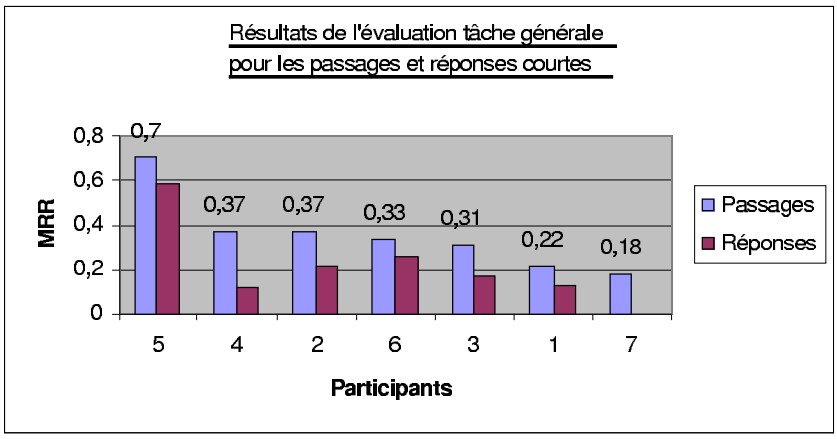
Identifiant du run	Nb de questions répondues [464]	Nb passages corrects	MRR sur tout les types sauf Listes	MRR sur Factuelles et Définitions	MRR sur Factuelles	MRR sur Définitions	MRR sur Oui/ Non	NIAP sur Listes
participant 5	464	312	0.58	0.58	0.57	0.69	0.67	0.71
participant 6	388	139	0.25	0.24	0.24	0.27	0.38	0.02
participant 2	463	131	0.22	0.22	0.24	0	0.25	0.02
participant 3	464	106	0.17	0.16	0.16	0.13	0.35	0
participant 1	333	80	0.13	0.15	0.16	0.01	0.04	0
participant 4	195	76	0.12	0.13	0.09	0.58	0	0

Figure 3 : Résultats de l'évaluation tâche générale pour les réponses courtes

Identifiant du run	Réponses courtes correctes											
	Nb Définitions [33]		Nb Factuelles [400]							oui non [31]	Total	
	org [19]	pers [14]	lieu [65]	man [26]	mes [77]	org [28]	autre/ objet [67]	pers [95]	date [42]		Total [464]	%
participant 5	14	14	46	5	63	19	28	67	34	22	312	67.24
participant 6	8	1	24	0	21	5	5	46	17	12	139	29.95
participant 2	0	0	24	1	24	9	4	39	22	8	131	28.23
participant 3	3	4	15	0	19	9	7	26	12	11	106	22.84
participant 1	1	0	15	1	14	4	3	27	11	4	80	17.24
participant 4	13	11	14	0	10	0	2	18	8	0	76	16.37

Figure 4 : Résultats de l'évaluation tâche générale pour les réponses courtes selon le type de réponse attendu

Pour une plus grande visibilité des résultats, nous avons fourni aux participants les résultats de la tâche générale sous forme de graphe.



4.2 Tâche spécialisée

Cinq groupes ont participé à la tâche spécialisée dans le domaine médical. Trois laboratoires publics : l'Université de Neuchâtel, le CEA-LIST/LIC2M et AP/HP en collaboration avec Paris XIII. Ainsi que deux institutions privées : France Télécom R&D et Synapse Développement.

Les travaux de la plupart de ces systèmes sont présentés plus en détail dans les pages suivantes des actes de l'atelier.

Au total, sept fichier-résultats ont été évalués. Un juge spécialiste de l'équipe du CISMeF (Catalogue et Index des Sites Médicaux Francophones) du CHU de Rouen a évalué les résultats. Les scores ont été calculés sur la base de 200 questions réparties comme suit : 81 « factuelles », 70 « définitions », 24 « oui/non » et 25 « listes ».

Les trois systèmes de question-réponse ayant obtenu les meilleurs résultats pour la tâche spécialisée lors de la campagne EQueR/EVALDA 2004 sont : pour les passages, les systèmes de Synapse Développement (participant 4), de l'Université de Neuchâtel (participant 2), et *ex-aequo* les systèmes de AP/HP-Paris XIII (participant 3) et de France Télécom R&D

(participant 1) ; pour les réponses courtes : le système de Synapse Développement, et *ex-aequo* les systèmes de AP/HP-Paris XIII et de l'Université de Neuchâtel.

Les résultats ont été fournis aux participants sous la forme d'un seul tableau, respectivement pour les réponses courtes et les passages. Il présente pour chaque fichier-résultats le nombre de questions traitées, le nombre de passages (ou réponses) corrects renvoyés, ainsi que les scores obtenus pour chaque type de question et combinaison.

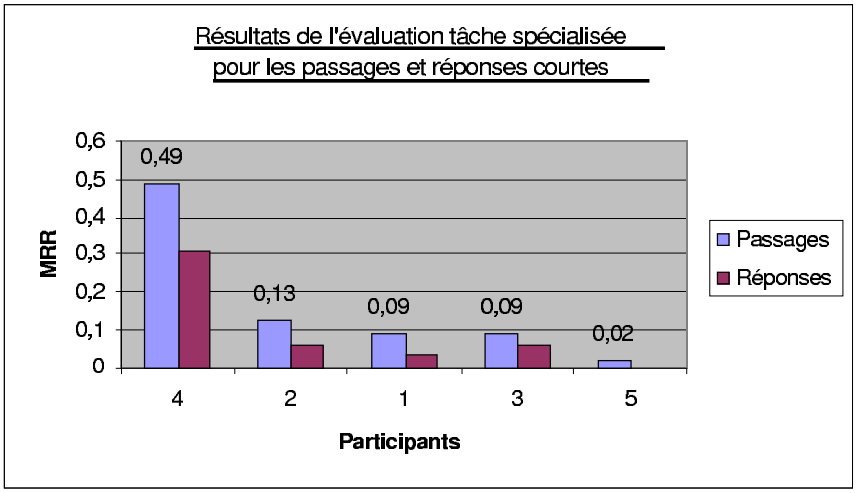
Identifiant du run	Nb de questions répondues [175]	Nb passages corrects	Nb passages incorrects	% passages corrects	MRR sur tout les types sauf Listes	MRR sur Factuelles et Définitions	MRR sur Factuelles	MRR sur Définitions	MRR sur Oui/ Non	NIAP sur Listes
participant 4	175	110	65	62.85	0.49	0.51	0.42	0.62	0.37	0.02
participant 2	175	27	148	15.42	0.13	0.13	0.23	0.02	0.08	0.02
participant 1	166	23	143	13.14	0.09	0.09	0.11	0.07	0.04	0
participant 3	112	16	96	9.14	0.09	0.05	0.02	0.08	0.33	0.01
participant 5	175	7	168	4	0.02	0.02	0.04	0	0	0

Figure 5 : Résultats de l'évaluation tâche spécialisée pour les passages

Identifiant du run	Nb de questions répondues [175]	Nb passages corrects	Nb passages incorrects	% passages corrects	MRR sur tout les types sauf Listes	MRR sur Factuelles et Définitions	MRR sur Factuelles	MRR sur Définitions	MRR sur Oui/ Non	NIAP sur Listes
participant 4	175	110	65	62.85	0.49	0.51	0.42	0.62	0.37	0.02
participant 2	175	27	148	15.42	0.13	0.13	0.23	0.02	0.08	0.02
participant 3	166	23	143	13.14	0.09	0.09	0.11	0.07	0.04	0
participant 1	112	16	96	9.14	0.09	0.05	0.02	0.08	0.33	0.01

Figure 6 : Résultats de l'évaluation tâche spécialisée pour les réponses courtes

Pour une plus grande visibilité des résultats, nous avons fourni aux participants les résultats de la tâche médicale sous forme de graphe :



4.3 Analyse des résultats

En premier lieu, nous avons pu constater de meilleurs résultats pour la tâche générale que pour la tâche spécialisée : les meilleurs scores des systèmes pour la tâche générale s'échelonnent entre 0,7 et 0,18 (selon la métrique adoptée : MRR, Moyenne des Réciproques du Rang, cf. paragraphe 3.2) alors que pour la tâche médicale les résultats s'échelonnent entre 0,49 et 0,02 (toujours d'après la métrique adoptée, MRR, cf. paragraphe 3.2). Ceci s'explique peut-être par la spécificité des textes liés au domaine médical contenu dans cette tâche. Mais ceci s'explique peut-être aussi en raison du délai de livraison du corpus médical (le corpus pour la tâche médicale n'a été distribué que quelques semaines avant l'évaluation).

De plus, l'ensemble des systèmes ont obtenu un meilleur score lors de l'évaluation des passages que lors de l'évaluation des réponses courtes. En effet, il paraît plus difficile pour un système d'extraire une réponse courte exacte et précise qu'un passage un peu plus long dans lequel finalement il est plus probable que se trouve la réponse attendue.

Si l'on compare l'ensemble des systèmes participants on s'aperçoit qu'ils allient tous plus ou moins massivement des technologies de Traitement Automatique des Langues. Pourtant, au vu des résultats, un système obtient des résultats nettement supérieurs aux autres participants, et ce, pour les deux tâches, générale et spécialisée. L'équipe de Synapse Développement présentant ses travaux dans les actes de la conférence principale TALN, on pourra y trouver des éléments d'explication.

Concernant la tâche générale, nous avons trouvé intéressant de faire connaître aux participants les résultats en fonction du type de réponse attendu. Ainsi, ils ont pu se rendre compte, sur quel type de question et de réponse, leur système avait été le plus performant lors de l'évaluation. Tous systèmes confondus, lors de l'évaluation des passages, les meilleurs résultats obtenus concernent les questions de type « définition », puis les questions de type « factuel » simple, les questions de type « oui/non » et enfin les questions de type « liste » pour lesquelles les systèmes ont rencontré le plus de difficultés. Concernant spécifiquement les questions de type « définition », les systèmes ont obtenu de meilleurs résultats lorsque la réponse attendue était une organisation plutôt qu'une personne. Concernant les questions de type « factuel » simple, les systèmes ont obtenu de meilleurs résultats lorsque la réponse attendue était de type « lieu », « organisation », « personne » ou « date » plutôt que « manière », « mesure » ou « objet ».

Pour la tâche générale, lors de l'évaluation des passages, le meilleur système a obtenu 81,46 % de bonnes réponses contre 51,07 % pour le deuxième système. Lors de l'évaluation des réponses courtes, la moyenne baisse avec 67,24 % de bonnes réponses pour le meilleur système et seulement 29,95 % pour le deuxième.

Pour la tâche spécialisée, les résultats baissent encore. Le meilleur système, lors de l'évalution des passages, a obtenu 62,85 % de bonnes réponses contre 15,42 % pour le deuxième système. Et lors de l'évaluation des réponses courtes, le meilleur système obtient seulement 40,57 % de bonnes réponses contre 7,42 % pour le deuxième.

Nous constatons bien une frontière entre les résultats du premier système et ceux des autres systèmes.

5 Conclusion et perspectives

En conclusion, cet article a décrit les principaux aspects de la première campagne d'évaluation de systèmes de question-réponse en France : EQueR.

Cette campagne a été un véritable succès avec la participation et l'intérêt croissant d'une très large majorité des acteurs académiques et industriels du domaine (au total, 7 participants français et 1 participant suisse). Certains participants n'avaient jamais fait d'évaluation question-réponse auparavant et jamais autant de groupes français n'avaient participé à une évaluation question-réponse de la sorte.

Concernant le domaine de l'évaluation, EQueR a innové avec un nouveau type de question, les questions de type « oui/non », qui ont suscitées beaucoup d'intérêt de la part des participants. EQueR a gagné aussi en proposant une tâche question-réponse dans un domaine spécialisé, ce qui a permis d'attirer d'autres participants intéressés plus particulièrement par le domaine médical.

Enfin, EQueR s'européanise avec la campagne d'évaluation CLEF[3] qui, depuis l'année dernière, offre une tâche spécialisée pour l'évaluation des systèmes de question-réponse en Europe. ELDA joue le rôle de coordinateur pour le français dans la campagne européenne CLEF ainsi que celui de distributeur pour l'ensemble des ressources européennes. Au vu des résultats de la campagne EQueR, nous pouvons constater que pour les meilleurs systèmes, les résultats sont comparables avec les résultats des meilleurs systèmes de la campagne CLEF 2004. Concernant la campagne CLEF 2005 qui débutera très prochainement, notre expérience de par la campagne EQueR a été très enrichissante aussi bien pour constituer les corpus de questions que pour discuter de la façon dont seront compilées les données, etc.

Notre souhait est de pouvoir voir en la campagne européenne CLEF l'avenir d'une campagne très enrichissante comme EQueR en France.

Références

AYACHE C. (2005), Rapport final de la campagne EVALDA/EQueR, Evaluation en Question-Réponse, http://www.technolangue.net/article61.html.

VALIN A., MAGNINI B., et AL.(2004), Overview of the CLEF 2004 Multilingual Question Answering Track, Actes de *Cross Language Evaluation Forum.*

VOORHEES E., HARMAN D., (1999), Overview of the Eight Text REtrieval Conference (TREC8), *National Institute of Standards and Technology*, page 1.

[3] CLEF, Cross Language Evaluation Forum, www.clef-campaign.org

SQuAr : Prototype de Moteur de Questions Réponses

Eric Blaudez (1,2), Eric Crestan (1,2) et Claude de Loupy (1,3)

(1) Sinequa Labs,
51-54, rue Ledru-Rollin,
92400 Ivry-sur-Seine, France
{*blaudez*, *loupy*, *crestan*}@sinequa.com

(2) Laboratoire Informatique d'Avignon,
B.P. 1228 Agroparc, 339 Chemin des Meinajaries,
84911 Avignon Cedex 9, France Adresse2

(3) Laboratoire MoDyCo - UMR 7114, Université de Paris 10,
Bâtiment L, 200, avenue de la République
92001 Nanterre Cedex, France

Mots-clés : Moteur de question réponse, Évaluation
Keywords: Question Answering, Evaluation

Résumé : Le système SQuAr est conçu autour de trois modules. Un module est chargé de l'analyse fine de la question. Le second retrouve les documents contenant potentiellement des réponses grâce au moteur de recherche *Intuition*. La dernière étape consiste à extraire les passages contenant une réponse correcte par un calcul de distance d'édition entre question/reformulations et les passages. De plus, les reformulations des questions en forme affirmative servent de patron d'extraction sur le corpus EQueR.

1 Introduction

Depuis le lancement de la première campagne d'évaluation des moteurs de question-réponse (*MQR*) en 1999 dans le cadre des campagnes TREC (Voorhees, Harman, 1999), l'engouement de la communauté pour cette tâche n'a cessé de croître. Au-delà du mode conventionnel de recherche documentaire, les MQR offrent un cadre complet faisant potentiellement intervenir tous les *corps de métier* du TAL. Alors que le nombre de participants aux campagnes TREC-QA est en nette régression depuis quelques années, la première campagne d'évaluation des MQR en français, EQueR (Évaluation en Questions-Réponses), a été menée dans le cadre du programme Evalda[1].

[1] Le projet EVALDA est financé par le Ministère français en charge de la Recherche, dans le cadre du programme Technolangue (http://www.technolangue.net/article20.html)

Le système SQuAr[2] est basé sur une approche en trois phases. Lors de la première phase, les questions sont analysées via des règles linguistiques pour déterminer le type de question et pour repérer les parties essentielles de celles-ci. Les règles sont formalisées sous la forme d'une cascade de 340 transducteurs pour 166 types de question identifiée. De ces types de question découlent des types de réponse attendue (généralement les entités nommées). L'analyse de la question permet également de générer la requête qui sera utilisée pour retrouver des documents susceptibles de contenir une réponse valide. De plus, pour les questions ayant eu une analyse complète, des reformulations sont générées. Le principe est de générer à partir d'une structure interrogative, des patrons d'extraction sous forme affirmative. Ces patrons ont été créés manuellement pour chaque type de question. La seconde phase permet d'extraire des documents de la base EQueR à partir des requêtes générées lors de la phase d'analyse de la question. Les requêtes comportent différents niveaux de contrainte suivant les éléments reconnus dans la question. Seul les 5 premiers documents retournés par le moteur ont été pris en compte à cause de la longueur du traitement d'extraction. La troisième et dernière phase consiste à extraire les passages répondant aux questions. Pour cela, une extraction d'entités est réalisée à l'aide d'une chaîne de transducteurs. Puis, les passages pertinents sont sélectionnés sur un critère d'une distance entre le passage et les reformulations proposées. La distance de *Levenshtein* (Wu, Manber, 1992) a été adaptée au besoin d'une recherche de réponse, en ajoutant une notion de distance entre entités. Les passages obtenant les meilleurs scores sont retournés en premier lieu. Les réponses courtes sont quant à elles extraites en sélectionnant la première entité qui correspond au type attendu. Cependant, quelques problèmes techniques n'ont pas permis d'extraire correctement ces dernières.

2 Système SQuAr

2.1 Analyse de la question

La phase la plus importante pour un MQR est l'analyse de la question. Cela est primordial dans le sens où une mauvaise analyse a de grandes chances d'engendrer des réponses incorrectes. Son rôle principal consiste à déterminer le type de la question et donc le type de la réponse attendue. Cette étape, bien qu'indispensable pour une extraction de réponse exacte, l'est moins pour une extraction par passage. Toutefois, la contrainte de la présence d'une entité candidate dans le passage est un indice fort.

Afin de déterminer le type de la question et d'identifier les éléments clés de celle-ci pour former la requête, des règles linguistiques ont été développées sous la forme d'une cascade de 340 transducteurs. 166 types de questions sont aussi distingués. Différents niveaux d'analyse résultent de cette première phase :

- Analyse complète : La question a été totalement analysée. Le type de question (donc le type de réponse attendue) a été déterminé et les éléments ont été extraits de la question pour former une requête ;
- Analyse incomplète : La question a été partiellement analysée, seul le type de la question a été déterminé avec une certaine latitude dans le type de réponse

[2] Sinequa's QUestions-AnsweRing system.

attendue. La requête sera formée directement à partir de la question après nettoyage des mots outils ;

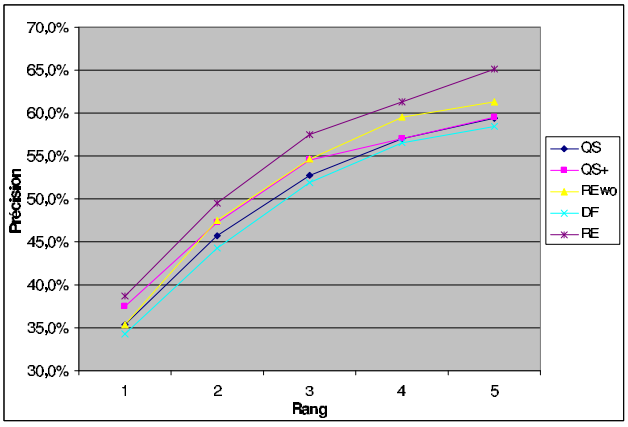
- Échec d'analyse : La question n'a pu être analysée. La requête sera là aussi formée en utilisant les termes présents dans la question.

La sortie de cette analyse se présente sous un format xml, intégrant à la fois les éléments de syntaxe et les éléments d'analyse. Celle-ci est utilisée afin de générer les requêtes pour le moteur de recherche d'un côté et les reformulations pour interroger directement la base et extraire les réponses de l'autre.

Les reformulations correspondent en fait à une transformation des questions à l'affirmatif. Cette transformation permet d'obtenir des séquences plus proches des passages répondant à la question, en terme de structure syntaxique. Par exemple, la question « *Qui est le président de la Zambie ?* » sera reformulée sous la forme des patrons suivants : « *président de la Zambie, {NPP}* » ou encore « *{NPP}, qui est le président de la Zambie* ».
L'élément entre '{…}' correspond au type de l'entité recherchée, dans le cas présent, une personne. La position de cette entité est également importante car elle définit une dépendance syntaxique de la réponse avec son contexte.

2.2 Sélection des documents candidats

Il serait inconcevable de traiter l'ensemble du corpus pour trouver la réponse à chaque question. Pour palier ce problème, nous avons pris le pari de n'analyser que les 5 premiers documents retournés par le moteur. Toutefois, les requêtes précises issus de l'analyse de la question, combinées avec le moteur de recherche sémantique de Sinequa, *Intuition*, permettent de maximiser les chances d'avoir un « *document pertinent*[3] » en tête de liste. Les requêtes générées à partir de l'analyse sont plus ou moins contraintes suivant la complétude de l'analyse. Plus l'analyse est « parfaite », plus la requête résultante sera contrainte (recherche avec présence stricte d'un mot, d'une séquence, ajout de mots supports pour la question…).



L'évaluation de la recherche de documents effectuée après la campagne EQueR montre, en moyenne, que l'utilisation des requêtes évoluées (*RE*, issues de l'analyse

[3] Nous nous référons, par *document pertinent*, aux documents contenant une réponse valide à la question considérée.

des questions) permet d'obtenir plus de document pertinent parmi les 5 premiers retournés. A 5 documents, notre approche utilisant le moteur Intuition permet de gagner près de 12% de précision en plus.

2.3 Extraction des réponses

Une fois les documents « pertinents » trouvés et les entités de chacun de ces documents détectées, la sélection des passages répondant à la question est effectuée. Un score est calculé pour chaque passage de 250 caractères de chaque document. Ils sont ensuite triés suivant ce score et les *N* meilleurs passages sont présentés comme réponses. Afin d'apparier les questions ou reformulation avec les passages, la distance de Levenshtein (Wu, Manber, 1992) a été employée. Cependant, celle-ci a été modifiée pour pourvoir calculer la distance par rapport aux mots et non au caractère comme communément utilisée. De plus, un thésaurus et un dictionnaire de synonyme permettent de prendre en compte les proximités sémantiques ou de domaine entre un mot de la question et un mot lié des passages.

Un second mode d'extraction de réponse possible a été mis au point à travers l'extraction par les reformulations. Elles sont utilisées comme des patrons d'extraction et appliqué directement sur les 100 premiers documents retournés par le moteur de recherche. Le poids de ces réponses est prépondérant sur celles extraites par le calcul de distance.

3 Évaluation

Lors de la campagne EQueR, les questions de type liste n'ont pas été traitées avec SQuAr. De plus, les efforts ont été menés seulement sur l'extraction des réponses sous forme de passages de 250 caractères, au détriment des réponses courtes. Cela s'explique par un maque de temps concédé pour ce développement.

	Définition	Factuelle	Booléenne	*Total*
Nb. Question	33	400	31	**464**
Nb. Correct	19	225	14	**237**

Ces résultats nous ont permis de terminer deuxième sur les 7 participants pour l'évaluation des passages avec 237 réponses correctes sur 464 questions, avec moyenne du rang inverse pour le système de 0,37.

4 Bibliographie

LEVENSHTEIN V. I. (1965), *Binary codes capable of correcting deletions, insertions and reversals.* Doklady Akademii Nauk SSSR 163(4) p845-848.

MANIGOT L., PELLETIER B. (1997), *Intuition, une approche mathématique et sémantique du traitement d'informations textuelles*. Actes de Fractal'1997. pp. 287-291.

VOORHEES E., HARMAN D. (1999), *Overview of the Eighth Text REtrieval Conference (TREC-8)*, National Institute of Standards and Technology, pp. 1.

WU S. AND MANBER U. (1992), *Agrep: a fast approximate pattern-matching tool.* Actes de USENIX Technical Conference, pp. 153-162.

Minimalisme et question-réponse : le système Œdipe

Antonio Balvet, Mehdi Embarek et Olivier Ferret

CEA – LIST/LIC2M
92265 Fontenay-aux-Roses Cedex
{balveta, embarekm, ferreto}@zoe.cea.fr

Résumé - Abstract

Cet article présente le système Œdipe, développé par le LIC2M pour sa participation à la campagne d'évaluation EQUER. Ce système se caractérise par une approche minimaliste au niveau des moyens utilisés. L'article détaille en particulier la stratégie adoptée pour la constitution de patrons d'analyse des questions et en évalue l'efficacité pour EQUER.

This article presents the Œdipe system, which was developed by the LIC2M for participating to the EQUER evaluation campaign and is characterized as relying on minimum means. This article gives more particularly details about the strategy we used for building the patterns for finding the type of questions and evaluates its efficiency for the EQUER questions.

1 Introduction

La campagne EQUER d'évaluation des systèmes de question-réponse en français a été l'occasion pour nous de développer un premier système de question/réponse, le système Œdipe, en s'appuyant sur les outils d'analyse linguistique qui avaient été développés au LIC2M durant les deux années précédentes. Ce développement s'inscrivait donc comme un point de départ et nos ambitions, tant en termes de degré de sophistication du système conçu qu'en termes de résultats attendus, étaient donc modestes. Dans sa forme actuelle, Œdipe peut ainsi être considéré comme un système de question/réponse minimaliste, fournissant en tant que réponse des passages de taille fixe à partir d'un ensemble de documents sélectionnés par un moteur de recherche.

2 Description du système Œdipe

L'architecture du système Œdipe est tout à fait classique. Elle s'appuie à la base sur l'analyseur linguistique LIMA (LIC2M Multilingual Analyzer) (Besançon et al., 2004) qui permet d'une part, de normaliser les mots apparaissant dans les documents et dans les questions et d'autre part, d'en extraire des entités nommées de type MUC (personnes, lieux, organisations, dates et unités de mesure ainsi que produits). La normalisation des mots est réalisée par une analyse morphologique et un étiquetage morpho-syntaxique. La reconnaissance des entités nommées est effectuée quant à elle par une série d'automates appliqués au résultat de l'analyse linguistique. Chaque question posée est analysée afin de déterminer si la réponse attendue est une entité nommée et le cas échéant, le type d'entité

nommée concerné. Cette analyse repose sur un ensemble d'environ 150 automates du même type que ceux définis pour la reconnaissance des entités nommées. Les mots pleins de la question sont par ailleurs pondérés afin de caractériser leur importance *a priori*.

Les documents font quant à eux l'objet d'un traitement en deux temps. Un premier traitement permet de localiser les extraits en relation directe avec la question en s'appuyant sur les zones de forte densité en mots de celle-ci. Chaque extrait se voit attribuer un poids en fonction des mots de la question qu'il contient et éventuellement, de la présence d'entités nommées correspondant au type attendu de réponse. Les extraits sont ensuite ordonnés suivant leur score et les N (égal à 20 pour EQUER) premiers sont sélectionnés. Le second module est chargé de localiser la réponse à la question dans les extraits retenus. Si la réponse attendue est une entité nommée, Œdipe retient comme réponse possible la partie de l'extrait considéré centrée sur une entité nommée du type attendu et qui présente le score le plus élevé, score comparable à celui calculé pour les extraits. Si la réponse attendue n'est pas une entité nommée, Œdipe applique une fenêtre glissante (égale à la taille souhaitée de la réponse) sur l'extrait en calculant pour chacune de ses positions un score comparable à celui de l'extrait. Il retient ensuite comme réponse possible la partie de l'extrait dans laquelle ce score est maximal. Un ensemble de réponses possibles dotées chacune d'un score est ainsi constitué. La liste finale des réponses est obtenue en ordonnant cet ensemble et en le tronquant en fonction du nombre de réponses désiré. Si le score de la meilleure réponse est trop faible, Œdipe suppose qu'aucune réponse n'existe dans les documents. Cette même heuristique est exploitée pour le traitement des questions polaires : une réponse négative est donnée lorsque le score de la meilleure réponse est trop faible. Œdipe prend également en compte les questions de type liste, sans autre particularité que de rechercher le nombre possible de réponses attendues lors du traitement de la question en prenant comme référence la première entité nommée numérique trouvée.

3 Une stratégie de découverte de patrons de questions

Une des parties importantes d'un système de question-réponse est l'analyse des questions. Dans le cas d'Œdipe, cette analyse repose sur l'application d'un ensemble de patrons morpho-syntaxiques. La stratégie adoptée pour leur découverte s'inspire de stratégies d'extraction de patrons couramment employées en extraction d'information (Riloff, 1994) et de travaux dans le domaine de l'apprentissage dit « par alignement » (Van Zaanen, 2001). La justification linguistique de cette approche est à chercher du côté des opérations de segmentation-commutation employées en linguistique structurale pour l'étude des régularités linguistiques en corpus, ainsi que du côté des stratégies d'analyse « naïve » mises en œuvre par les locuteurs d'une langue et qualifiées par Saussure de « fausse analogie ». Concrètement, cette stratégie de co-analyse suit la procédure schématisée ci-dessous :

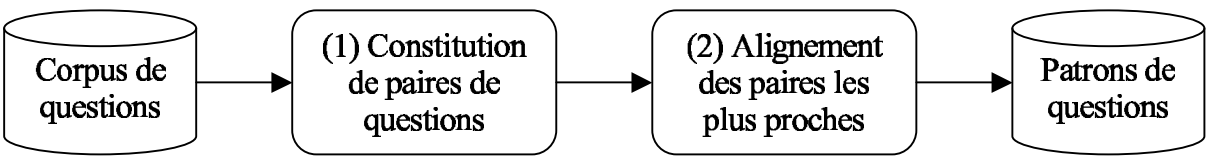


Figure 1 : Étapes pour la constitution d'une base de données de patrons de questions

L'étape (1) se base sur la mesure de la distance d'édition entre deux chaînes de caractères, dite distance de Levenstein, calculée à partir d'opérations d'insertion, d'élision et de

déplacement. Elle aboutit à une liste de paires de questions associées à un score de distance d'édition. L'étape (2) cherche, pour toutes les paires, la plus longue sous-chaîne commune de mots en s'inspirant de l'algorithme Longest Common Substring. Le résultat de ces deux étapes est une liste de paires de questions, chaque paire étant caractérisée par un score de distance d'édition, ainsi que des scores dérivés de celui-ci, et par la plus longue sous-chaîne de mots commune aux deux questions de la paire. Par exemple, la recherche de la plus longue sous-chaîne commune de mots pour la paire de questions ci-dessous donne le patron suivant, où les '_' marquent des positions possibles dans la séquence de mots analysée[1] :

Quelle est la capitale de la Yougoslavie ?
Quelle est la capitale de Madagascar ? → *Quelle est la capitale de _ _ ?*

Le patron extrait est ensuite traduit sous la forme d'une expression régulière typée qui est utilisée lors de l'identification du type d'une question. Par ailleurs, les mots non alignés peuvent être considérés comme les membres d'un même paradigme (*i.e.* des noms de pays pour *Yougoslavie* et *Madagascar*).

L'approche adoptée ici, concrétisée par la plate-forme CoPT[2], est donc une approche de surface, ne mettant en œuvre aucune connaissance linguistique explicite (*i.e.* morphologique, syntaxique ou sémantique) autre que des récurrences de chaînes de caractères et des coïncidences de position pour ces chaînes de caractères. De ce fait, elle est susceptible d'être appliquée sur tout type de corpus de spécialité, y compris dans un contexte multilingue : la stratégie adoptée est en effet indépendante du jeu de caractères employé. Elle requiert simplement une certaine stabilité dans les patrons morpho-syntaxiques employés[3], conformément aux présupposés de l'approche structuraliste. Il reste cependant à améliorer l'algorithme d'appariement utilisé car certains appariements sont impossibles dans la version décrite ici : pour les chaînes « A B C » et « A X B X C », où « X » représente une insertion quelconque (un mot), le patron extrait est « A _ _ _ _ ». Autrement dit, les insertions en position centrale sont mal détectées. Ce problème est néanmoins en voie de résolution.

4 Résultats et analyse

Dans le cadre d'EQUER, le LIC2M a participé à la fois à la tâche générale et à la tâche médicale, en utilisant dans les deux cas exactement le même système et en traitant tous les types de questions (le lecteur pourra se reporter à (Ayache, 2005) pour plus de détails concernant la description des tâches et des métriques utilisées). En revanche, seules des réponses prenant la forme de passages de 250 caractères ont été renvoyées. La Table 1 synthétise les résultats du système Œdipe. Pour la tâche générale, la MRR globale (moyenne de l'inverse des rangs de la première bonne réponse) se situe à 0,7 pour le meilleur système et aux alentours de 0,3 pour la majorité des systèmes, à comparer à 0,5 et 0,1 pour la tâche médicale. Comparativement, les résultats d'Œdipe sont donc faibles et même très faibles pour le domaine médical. Ceci peut s'expliquer bien sûr par le minimaliste d'Œdipe mais la Table 1 montre également un nombre anormalement élevé de questions jugées sans réponse dans le corpus d'évaluation par le système (ce qui explique d'ailleurs le relatif bon score obtenu pour les questions polaires). Il est donc probable qu'une part de nos faibles résultats

[1] La plus longue sous-chaîne commune est alignée sur la plus longue des séquences traitées.
[2] Corpus Processing Tools, disponible à l'adresse http://copt.sourceforge.net
[3] Nous avons pu constater que cette stabilité est plus rare dans des corpus littéraires par exemple.

Tâche	Passages corrects / # questions	MRR Sauf listes	MRR Sauf listes et polaires	MRR Polaires	Précision moyenne Listes	Détection d'absence de réponse (#)
générale	113 / 464	0,18	0,17	0,38	0,13	236 précision : 0 rappel : 0,4
médicale	7 / 175	0,02	0,02	0	0	n/a

Table 1 : Résultats du système Œdipe pour l'évaluation EQUER

puisse s'expliquer par des valeurs de seuil inadéquates à ce niveau. Afin d'éclaircir l'origine des insuffisances d'Œdipe, nous avons mené une analyse manuelle concernant les performances du typage des questions, analyse dont les résultats sont donnés dans la Table 2. Celle-ci laisse apparaître que le typage des questions effectué par le système Œdipe se révèle assez efficace. Si l'on prend en compte à la fois les cas dans lesquels Œdipe trouve le type d'entité nommée attendue comme réponse (première ligne) et les cas dans lesquels il considère que la réponse n'est pas une entité nommée (seconde ligne), on constate qu'il se trompe dans 28,6% des cas pour les 469 questions analysées (hors questions polaires) du domaine général et dans 10,8% des cas pour les 176 questions analysées du domaine médical.

Typage	Jugement manuel	Général	Médical
Type identifié par Œdipe	correct	215 / 254 (45,8%)	17/ 29 (9,7%)
	incorrect	39 / 254 (8,3%)	12 / 29 (6,8%)
Type non identifié par Œdipe ≡ réponse non factuelle	incorrect	95 / 215 (20,3%)	7 / 147 (4%)
	correct	120 / 215 (25,6%)	140 / 147(79,5%)

Table 2 : Résultats de l'analyse manuelle du typage des questions par Œdipe

Un travail d'analyse plus poussé reste donc à mener pour déterminer à quel point le minimalisme du système Œdipe est à l'origine de ses faibles performances et contribuer ainsi sur un plan plus général à éclaircir le rapport, pour un système de question-réponse, entre les moyens engagés et les performances à en attendre.

Références

AYACHE C. (2005), *Campagne EVALDA/EQUER : Evaluation en Question-Réponse*, rapport final de la campagne EVALDA/EQUER (http://www.technolangue.net/article61.html).

BESANÇON R., DE CHALENDAR G., FERRET O., FLUHR C., MESNARD O., NAETS H. (2004), Concept-Based Searching and Merging for Multilingual Information Retrieval: First Experiments at CLEF 2003, LNCS 3237, Springer, pp. 174-184.

RILOFF E. (1994), *Information Extraction as a Basis for Portable Text Classification Systems*, Ph.D. de l'Université du Massachusetts Amherst.

VAN ZAANEN M. (2001), *Bootstrapping Structure into Language: Alignment-Based Learning*, Ph.D. de l'Université de Leeds.

Le LIA à EQueR

L. Gillard, P. Bellot, M. El-Bèze

Laboratoire d'Informatique d'Avignon (LIA)
339 ch. des Meinajaries, BP 1228 ; F-84911 Avignon Cedex 9 (France)
{ laurent.gillard, patrice.bellot, marc.elbeze } @ univ-avignon.fr

Résumé Cet article présente un système de Question Réponse pour le français développé dans le cadre de notre participation à la campagne d'évaluation EQueR 2004.

Abstract This paper describe a Question Answering system for French which was developed for our participation to the EQueR 2004 evaluation campaign.

1 Introduction

Un système de Question Réponse (sQR) permet, à partir d'une question exprimée en langue naturelle, d'obtenir automatiquement une réponse concise à cette question. Ainsi, la campagne EQueR est la première campagne d'évaluation sur le français de ces systèmes. Elle se propose d'établir un référentiel permettant leur émulation et leur comparaison, et rejoint en cela les objectifs des campagnes comme TREC-QA (précurseur du domaine, sur l'anglais) ou CLEF-QA (langues européennes).

Dans cet article, nous présentons le sQR que nous avons développé pour EQueR. L'architecture du système correspond à un découpage séquentiel en modules, dont les principaux concernent l'étiquetage des questions, la recherche de segments de documents et l'extraction d'une réponse avec ou sans exploitation de bases de connaissances. Chacun de ces constituants donnera lieu à une description succincte avant la présentation des résultats.

2 Architecture générale du sQR : par composants

2.1 Analyse des questions

Le sQR du LIA, comprend un composant d'étiquetage hiérarchique des questions, capable d'effectuer un appariement entre une question et un ou plusieurs types d'entités réponses attendues (*ERa*). La hiérarchie utilisée a été inspirée par celle proposée par (Sekine *et al.*, 2002), dont elle est un sous ensemble. Ce sous-ensemble a été choisi selon une observation de la fréquence des questions associées à une entrée de cette hiérarchie dans les questions françaises proposées lors des campagnes d'évaluation QR de CLEF. Concrètement, cet étiquetage se déroule après une étape d'uniformisation, à base de règles et de lexiques, permettant de réduire différentes variantes à une même écriture. Cela permet de diminuer le nombre de règles d'étiquetage, mais également d'en faciliter l'écriture (qui est manuelle).

2.2 Pré-traitements : Filtrage du corpus et Reconnaissance des Entités

Le filtrage du corpus consiste à restreindre, à partir de la question ou d'une variante, l'espace de recherche de la collection des documents à un sous ensemble de celle-ci. Pour EQueR, il était proposé comme facilité, et pour chaque question, les 100 premiers documents retournés par le moteur de recherche de la société Pertimm. Le sQR décrit ici n'utilise qu'une partie de ces listes de documents : en effet, seules les collections « Le Monde » et « Le Monde Diplomatique » ont été considérées suite à différents problèmes d'ingénierie.

Avec un objectif similaire de réduction des possibilités, la tâche d'étiquetage des entités permet dans un texte, de marquer certaines de ses parties comme des réponses candidates. Etant donné que notre sQR ne fonctionne que par extraction des entités (pas de processus de synthèse ou de raisonnement), la qualité et la granularité de cet étiquetage est crucial. Il est effectué par deux outils dont la couverture est complémentaire : le premier construit à partir de la plateforme GATE (Cunningham et al., 2002) dans le cadre d'une collaboration avec la société iSmart ; le second étant également à base de transducteurs et de lexiques. Au final, le nombre d'entités nommées ou génériques reconnues est d'environ 70, et entre en correspondance avec un sous ensemble des *ERa* issues de l'analyse des questions. Cette reconnaissance d'entités est effectuée à la fois sur les questions et sur les documents, de même qu'un étiquetage syntaxique obtenu grâce au TreeTagger (Schmid, 1994).

2.3 Recherche de passages

Afin de mieux localiser les entités candidates, notre sQR effectue une recherche de passages dans les documents préalablement filtrés et étiquetés.

Une requête est constituée d'un ensemble « d'objets » provenant de la question : les lemmes des mots, à l'exception des mots outils, les étiquettes d'entités présentes dans la question ainsi que celles des *ERa*. Ensuite, un score normalisé (F 1) est calculé à partir d'une distance moyenne μ (évaluée en nombre de mots) entre une occurrence d'un objet, et les autres objets de la requête (ou de leur plus proche occurrence en cas de présence multiple), cela dans chacun des documents issus du filtrage et pour chacun des objets.

$$score(o_i) = \frac{\log[\mu(o_i) + (\text{tailleRequête} - \text{nbObjetsTrouvés}) * \textit{pénalité}]}{\text{tailleRequête}} \qquad \text{(F 1)}$$

La pénalité est fixée empiriquement afin de plus ou moins favoriser le nombre d'objets communs entre la question et le texte par rapport à la distance moyenne qui les sépare.

Le score d'une phrase correspond au meilleur score des objets qu'elle contient. Pour chaque phrase, un « passage » est constitué à partir de cette dernière mais aussi de la phrase qui la précède et qui la suit, lorsqu'elles existent (cela afin d'essayer de compenser une éventuelle perte d'information sur les phrases courtes ou utilisant des référents trans-phrase). Le score d'un passage est le score de sa phrase centrale. Les meilleurs passages (au plus 1 000) pour l'ensemble des documents trouvés sont proposés en entrée de l'étape suivante.

2.4 Sélection de la Réponse

Enfin, la sélection d'une réponse est l'étape ultime de notre sQR. Pour cela, il dispose des passages ordonnées, suivant le score présenté en (2.3), contenant des entités (2.2), et d'entités réponses attendues (sortie de l'étiquetage 2.1).

Pour certaines questions, il dispose également de l'appui d'un module de base de connaissances (BC), sorte de couples pré-enregistrés de QR utilisés pour crédibiliser une entité candidate. En effet, les réponses contenues dans ces BC sont sous la forme d'expressions régulières et permettent l'extraction d'une réponse dans un passage déjà sélectionné comme un « support susceptible d'être correct ». Ce module avait été construit suite au constat que certaines questions assimilables à des thématiques de culture générale (telles que les capitales géographiques), et aux réponses peu variables, étaient relativement fréquentes dans les campagnes QR. Cette observation nous avait conforté dans le choix de mettre en place un tel composant lors de notre participation à TREC-11 (Bellot et al., 2002). Il a été « francisé » et sa couverture légèrement augmentée : ainsi, d'environ 20% sur le jeu des questions traduites de TREC-11, il permet d'atteindre au mieux 12% de couverture sur EQueR.

Cependant, l'essentiel de la sélection d'une réponse repose sur une autre approche : l'approximation que le candidat optimal peut être extrait à partir d'une « compacité moyenne » des mots. Par compacité moyenne, nous entendons une mesure normalisée du nombre de mots communs entre une question et un passage à l'intérieur d'une fenêtre glissante centrée sur une entité candidate d'un type compatible avec l'une des entités réponses attendues. Bien que sa formulation soit différente, cette mesure est de même nature que celle utilisée pour la sélection des passages.

3 Résultats

Le détail des métriques d'évaluation ainsi que celui des résultats pour tous les participants sont présentés dans le rapport final de l'évaluation (Ayache *et al.*, 2005). Aussi, dans cette section n'est envisagée qu'une partie des résultats obtenus par notre système.

Le tableau 1 présente le nombre de réponses correctes retournées par type (questions définitoires, *« Qu'est-ce que l'Unscom ? » ;* q. factuelles, *« Où se trouve le siège d'Adidas en France ? » ;* et q. booléennes, *« Est-ce que Bernard Dort a rencontré Bertolt Brecht ? »*) à la fois pour les évaluations « passages » et « courtes » pour les 2 soumissions (*« run »*) officielles effectuées lors de notre participation : les différences se situent au niveau du prétraitement c.à.d. dans le nombre de documents pris en compte et provenant de l'étape de filtrage du corpus, au plus 30 pour le *« run1 »* et au plus 100 pour le *« run2 »* ; ainsi que dans la finesse de l'étiquetage des entités, le premier, est basé sur une hiérarchie moins fine que le second. Cependant, ce *« run2 »* était entaché d'un problème de format qui n'a été corrigé qu'après l'échéance d'EQueR, aussi, bien que plus abouti, il a donné des résultats moins satisfaisant qu'attendu. Le *« run2 corrigé »* est également présenté, il a été évalué à l'aide de motifs construits d'après l'ensemble des réponses jugées et soumises par tous les participants.

Il est a noté que l'aspect séquentiel du sQR a également entraîné différents silences au niveau de chacun de ses composants, ce qui s'est traduit par un nombre final de questions répondues d'au mieux 407 réponses sur l'ensemble des 464 questions de l'évaluation (*cf.* avant dernière colonne du tableau 1). Nous excluons volontairement les questions « listes » dans ces décomptes, notre approche embryonnaire de celles-ci étant à améliorer.

En terme de score, pour les réponses « courtes » – par opposition à celles « passages », c.à.d. moins localisées et contenues dans un bloc de 250 caractères – le MRR (Moyenne des réciproques du rang, elle correspond à la moyenne des inverses du rang de la première bonne

réponse parmi les 5 autorisées, ou zéro en cas d'absence d'une réponse correcte) est de 0,25 sur le « *run1* » est de 0,23 sur le « *run2* » (*cf.* dernière colonne du tableau 1).

A titre de comparaison, sur les réponses « courtes », le meilleur système obtient un MRR de 0,58 et répond correctement à 312 sur 464 questions. Notre sQR se positionne en seconde position sur les réponses « courtes », mais en 5ième sur les 7 systèmes participants dans le cas des réponses « passages ».

	# Q Définitoires (33)	**# Q Factuelles (400)**	**# Q Booléennes (31)**	**# réponses CORRECTES**	**soit %**	**# réponses répondues (464)**	**MRR**
PASSAGE – run1	17	153	12	182	39,2	388	0,33
COURTE – run1	9	118	12	139	29,5	388	0,25
PASSAGE– run1	14	137	11	162	34,9	354	0,29
COURTE – run2	8	111	11	130	27,6	354	0,23
PASSAGE – run2Corr.	14	194	11	219	47,2	407	0,39
COURTE - run2Corr.	4	155	11	170	36,6	407	0,29

Tableau 1 : Nombre de réponses correctes par soumission et par type ; MRR global

4 Conclusion et perspectives

Nous avons présenté les grandes lignes d'un système de QR pour le français principalement basé sur des métriques de compacité moyennes pour la sélection des passages et des réponses ; ainsi que les résultats obtenus lors de la campagne EQueR.

En outre, le fait de disposer d'un sQR complet et d'un corpus de référence en français permettra d'autres expériences, l'évaluation précise (en cours) de chacun des composants de ce système, et ainsi de répondre à la question concernant la probable amélioration de ses performances, actuellement plus qu'acceptables si on considère le classement, largement perfectibles si l'on tient compte de la précision.

Références

AYACHE C., CHOUKRI K., GRAU B. (2005). Campagne EVALDA/EQueR Evaluation en Question-Réponse. *http://www.technolangue.net/IMG/pdf/rapport_EQueR_1.2.pdf*

BELLOT P., CRESTAN E., EL-BÈZE M., GILLARD L., DE LOUPY C. (2002). Coupling Named Entity Recognition, Vector-Space Model and Knowledge Bases for TREC-11 Question Answering Track. *in Proceedings of the 11th Text REtrieval Conference.* Gaithersburg, Maryland, USA pp. 398-406.

CUNNINGHAM H., MAYNARD D., BONTCHEVA K., TABLAN V. GATE: (2002). A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02).* Philadelphia, July 2002. *http://gate.ac.uk/sale/acl02/acl-main.pdf*

ISMART, *http://ismart.fr/*.

PERTIMM, *http://www.pertimm.fr*.

SEKINE S., SUDO K., NOBATA C. (2002). Extended Named Entity Hierarchy. *In Proceedings of the LREC-2002 Conference*, pp. 1818–1824.

SCHMID H. (1994). Probablistic Part-of-Speech Tagging Using Decision Trees. *In Proceedings of the First International Conference on New Methods in Natural Language Processing (NemLap-94)*, Manchester, U.K., pp. 44-49.

FRASQUES, le système du groupe LIR, LIMSI

B. Grau (1), G. Illouz (1), L. Monceaux (2), P. Paroubek (1), O. Pons (3), I. Robba (1), A. Vilnat (1)

(1) Groupe LIR – LIMSI
BP 133, 91403 Orsay Cedex
{grau, illouz, pap, robba, vilnat}@limsi.fr
(2) LINA – Université de Nantes
2 rue de la Houssinière, BP 92208, 44322 Nantes Cedex 03
Laura.Monceaux@lina.univ-nantes.fr
(3) CEDRIC-IIE – CNAM
IIE, 18 allee Jean Rostand, 91025 Evry Cedex
pons@cnam.fr

Mots-clés : système de question-réponse, évaluation
Keywords: question-answering system, evaluation

Résumé Le système FRASQUES qui a participé à l'évaluation EQueR est présenté ici en comparaison avec notre système QALC dédié à l'anglais. Ses résultats sont commentés et une évaluation des différents modules est exposée.

Abstract We present our system FRASQUES in comparison to QALC our system for English. The results that FRASQUES obtained at EqueR are presented, and an evaluation of its modules is given.

1 Présentation de FRASQUES

Comme QALC notre système pour l'anglais, (Ferret *et al.* 2002), FRASQUES s'organise en quatre principaux modules présentés dans la figure 1 : l'analyse des questions, la sélection des documents par un moteur de recherche, et le traitement des documents pour en extraire les phrases et les réponses finales. Nous allons tout d'abord présenter globalement notre système en précisant comment s'est faite l'adaptation de QALC au français, puis nous en donnerons ses résultats, et ce pour les différents modules.

L'analyse des questions est réalisée en deux étapes. L'analyseur XIP de Xerox (Aït-Mokthar *et al.* 2002) construit les segments syntaxiques et établit les relations entre eux. A partir de ces

données, des informations telles que le focus ou le type attendu de la réponse sont calculées. Ce module a été réécrit pour traiter les questions en français, mais les règles de reconnaissance ont été transposées depuis l'anglais de manière très directe.

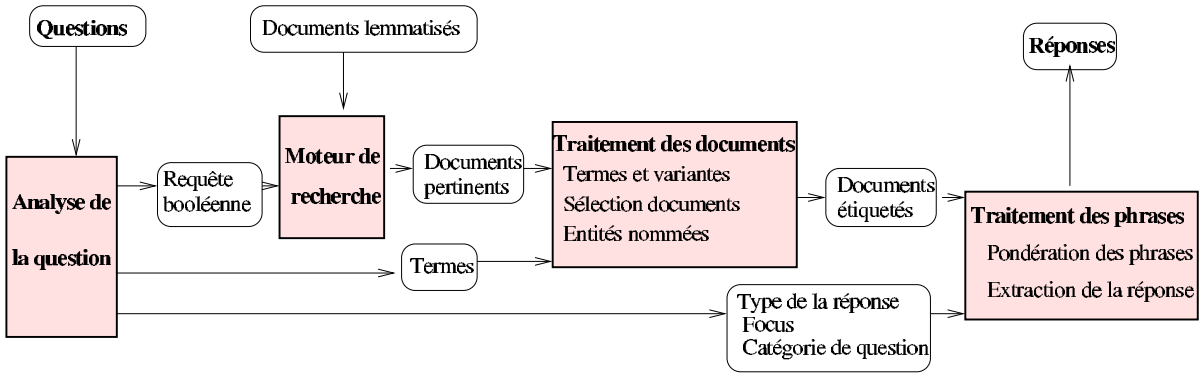


Figure 1 : Le système FRASQUES

Le moteur de recherche utilisé, Lucene[1], est un moteur booléen ; il nous a permis d'indexer le corpus, qui a été au préalable lemmatisé par le Tree Tagger[2] et l'analyseur morphologique de XIP. QALC, pour sa part, faisait appel à un moteur vectoriel utilisant le stemming. Lors de l'interrogation, Lucene reçoit un ensemble de requêtes constituées des mots non vides de la question. Les requêtes les plus larges, n'étant utilisées que si les plus précises ne retournent pas assez de documents. Les documents trouvés par Lucene, sont traités par Fastr[3] qui permet de reconnaître des variantes morphologiques, syntaxiques et sémantiques des termes simples et composés de la question et de pondérer les documents selon les termes trouvés. Les documents sont ainsi réordonnés et un sous-ensemble est extrait sur lequel on applique alors le module d'extraction des entités nommées. Cette partie de la chaîne de traitement est la même pour FRASQUES et QALC, seules diffèrent les ressources qui lui sont données.

Enfin le module d'extraction de la réponse est appliqué ; il procède différemment selon que la question attend ou non pour réponse une entité nommée. Les phrases sont pondérées en fonction du taux de présence des mots de la question dans la phrase, et de la présence ou non de l'entité nommée attendue. Dans FRASQUES, comme dans la plupart des systèmes de QR, les questions à entités nommées obtiennent de meilleurs résultats que les autres, car, de par leur nature, les entités nommées sont plus facilement repérées dans les documents. L'extraction de la réponse exacte a été réalisée différemment dans FRASQUES, puisque nous avons utilisé l'analyseur SCOL[4], afin d'appliquer des patrons d'extraction. Ces patrons, écrits sous forme de règles, s'appuient sur un étiquetage morpho-syntaxique des phrases spécialisant les caractéristiques de la question qui sont présentes dans la phrase sous leur forme initiale ou sous forme de synonyme.

---

[1] http://jakarta.apache.org/lucene/docs/index.html

[2] http://www.uni-stuttgart.de/lingrom/stein/forschung/resource.html

[3] http://www.limsi.fr/Individu/jacquemi/FASTR/

[4] http://www.sfs.nphil.uni-tuebingen.de/Staff-Old/abney

La campagne EQueR proposait des questions booléennes ainsi que des questions dont la réponse devait être une liste de réponses. La stratégie mise en œuvre pour répondre à ces questions était très simple : si la phrase réponse comportait tous les noms de la question et le verbe principal, la réponse était positive. En ce qui concerne les listes, le choix de la réponse finale consistait à reclasser les réponses extraites, en favorisant le plus grand ensemble de réponses dans un même document et en fixant un nombre de réponses proposées en fonction du nombre de réponses demandées dans la question.

Nous allons maintenant donner les résultats de FRASQUES et les évaluations de chacun des modules réalisés. Pour cela, nous avons recensé pour chaque question les différentes réponses données par les participants et l'organisateur, et nous avons testé la présence de ces réponses dans les documents et passages retournés par notre système[5].

2 Sélection des documents

En ce qui concerne la sélection des documents, le moteur de recherche ne retourne de documents contenant la réponse que pour 73 à 76% des questions pour les deux tests soumis. Cela s'explique par plusieurs facteurs : imprécision de la sélection des mots-clefs des questions, qui sont retenus uniquement en fonction de leur étiquette morpho-syntaxique ; erreurs de lemmatisation problèmes de référence... Ces difficultés étaient déjà présentes dans QALC. La différence entre les deux tests tient au fait que pour le deuxième, le nombre de documents retournés était limité à 200.

La sélection par Fastr de 50 documents en fonction des reformulations des multi-termes (et de synonymes mono-termes pour le test 2) trouvés n'entraîne pas de perte de bons documents. Le second test ayant retenu les synonymes mono-termes, on pourrait s'attendre à ce que son rappel soit meilleur, mais il n'en est rien. Ce phénomène peut s'expliquer par le fort degré de ressemblance des questions avec les phrases réponses, et par le bruit introduit par la recherche de « mauvais » synonymes.

	Réponses longues	MRR	MRR2	Réponses courtes	MRR	MRR2	Phrases Rang 1-5	MRR
Test 1	210 (42%)	0,37	0,38	131 (26%)	0,22	0,22	253 (60%)	0,48
Test 2	187 (37%)	0,32	0,33	118 (24%)	00,2	0,2		
Trec9	393 (56%)		0,407					
Trec11				139 Rang 1 (28%)				

Tableau 1 : Résultats officiels de FRASQUES à EqueR et comparaison avec QALC à TREC

[5] Le nombre des questions est ici ramené à 450 : les questions booléennes et les questions dont la réponse est une liste n'ayant pas été prises en compte.

3 Pondération des phrases et extraction de la réponse

Dans le tableau 1, qui donne les résultats officiels, nous avons comparé nos résultats à ceux de QALC. Il est peu surprenant de voir que les systèmes obtiennent des performances similaires. En effet, les résultats officiels à Trec11 (165 bonnes réponses sur 500 questions) provenaient de la fusion de 2 sources de réponses et augmentaient le nombre de réponses correctes (Berthelin *et al.* 2003). En outre, nous avons appliqué les patrons de réponses sur les 5 premières phrases du test 1, et obtenons alors un résultat de 60% de bonnes réponses contre 42% de réponses longues dues à des phrases tronquées par erreur.

Des résultats issus d'EqueR, nous avons mené deux études. La première mesure le taux de présence des termes à l'identique et celui des synonymes, dans nos phrases réponses et dans celles des participants (Grau et al . 2005). On voit que très peu de synonymes sont présents dans les phrases retenues et cela amène à poser la question du type de connaissance à utiliser. La seconde étude (réalisée par A.-L. Ligozat, groupe LIR) vise à déterminer les phénomènes responsables de l'extraction d'une réponse erronée quand on dispose d'une phrase correcte. Aussi, nous avons retenu notre ensemble de réponses longues correctes, pour lesquelles nous avons recherché la phrase d'origine. Il en ressort que parmi 74 questions, 25 réponses incorrectes sont dues à l'application d'un mauvais patron, 33 à un mauvais étiquetage de la réponse attendue : 15 ont une mauvaise étiquette, 11 sont dues à une absence d'étiquette et 7 à un étiquetage présent à tort, 11 EN sont absentes des documents et 5 erreurs diverses.

4 Conclusion

Même s'il est encore perfectible, FRASQUES notre système mis au point dans le cadre d'EqueR, apporte une brique fondamentale à notre système multilingue MUSCAT qui jusqu'à présent ne possédait que peu de modules véritablement multilingues. En dehors des résultats et des comparaisons possibles avec les participants, la finalisation de systèmes est un des apports majeurs d'une campagne d'évaluation telle EQueR.

Références

AÏT-MOKTHAR S., CHANOD J.P, ROUX C., (2002), Robustness beyond shallowness : incremental deep parsing, *Journal of Natural Language Engineering*, Vol. 8, n°3-2.

BERTHELIN J.B., DE CHALENDAR G., ELKATEB-GARA F., FERRET O. GRAU B., HURAUL-PLANTET M., MONCEAUX L., ROBBA I., VILNAT A. (2003), Getting reliable answers by exploiting results from several sources of information, Actes de CoLogNET-ElsNET Symposium (Question and Answers: Theoretical and Applied Perspectives), Amsterdam.

FERRET O., GRAU B., HURAULT-PLANTET M., ILLOUZ G., JACQUEMIN C., MONCEAUX L., ROBBA I., VILNAT A. (2002) How NLP Can Improve Question Answering, *Knowledge Organization*, Vol. 29 (2002), N°3-4, pages 135-155

GRAU B., LIGOZAT A. L., ROBBA I., VILNAT A., ELKATEB-GARA F., ILLOUZ G., MONCEAUX L. PAROUBEK P., PONS O., (2005) De l'importance des synonymes pour la sélection de passages en question-réponse, Actes de CORIA, Grenoble.

Le système STIM/LIPN à EQueR 2004, tâche médicale

Thierry Delbecque[1, 2], Pierre Zweigenbaum[1, 2, 3], Jean-François Berroyer[4, 5], Thierry Poibeau[4, 5]
(1) INSERM, U729, 75006 Paris
(2) INALCO, CRIM, 75343 Paris Cedex 07
(3) Assistance Publique - Hôpitaux de Paris, STIM/DSI, 75674 Paris Cedex 14
(4) CNRS, UMR 7030, 93430 Villetaneuse
(5) Université Paris 13, LIPN, 93430 Villetaneuse

Mots-clefs : Systèmes de questions-réponses, médecine, projet EQueR

Keywords: Question-answering systems, medicine, EQueR project

Résumé Nous présentons les principes du système de questions-réponses STIM/LIPN, fruit d'une collaboration entre le STIM (AP-HP & INSERM U729) et le LIPN (CNRS & Université Paris 13), qui a pris part à la tâche médicale de l'évaluation EQueR 2004 (Ayache, 2005).

Abstract We present the principles of the question-answering system STIM/LIPN, the result of joint work by STIM (AP-HP & INSERM U729) and LIPN (CNRS & Université Paris 13), which participated in the medical track of the EQueR 2004 evaluation (Ayache, 2005).

1 Segmentation et indexation du corpus

1.1 Segmentation

Nous avons fait l'hypothèse que la réponse à une question pouvait être trouvée au sein d'une même phrase. Nous avons donc décomposé le corpus en phrases, plus précisément en unités que nous avons appelées *prédicats*, dont l'organisation comporte (figure 1) (i) une tête verbale ; (ii) un modérateur, pour exprimer les négations, intensité, etc. (iii) un sujet ; (ix) un ou plusieurs compléments. Pour cela, le corpus dans sa totalité à été soumis à TreeTagger[1] ; la détermination et la construction des prédicats se fait sur la base de patrons de parties du discours. Le résultat intermédiaire est un document XML.

Parallèlement à cette analyse, une détection des définitions d'abréviations est effectuée ; nous disposons ainsi d'un dictionnaires d'acronymes, avec pour chaque entrée les définitions possibles, les fréquences et localisations des définitions dans le corpus.

[1] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html`.

```
<form id="57">
        <predicat> être </predicat>
        <moderator> souvent </moderator>
        <subject from="1339" to="1343">
                <SUI>1:S0226498</SUI> <CUI>C0000726</CUI> <TUI>T029</TUI>
                L examen de l abdomen </subject>
        <arguments> <item from="1346" to="1361">
                        <SUI>1:S0234187</SUI> <CUI>C0022828</CUI> <TUI>T007</TUI>
                        normal ( formes basses ) ou montre
                        parfois une sensibilité de la fosse iliaque gauche . </item> </arguments>
</form>
```

FIG. 1 – Exemple de prédicat

1.2 Repérage d'entités nommées générales et médicales

Entités nommées médicales Nous avons choisi, comme source d'EN médicales, d'utiliser les types sémantiques et les relations sémantiques fournies par l'UMLS (`http://www.nlm.nih.gov/research/umls/`). Ainsi, nous espérions faire ressortir aussi bien les prédicats portant par exemple sur une pathologie (type sémantique *Pathologic Function* :T046), que ceux pouvant exprimer un traitement en action (relation sémantique *treats* :T154)[2].

La projection des types sémantiques sur les prédicats nécessite la détermination des termes simples et des termes composés, dans les sujets et compléments des prédicats ; puis la recherche de ces termes, ou de termes hyperonymes, au sein de la partie francophone de l'UMLS[3]. On remonte des termes aux concepts, puis aux types sémantiques correspondants. Les prédicats sont ainsi enrichis avec les types sémantiques dont la présence est soupçonnée, ainsi qu'on le voit sur un extrait (figure 1).

Les autres entités nommées Les autres entités nommées sont reconnues grâce à un outil appelé TagEN, développé au LIPN. L'analyse effectuée est très classique : elle se fonde sur un ensemble de données lexicales définies dans des dictionnaires et sur des automates permettant de regrouper les unités pertinentes en fonction des informations contenues dans les dictionnaires. Les types d'entités classiquement définis ont été repris — noms de personnes, de lieux, dates, durées —, et d'autres ont été affinés — pour les entités chiffrées, on distingue des dosages, des posologies, etc. L'analyse fait alors appel aux algorithmes classiques sur les automates, tels qu'ils sont développés au sein de la boîte à outils Unitex (`http://www-igm.univ-mlv.fr/~unitex/`). En cas d'ambiguïté dans l'analyse, seule la séquence la plus longue est retenue.

1.3 Indexation

Nous gérons directement l'ensemble des prédicats dans une base de données MySQL ; l'indexation remplace le recours à un moteur de recherche. Quatre tables d'index sont créées : sur les têtes verbales (lemmatisées), sur les mots pleins (formes originales et lemmatisées), sur les EN générales ; sur les types et relations sémantiques projetés depuis l'UMLS. Un attribut du schéma relationnel est consacré à la position dans le corpus du prédicat, pour pouvoir extraire a posteriori le passage correspondant.

[2]Le réseau sémantique de l'UMLS contient 134 types sémantiques et 54 relations sémantiques, hiérarchisées par un lien *is-a*.

[3]Ce qui constitue en soi une limitation : le français ne couvre qu'à peine plus de 2% des concepts de l'UMLS (version 2002-AA).

2 Exécution d'un *run*

2.1 Analyse et transformation de la question

L'analyse d'une question doit permettre de décider de sa catégorie (en particulier, si l'on cherche une définition), du type d'entité nommée s'il y a lieu, et de l'ensemble des mots clés rencontrés (« mots d'ancrage »). Lorsqu'une question concerne la définition d'un acronyme, elle est traitée de manière particulière (voir la section 2.2)[4].

Le typage de la question est effectué en appliquant une série d'automates sur les questions, suivant la même stratégie que pour le repérage des entités nommées. L'analyse de la question produit un objet XML.

Dans le cas où la question ne porte pas sur la définition d'un acronyme, l'objet XML est transformé en requête dans un langage intermédiaire. Lors de cette transformation, la requête peut être étendue, en fonction du type d'entité nommée cherché, en particulier s'il s'agit d'une entité médicale. Par exemple si la question porte sur une pathologie, on adjoindra à la requête « *syndrome* » comme mot clé pertinent ; de telles extensions portent également sur la tête verbale. On associe également à la requête les poids à affecter aux réponses, en fonction de la présence ou non d'indices (EN, mots clés, etc.).

Enfin, une analyse complémentaire de la question y recherche des mots clés du thésaurus MeSH. Ces mots clés sont utilisés plus tard (voir la section 2.2) pour obtenir un indice de confiance supplémentaire dans certains documents.

2.2 Requête et tri des réponses

Lorsque l'analyse de la question a établi qu'il s'agit de trouver la définition d'un acronyme, la procédure de recherche consiste simplement à consulter le dictionnaire d'acronymes (1.1).

Dans le cas général, la requête en langage intermédiaire est traduite en SQL, puis soumise à MySQL ; les résultats sont utilisés pour isoler les fragments de texte adéquats dans le corpus, qui constitueront les résultats finals[5]. Ceux-ci sont alors pondérés (notés) selon différents critères.

L'un de ces critères tient compte du fait que la thématique de la question est fortement représentée dans le document où la réponse a été trouvée. Pour cela, nous nous aidons de l'indexation thématique faite par le portail CISMeF (`http://www.chu-rouen.fr/cismef/`) à l'aide de termes MeSH sur certains documents du corpus EQueR médical. Si la question contient des mots clés du thésaurus MeSH, une requête est envoyée au moteur de recherche du catalogue CISMeF. Les documents de CISMeF indexés par les mots clés MeSH de la question, lorsqu'il y en a, sont ainsi recensés. Si une réponse proposée à la question est trouvée dans l'un de ces documents, elle reçoit un bonus : sa note globale est augmentée d'un point.

Les autres critères portent sur la présence ou non d'entités nommées, de verbes ou de mots clés précis[6] ; la note globale est réévaluée en conséquence, conformément à ce qui est spécifié dans la requête intermédiaire. Un score final est affecté à chaque réponse, ce qui permet de les trier.

[4] Après l'évaluation, nous avons aussi mis au point un traitement spécifique pour les questions « définitoires » (Malaisé *et al.*, 2005).

[5] Sauf dans le cas des réponses oui/non, qui réclament un traitement supplémentaire, présenté dans la section 2.3.

[6] Dans le prédicat, et non dans le passage.

2.3 Décisions finales pour les réponses

La compétition prévoyait trois formats de réponses : longues, courtes, et binaire, ce qui a nécessité un reformatage des réponses.

Le mécanisme de construction d'une réponse courte n'était pas encore finalisé lorsque le système a été présenté. Nous avons donc décidé de ne pas demander l'évaluation des réponses courtes, sauf pour les cas suivants qui étaient prêts :
- les réponses booléennes, qui n'auraient pas de sens sinon ;
- les définitions d'acronymes : questions de type définition où le passage trouvé contenait un sigle (au moins deux lettres majuscules consécutives).

Toutes les autres réponses courtes ont été mises à « NUL ».

Pour les questions booléennes, il faut déterminer, sur la base des passages trouvés, si la réponse est oui ou non. Nous utilisons pour cela une heuristique simple. Le système, tel que présenté à l'évaluation, renvoie pour chaque question booléenne au plus un passage, dans lequel il a trouvé un ou plusieurs « mots d'ancrage » (mots de la question). Si au moins trois mots d'ancrage (de plus de trois lettres) ont été trouvés, la réponse fournie est « oui », et « non » dans le cas inverse.

3 Conclusion

Ce travail a été pour nous l'occasion de proposer l'utilisation de ressources terminologiques médicales dans une optique de QR. L'un des enjeux était l'utilisation de l'UMLS comme source d'entités nommées spécifiques au domaine. Sur 200 questions, le système a proposé 112 passages, dont 16 corrects (3^{e} ex æquo, moyenne globale de l'inverse des rangs de 0,09), et 35 réponses courtes, dont 12 correctes (2^{e} ex æquo, MRR = 0,06). Les réponses correctes obtenues portaient sur des questions booléennes (MRR = 0,33) et définitions (en l'occurrence, des sigles).

Le système est encore très jeune ; tous les composants ont été développés pour l'occasion, et il a été assemblé juste à temps pour l'évaluation, sans phase de test préalable. Parmi ses points faibles, outre sans doute des bogues à éliminer, on trouve certainement la qualité insuffisante de la construction des prédicats, qu'une véritable analyse syntaxique devrait améliorer. D'autre part, seule une infime partie des apports possibles de l'indexation par « concepts » de l'UMLS réalisée ici a été exploitée. Enfin, une méthode simple, de type recherche de passages en texte intégral, pourrait sans doute aider à obtenir des réponses lorsque la méthode présentée ici reste silencieuse. L'analyse *post mortem* des résultats du système par rapport aux réponses attendues va nous permettre d'établir des priorités pour nos efforts futurs.

Références

AYACHE C. (2005). *Campagne EVALDA/EQueR – Évaluation en Question-Réponse, rapport final*. Rapport interne, ELDA, Paris. Disponible à`http://www.technolangue.net/IMG/pdf/rapport_EQUER_1.2.pdf`.

MALAISÉ V., DELBECQUE T. & ZWEIGENBAUM P. (2005). Recherche en corpus de réponses à des questions définitoires. In *Actes Traitement automatique des langues naturelles (Traitement automatique des langues naturelles)*, Dourdan. *À paraître*.

TALN 2005 - RECITAL 2005

12ème conférence annuelle sur le Traitement Automatique des Langues Naturelles

9ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues Naturelles

---

ATELIER

DEFT

DÉFI FOUILLE DE TEXTES

Atelier des conférences TALN'05/RECITAL'05

DEFT'05 (DÉfi Fouille de Textes)

Page Web : http://www.lri.fr/ia/fdt/DEFT05/

10 juin 2005, 14h-17h30, Dourdan (91)

1 Motivations

Le défi proposé est motivé par le besoin de mettre en place des techniques de fouille de textes permettant soit d'identifier des phrases non pertinentes dans des textes, soit d'identifier des phrases particulièrement singulières dans des textes apparemment sans réel intérêt.

Le but du défi proposé consiste à supprimer les phrases non pertinentes dans un corpus de discours politiques en français. Ce défi porte le nom de DEFT (**DÉ**fi **F**ouille de **T**extes).

Dans les corpus spécialisés (biologie, médecine, etc.) un travail conséquent est dédié à l'identification des phrases pertinentes pour ensuite y rechercher des informations spécifiques. Ce type de tâche consistant à effectuer un premier filtrage des textes est une étape préliminaire essentielle à effectuer pour la constitution de corpus pertinents et homogènes. Le défi DEFT'05 est relatif à une telle tâche.

L'étape suivante peut consister à rechercher des informations précises dans ces textes filtrés. Le défi proposé ne s'intéresse pas à ce travail qui fait l'objet d'autres défis telle que la tâche *questions/réponses* du défi international TREC[1].

Le défi que nous proposons est plus proche de la tâche *Novelty* du challenge TREC. La première partie de la tâche *Novelty* de TREC consiste à identifier les phrases pertinentes

[1] http ://trec.nist.gov

puis, parmi celles-ci, les phrases nouvelles d'un corpus d'articles journalistiques. DEFT'05 qui consiste à supprimer les phrases non pertinentes d'un corpus de discours politiques est assez proche du travail d'identification des phrases pertinentes de la tâche *Novelty* du challenge TREC.

Nous proposons ici une liste non exhaustive de tâches similaires à celle proposée dans DEFT'05 et pour lesquelles il devrait être possible de réutiliser avec peu de modifications les approches mises en œuvre pour répondre à DEFT'05.

- détection des passages les plus singuliers dans des textes quelconques (rupture de style, changement de contexte) ;
- détection de plagiats possibles dans des textes ;
- détection des informations générales dans des corpus techniques.

2 Tâches à réaliser pour DEFT'05

Un corpus de textes, issus de la Présidence de Jacques Chirac (1995-2005), est fourni aux participants. Ce corpus est composé d'allocutions officielles du Président. Dans ce corpus, des passages issus d'un corpus d'allocutions du Président de la République François Mitterrand (1981-1995) sont insérés. Les passages d'allocutions de F. Mitterrand insérés sont composés d'au moins deux phrases successives.

Les passages de F. Mitterrand introduits traitent d'une thématique différente. Par exemple, dans les allocutions de J. Chirac évoquant la politique internationale, les phrases de F. Mitterrand introduites sont issues de discours traitant de politique nationale. Ainsi, la rupture thématique peut être une des manières de détecter les phrases issues du corpus de F. Mitterrand.

Un corpus avec des passages extraits d'allocutions de F. Mitterrand introduits dans les textes de J. Chirac est alors constitué. Certaines informations sont supprimées de ce corpus (années et noms de personnes) afin de constituer les données ci-dessous :

- **Corpus 1 :** Corpus sans la présence d'années ni de noms de personnes : les années et les noms de personnes sont remplacés par les balises `<date>` et `<nom>`.

- **Corpus 2 :** Corpus sans années : les années sont remplacées par la balise `<date>`.

- **Corpus 3 :** Corpus avec la présence des années et des noms de personnes.

Le but du défi consiste à déterminer les phrases issues du corpus de F. Mitterrand introduites dans le corpus composé d'allocutions de J. Chirac.

3 Comités

3.1 Comité d'Organisation

Responsables : Jérôme Azé (LRI - IA) et Mathieu Roche (LRI - IA)

Membres :

- Thomas Heitz (LRI - IA)
- Amar-Djalil Mezaour (LRI - IASI)
- Érick Alphonse (INRA - MIG)
- Ahmed Amrani (ESIEA & LRI - IA)

3.2 Comité de Programme

Présidents : Violaine Prince (LIRMM - TAL) et Yves Kodratoff (LRI - IA)

Membres :

- Nathalie Aussenac-Gilles (IRIT)
- Valérie Beaudouin (France Telecom)
- Catherine Berrut (CLIPS - MRIM)
- Béatrice Daille (LINA - LeC)
- Patrick Gallinari (LIP6 - Connexioniste)
- Éric Gaussier (Xerox Research)
- Thierry Hamon (LIPN - RCNL)
- Fidélia Ibekwe (ERSICOM)
- Michèle Jardino (LIMSI - LIR)
- Éric Laporte (IGM-LabInfo - Informatique linguistique)
- Josiane Mothe (IRIT, SIG)
- Xavier Polanco (INIST - URI)
- Pascal Poncelet (LGI2P - EMA)
- Christian Retoré (LABRI - SIGNES)
- Christophe Roche (LISTIC - Condillac)
- Pascale Sébillot (IRISA - TEXMEX)
- Yannick Toussaint (LORIA - Orpailleur)
- François Yvon (ENST)

Remerciements

Nous remercions les organisateurs de TALN'05 de nous avoir permis de mettre en œuvre cet atelier dans les meilleures conditions possibles. Nous remercions l'association **AFIA**[2] (Association Française d'Intelligence Artificielle) pour le prix de 300 euros attribué au gagnant de DEFT'05 ainsi que Jérémie Mary pour la conception du logo de ce défi. Enfin, n'oublions pas de remercier et surtout de féliciter les onze équipes issues de neuf laboratoires différents qui ont participé à DEFT'05.

[2] http ://afia.lri.fr/

Préparation des données et analyse des résultats de DEFT'05

Érick Alphonse (1), Ahmed Amrani (2,3), Jérôme Azé (3),
Thomas Heitz (3), Amar-Djalil Mezaour (4), Mathieu Roche (3)

(1) MIG - INRA
Domaine de Vilvert, 78350 Jouy en Josas Cedex
Erick.Alphonse@jouy.inra.fr
(2) ESIEA Recherche
9 rue Vésale - 75005 Paris
amrani@esiea.fr
(3) Équipe IA, LRI - Université Paris-Sud
Bât. 490, 91405 Orsay Cedex
{amrani,aze,heitz,roche}@lri.fr
(4) Équipe IASI, LRI - Université Paris-Sud
Bât. 490, 91405 Orsay Cedex
mezaour@lri.fr

Résumé Le **DÉ**fi **F**ouille de **T**extes a consisté à supprimer les phrases non pertinentes dans un corpus de discours politiques en français. Il a eu lieu en 2005 et réuni onze équipes, totalisant une trentaine de participants. Cet article décrit les prétraitements effectués sur les corpus de F. Mitterrand et de J. Chirac dans le cadre de ce défi. Notamment, la conversion au format texte, le découpage en phrase, le classement des discours, l'introduction de phrases de F. Mitterrand dans les discours de J. Chirac et l'identification des dates et noms de personnes. Les résultats obtenus par les onze équipes participantes sont aussi présentés.

Abstract The text-mining challenge (DEFT) consisted of removing non relevant sentences from French corpora of political speeches. It took place in 2005 and brought together about thirty participants from eleven teams. This paper describes the preprocessings carried out on the corpora of F. Mitterrand and J. Chirac within the framework of this challenge. In particular, conversion to text format, sentence segmentation, classification of the speeches, introduction of F. Mitterrand's sentences into J. Chirac's speeches and identification of dates and people's names. The results obtained by the eleven participating teams are also presented.

1 Introduction

Le but du défi proposé consiste à supprimer les phrases non pertinentes dans un corpus de discours politiques en français. Ce défi porte le nom de DEFT (**DÉ**fi **F**ouille de **T**extes).

Ce défi, proche de la tâche *Novelty* du challenge TREC[1] (Soboroff, Harman, 2003; Amrani *et al.*, 2004), est motivé par le besoin de mettre en place des techniques de fouille de textes permettant soit d'identifier des phrases non pertinentes dans des textes, soit d'identifier des phrases particulièrement singulières dans des textes apparemment sans réel intérêt. Cette étape est préliminaire à tout processus d'extraction d'informations.

Par exemple, dans les corpus spécialisés (biologie, médecine, *etc.*) un travail conséquent est dédié à l'identification des phrases pertinentes pour ensuite y rechercher des informations spécifiques. Ce type de tâche consistant à effectuer un premier filtrage des textes est une étape préliminaire essentielle à effectuer pour la constitution de corpus pertinents et homogènes.

Ce type de prétraitements est aussi utilisé dans des tâches type *questions/réponses* (voir TREC).

Nous proposons ici une liste non exhaustive de tâches similaires à celle proposée dans DEFT'05 et pour lesquelles il devrait être possible de réutiliser avec peu de modifications les approches mises en œuvre pour répondre à DEFT'05.

- détection des passages les plus singuliers dans des textes quelconques (rupture de style, changement de contexte);
- détection de plagiats possibles dans des textes;
- détection des informations générales dans des corpus techniques.

Un corpus de textes, issus de la Présidence de Jacques Chirac (1995-2005), a été fourni aux participants de DEFT'05. Ce corpus est composé d'allocutions officielles du Président. Dans ce corpus, des passages issus d'un corpus d'allocutions du Président de la République François Mitterrand (1981-1995) sont insérés. Les passages d'allocutions de F. Mitterrand insérés sont composés d'au moins deux phrases successives. Chaque discours de J. Chirac contient zéro ou un passage extrait d'une allocution de F. Mitterrand.

Les passages de F. Mitterrand introduits traitent d'une thématique différente. Par exemple, dans les allocutions de J. Chirac évoquant la politique internationale, les phrases de F. Mitterrand introduites sont issues de discours traitant de politique nationale. Ainsi, la rupture thématique peut être une des manières de détecter les phrases issues du corpus de F. Mitterrand.

Cet article décrit plus spécifiquement les prétraitements effectués sur les corpus de F. Mitterrand et de J. Chirac.

La figure 1 illustre l'ensemble des traitements effectués pour DEFT'05.

[1] http://trec.nist.gov

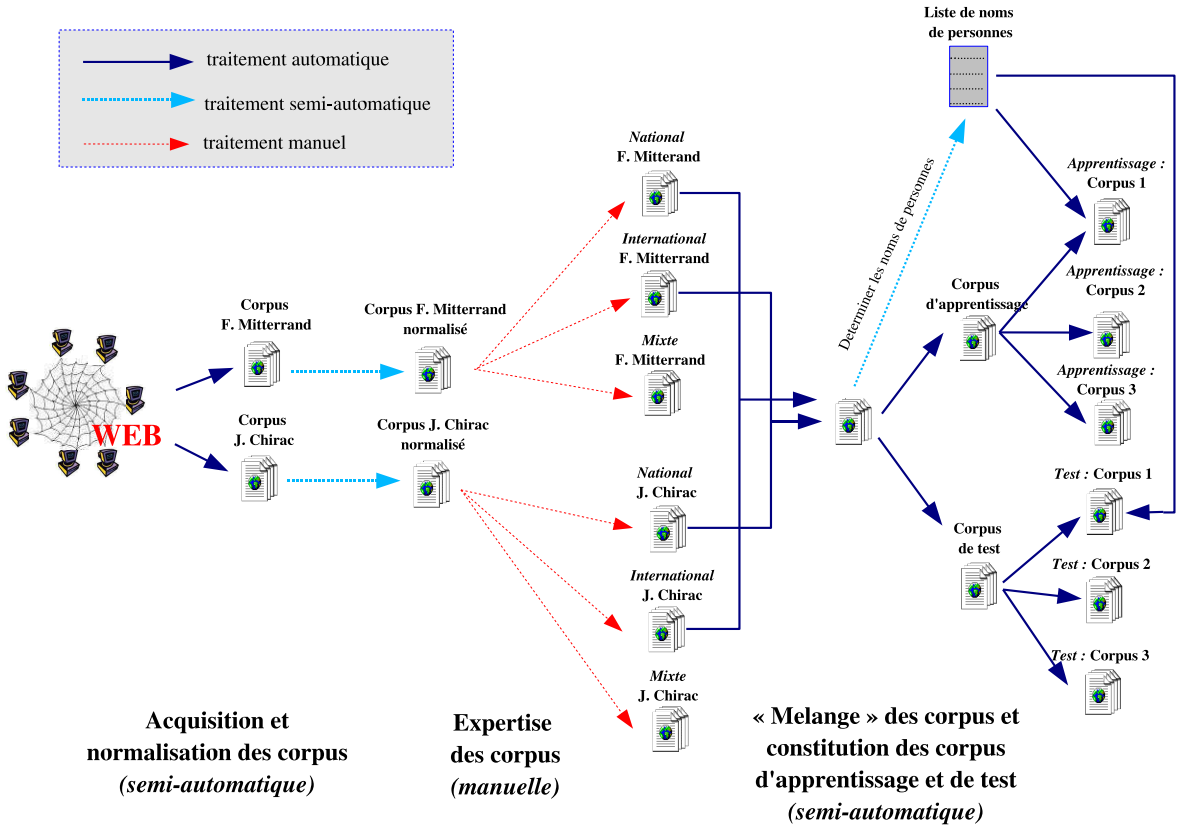


Figure 1: Chaîne globale de traitements de DEFT'05.

2 Acquisition des corpus

Les corpus composés des allocutions de J. Chirac et de F. Mitterrand ont été obtenus à partir des sites Web suivants :

- Corpus de J. Chirac (14 Mo sans considérer les balises HTML) : http://elysee.fr/
- Corpus de F. Mitterrand (17 Mo sans considérer les balises HTML) : http://discours-publics.ladocumentationfrancaise.fr/

3 Normalisation des corpus

Les corpus d'allocutions ont demandé un nombre de prétraitements important. Après avoir supprimé les commentaires et les balises HTML, les en-têtes des allocutions ont été enlevées (dates, lieux, *etc.*). Puis les entités au format SGML ont été transformées en caractères ISO8859-1. Par exemple, les entités « é » ont été remplacées par le caractère « é ».

Chacune des lignes des corpus fournis aux participants est composée d'une seule phrase. Pour identifier les phrases, il est nécessaire de repérer les ponctuations de fin de phrases (point final, point d'exclamation, point d'interrogation). Notons que comme dans les travaux de (Smadja, 1993), cette tâche nécessite le fait de ne pas considérer tous les points comme des ponctuations de fin de phrases (par exemple, les abréviations telles que R.M.I. pour Revenu Minimum d'Insertion ou M. pour Monsieur). De manière similaire aux travaux de (Rudolf, Świdziński, 2004), nous pouvons considérer que les points peuvent avoir des rôles spécifiques et sont utilisés dans différentes situations : abréviations, adresses internet, numéros de sections, *etc.*

Chaque locuteur peut utiliser régulièrement des phrases types du domaines qui pourraient permettre d'identifier les allocuteurs. À titre d'exemple, nous avons supprimé l'expression « Vive la République » qui est plus fréquente dans le corpus de F. Mitterrand (216 fois dans le corpus de F. Mitterrand contre 48 fois dans le corpus de J. Chirac).

Enfin, chaque phrase a été indexée rigoureusement grâce à une numérotation spécifique.

4 Expertise des corpus

Une étape d'expertise manuelle a alors été effectuée à partir des corpus normalisés. Le but de cette expertise a consisté à associer une catégorie à chacun des textes du corpus. Trois catégories ont été déterminées par le comité d'organisation (les auteurs de cet article) :

- Catégorie nationale
- Catégorie internationale
- Catégorie mixte ou ambiguë

Un discours traitant à 80% (estimation) d'une thématique déterminée sera associée à cette catégorie. Les discours contenant moins de 80% d'une thématique ont été associés à la catégorie mixte et ont été supprimés des données utilisées pour créer les corpus fournis aux participants.

Au total, 2523 textes ont été expertisés par les six organisateurs : 1200 allocutions de J. Chirac et 1323 allocutions de F. Mitterrand. Sur ces 2523 textes, 36.6% des textes ont été associés à la catégorie Nationale, 47.2% à la catégorie Internationale et 16,2% à la catégorie Mixte (voir tableau 1). Le détail complet des résultats donnés dans le tableau 1 montre notamment que les discours de F. Mitterrand ont davantage été associés à la catégorie Nationale que les allocutions de J. Chirac.

	F. Mitterrand	J. Chirac	**Global**
National	40.8%	31.9%	**36.6%**
International	45.0%	49.7%	**47.2%**
Mixte	14.1%	18.4%	**16.2%**

Table 1: Répartition des expertises par allocuteur.

Précisons que d'une période à l'autre, la répartition peut différer significativement. À titre d'exemples, les allocutions officielles de J. Chirac en 2002, année de l'élection présidentielle ont davantage été associées à la catégorie nationale (125 allocutions de J. Chirac en 2002 : 63 (50%) appartiennent à la catégorie Nationale, 46 (36.7%) appartiennent à la catégorie Internationale et 16 (12.8%) ont été associées à la catégorie Mixte).

5 Introduction des phrases de F. Mitterrand dans le corpus de J. Chirac

L'introduction des extraits de discours de F. Mitterrand dans les discours de J. Chirac a été réalisée en suivant les règles suivantes :

- Croisement des thématiques identifiées (politique nationale vs politique internationale)
- Sélection des extraits de discours de F. Mitterrand les plus "proches" des discours de J. Chirac pour l'introduction (voir paragraphe 5.2)
- Introduction d'au plus un passage de F. Mitterrand dans chaque discours de J. Chirac (voir paragraphe 5.3).

Le croisement des thématiques est lié à l'analyse présentée dans le tableau 1.

La distance entre un extrait de F. Mitterrand et un discours de J. Chirac est calculée en fonction des Ngrams de caractères et Ngrams mots. Nous avons calculé de manière systématique les Ngrams de caractères et de mots (pour n=1, 2 et 3) des discours de J. Chirac et des parties de discours de F. Mitterrand candidates à l'insertion (c-à-d. toutes les parties de discours sauf la première et la dernière).

Puis, nous avons comparé les discours de J. Chirac et parties de discours de F. Mitterrand (en tenant compte du croisement thématique) sur la base de ces Ngrams.

5.1 Comparaison des discours et des passages

Ces éléments sont comparés sur la base du score suivant :

$$score(d_C^{cat}, p_M^{\overline{cat}}) = score_{car}(d_C^{cat}, p_M^{\overline{cat}}) + score_{mot}(d_C^{cat}, p_M^{\overline{cat}})$$

avec

$$\begin{cases} cat & international\ ou\ national \\ \overline{cat} & catégorie\ opposée\ à\ cat \\ d_C^{cat} & discours\ de\ J.\ Chirac\ appartenant\ à\ cat \\ p_M^{cat} & partie\ de\ discours\ de\ F.\ Mitterrand\ appartenant\ à\ cat \end{cases}$$

$$score_{car}(d_C^{cat}, p_M^{\overline{cat}}) = \sum_{n=1}^{3} \left(\frac{1}{n}\right) \times 2 \times \frac{commun(d_C^{cat}, p_M^{\overline{cat}})}{|d_C^{cat}| + |p_M^{\overline{cat}}|}$$

où

$$\begin{cases} |x| & nombre\ de\ mots\ ou\ caractères\ de\ x \\ commun(d_C^{cat}, p_M^{\overline{cat}}) & nombre\ de\ mots\ ou\ caractères\ communs\ entre\ d_C^{cat}\ et\ p_M^{\overline{cat}} \end{cases}$$

$score_{mots}(d_C^{cat}, p_M^{\overline{cat}})$ est calculé selon la même formule mais sur la base des Ngrams[2] entre mots et non pas entre caractères.

5.2 Sélection des passages à insérer

Ayant calculé ce score pour tous les couples possibles $(d_C^{cat}, p_M^{\overline{cat}})$, nous retenons pour chaque d_C^{cat} les vingt "meilleurs" $p_M^{\overline{cat}}$ (c-à-d. tels que $score(d_C^{cat}, p_M^{\overline{cat}})$ soient les plus élevés). Ces vingt candidats à l'insertion sont triés par valeurs décroissantes du score.

Puis, les discours de J. Chirac sont parcourus aléatoirement et les insertions de passages de discours de F. Mitterrand sont réalisées de la manière suivantes :

Soient d_C^{cat} le discours de J. Chirac étudié et $\mathcal{L}_{p_M^{\overline{cat}}}^{d_C^{cat}}$ la liste des passages candidats à l'insertion. Soit $\mathcal{E}_{p_M^{\overline{cat}}}$ l'ensemble des passages de discours de F. Mitterrand déjà introduits dans des discours de J. Chirac.

La liste ordonnée $\mathcal{L}_{p_M^{\overline{cat}}}^{d_C^{cat}}$ est parcourue depuis le premier passage vers le dernier jusqu'à trouver un passage qui soit absent de $\mathcal{E}_{p_M^{\overline{cat}}}$. Si un tel passage existe, il est introduit dans d_C^{cat}, puis dans $\mathcal{E}_{p_M^{\overline{cat}}}$. Par contre, si aucun passage n'est trouvé alors le discours de J. Chirac étudié n'est pas modifié (c-à-d. le discours est donc "non bruité").

5.3 Insertion d'un passage de F. Mitterrand dans un discours de J. Chirac

La position d'un passage à insérer est déterminée en respectant les contraintes suivantes :

- ni avant le premier, ni après le dernier paragraphe[3] du discours de J. Chirac.

[2] L'outil **nsp-v0.71** a été utilisé pour calculer les Ngrams (`http://www.d.umn.edu/~tpederse/nsp.html`).

[3] Un paragraphe correspond à un bloc de texte entre balises HTML <p> ou séparé par deux balises <br>.

- aléatoirement dans le reste du discours et entre deux paragraphes

Le corpus ainsi constitué a été divisé en deux sous-ensembles : le corpus d'apprentissage et le corpus de test. Nous avons utilisé 70% des discours pour constituer le corpus d'apprentissage et les 30% restant pour le test. Les discours ont été choisis de manière aléatoire et stratifiée. En effet, nous avons garanti par construction que les proportions de discours "bruités" et "non bruités", dans les corpus de test et d'entraînement, sont identiques à celles observées dans le corpus initial.

5.4 Remarque

Il peut arriver que deux thématiques identiques (Nationale ou Internationale) soient insérées dans un même texte. Ceci peut s'expliquer par le fait qu'un texte de F. Mitterrand associé à une catégorie Nationale (resp. Internationale) peut comporter des passages d'une catégorie Internationale (resp. Nationale). Ces passages de F. Mitterrand de la catégorie Internationale (resp. Nationale) bien que minoritaires dans l'allocution associée à la catégorie Nationale (resp. Internationale) pourraient alors être introduits dans une allocution de J. Chirac de la catégorie Internationale (resp. Nationale).

6 Constitution des trois corpus avec et sans informations relatives aux noms de personnes et aux années

Nous rappelons que le défi DEFT'05 comporte trois tâches distinctes :

- **Tâche 1 :** Identifier les phrases de F. Mitterrand dans le corpus de test numéro 1 (corpus ne comportant ni années, ni noms de personnes).
- **Tâche 2 :** Identifier les phrases de F. Mitterrand dans le corpus de test numéro 2 (corpus ne comportant pas d'années).
- **Tâche 3 :** Identifier les phrases de F. Mitterrand dans le corpus de test numéro 3 (corpus avec la présence des années et des noms de personnes).

Pour constituer les corpus 1 et 2, nous avons du identifier les dates (années) ainsi que les noms de personnes. Ces identifications sont détaillées ci-dessous.

6.1 Identification des dates

Seules les années situées dans l'intervalle $[1900 : 2099]$ ont été identifiées. Ces années pourraient en effet faciliter l'identification des phrases issues du corpus de F. Mitterrand.

Ainsi les années de la forme 19xx et 20xx où « x » est un chiffre quelconque ont été identifiées et remplacées par une balise <date>. De même, les intervalles entre années ont été reconnus : 19xx-19xx, 19xx-20xx, 20xx-20xx et xx-xx.

Chacune des dates de ces intervalles ont également été remplacées par une balise <date>.

Les dates au format "1er février 2004" n'ont pas été identifiée et peuvent donc figurer dans les corpus, sous la forme "1er février <date>"

Ce traitement a permis de constituer les corpus utiles pour les tâches 1 et 2.

6.2 Identification des noms de personnes

Une liste de noms de personnes a dû être établie manuellement. Les membres du Comité d'Organisation de DEFT'05 ont ainsi analysés les suites de mots suivants afin d'identifier les noms de personnes :

- couples de mots commençant par une majuscule.
- couples de mots commençant par une majuscule avec une particule intercalée entre les deux mots.
- particule suivi d'un mot en majuscules.

Les particules utilisées (avec et sans majuscules) sont les suivantes : Abd, Al, Ap, Ben, Bin, D', Da, Dalle, Dall', Dell', De, De La, De Los, Del, Dela, Della, Delle, Den, Der, Di, Du, El, Ibn, La, Le, Li, Lo, Mac, Mc, O', Of, Saint, San, Van, Van Den, Van Der, Von, Von Der, y.

De plus, un dictionnaire de noms de personnes composés d'un seul mot a été constitué (par exemple, Picasso, Dali, *etc.*).

Les noms de personnes étant identifiés, nous les avons remplacés par une balise <nom>.

Ce traitement a permis de constituer le corpus utile pour la tâche 1.

7 Traitement final des corpus

Le dernier traitement a consisté à maintenir en majuscule seulement la première lettre des noms de personnes. En effet, dans le corpus de J. Chirac la plupart des noms de personnes sont écrits en majuscules (Jacques CHIRAC, François MITTERRAND, *etc.*). Ainsi, l'identification des noms en majuscules aurait pu être une règle simple mais efficace pour reconnaître les phrases issues du corpus de F. Mitterrand et de J. Chirac des tâches 1 et 2. Pour corriger cette situation, nous avons uniformisé l'écriture des noms de personnes en écrivant seulement en majuscule la première lettre du nom de personne : MITTERRAND → Mitterrand. Bien entendu les acronymes (PS, RPR, EDF, *etc.*) sont maintenus en majuscules. Certains noms de personnes écrits en majuscules contiennent moins d'informations qu'un même nom écrit en majuscules. En effet, dans des cas relativement nombreux, les noms en majuscules ne comportent pas d'accents (par exemple « JUPPE » qui correspond en lettres minuscules à « Juppé »). Les accents doivent donc être restitués lors du passage majuscules/minuscules. Une manière semi-automatique de procéder consiste à relever la présence des noms de personnes (nom commençant par une majuscule) que l'on trouve dans le texte avec des accents. Dans ce cas, nous pouvons décider d'apposer par défaut l'accent omis. Si aucun mot similaire (avec accents) n'est repéré dans le corpus, et sans utiliser de ressources extérieures, il est nécessaire d'expertiser ces

	Corpus d'apprentissage	Corpus de test
Taille moyenne des phrases successives de F. Mitterrand	18.8	19.1
Pourcentage d'allocutions sans phrases de F. Mitterrand insérées	31.9% (187/587)	32.3% (95/294)
Nombre d'étiquettes <nom> par rapport au nombre de mots du corpus	0.18% (2511/1420833)	0.21% (1331/616584)
Nombre d'étiquettes <date> par rapport au nombre de mots du corpus	0.13% (1846/1420833)	0.12% (774/616584)

Table 2: Comparaison des corpus d'apprentissage et de test.

noms et d'y apposer manuellement les accents manquants.

8 Similarités entre les corpus d'apprentissage et de test

Les corpus d'apprentissage et de test ont été constitués simultanément, c'est la raison pour laquelle ils ont des caractéristiques similaires. Ainsi, les méthodes mises en œuvre sur les corpus d'apprentissage peuvent être appliquées sur les corpus de test sans nécessiter d'adaptations spécifiques. Nous donnons dans le tableau 2 les caractéristiques essentielles des corpus d'apprentissage et de test.

Remarquons que le pourcentage d'étiquettes <nom> dans les corpus de test est plus élevé que dans les corpus d'apprentissage (voir tableau 2). Cela peut s'expliquer par le fait que nous avons apporté une attention toute particulière à la préparation des corpus de test pour lesquels les participants avaient seulement deux à quatre jours de traitements possibles.

9 Résultats obtenus par les équipes participantes

Onze équipes ont participé à DEFT'05. Ces onze équipes sont issues de neuf laboratoires différents et totalisent une trentaine de participants.

Les résultats obtenus sont assez variés (en termes de précision, rappel et Fscore) et tendent donc à montrer que les tâches à réaliser étaient non triviales, tout en restant faisables. Nous rappelons que toutes les exécutions ont été évaluées en calculant le F_{score} (avec $\beta = 1$, voir formule (1)).

$$F_{score}(\beta) = \frac{(\beta^2 + 1) \times Précision \times Rappel}{\beta^2 \times Précision + Rappel} \quad (1)$$

Le tableau 3 présente les Fscores obtenus par les différentes équipes pour chaque tâche (nous avons présenté les Fscores moyens calculés à partir des différentes exécutions soumises). Ces résultats sont triés par Fscore décroissant sur la base de la première tâche. Les résultats indiqués en italiques correspondent à des équipes n'ayant soumis qu'une seule exécution pour la tâche concernée.

	tâche 1	tâche 2	tâche 3
équipe 1	0.870239	0.884376	0.880458
équipe 2	0.860471	0.851721	0.866254
équipe 3	0.819636	0.821018	0.819075
équipe 4	0.759816	0.741895	0.745863
équipe 5	0.751278	0.754767	0.755094
équipe 6	0.731591	0.793889	0.788093
équipe 7	0.561811	0.559366	0.573122
équipe 8	0.493809	0.521142	0.507484
équipe 9	0.49241	0.560066	0.563089
équipe 10	0.325341	0.306647	0.305337
équipe 11	0.176951	0.176951	0.41729

Table 3: Meilleurs Fscores des différents équipes pour chaque tâche.

L'analyse des résultats détaillés (par exécution) sur la base du Fscore avec $\beta = 1$ permet de voir que la plupart des équipes ont amélioré leurs résultats au fur et à mesure des tâches (voir les courbes 2). Ainsi, l'ajout d'informations (noms de personnes, puis années) représente une aide réelle pour les différents systèmes représentés dans ce défi.

De plus, l'analyse du front de Pareto associé à la tâche 1 (voir Figure 3) montre que plusieurs équipes se trouvent sur le front de Pareto et donc qu'en fonction de la valeur de β choisie, l'ordre des approches en fonction du Fscore peut être modifié. Nous obtenons les mêmes résultats pour les tâches 2 et 3.

10 Conclusion

La problématique abordée dans DEFT'05 est relative à une tâche importante dans tout processus de fouille de données et constitue une étape préliminaire aux phases d'extraction d'informations.

L'implication de nombreuses équipes de recherche dans ce défi montre l'intérêt réel de la communauté pour ce problème et notamment pour la comparaison et l'évaluation de différentes méthodes de prétraitement des données et d'extraction d'informations.

La diversité des résultats obtenus par les équipes ayant participé montre que cette tâche représente une réelle difficulté pour la communauté et l'un des avantages de DEFT'05 est lié à la nature artificielle du corpus qui permet ainsi une évaluation plus objective des résultats obtenus par les différentes équipes.

L'engouement de la communauté pour ce défi et les différentes propositions d'extensions de DEFT vont permettre la poursuite de ce défi l'année prochaine. Les tâches précises restent à déterminer mais DEFT'05 a réussi à fédérer la communauté francophone de fouille de textes. Une extension envisageable et intéressante serait liée à l'étude de données réelles et à la définition d'un nouveau problème pour DEFT'06.

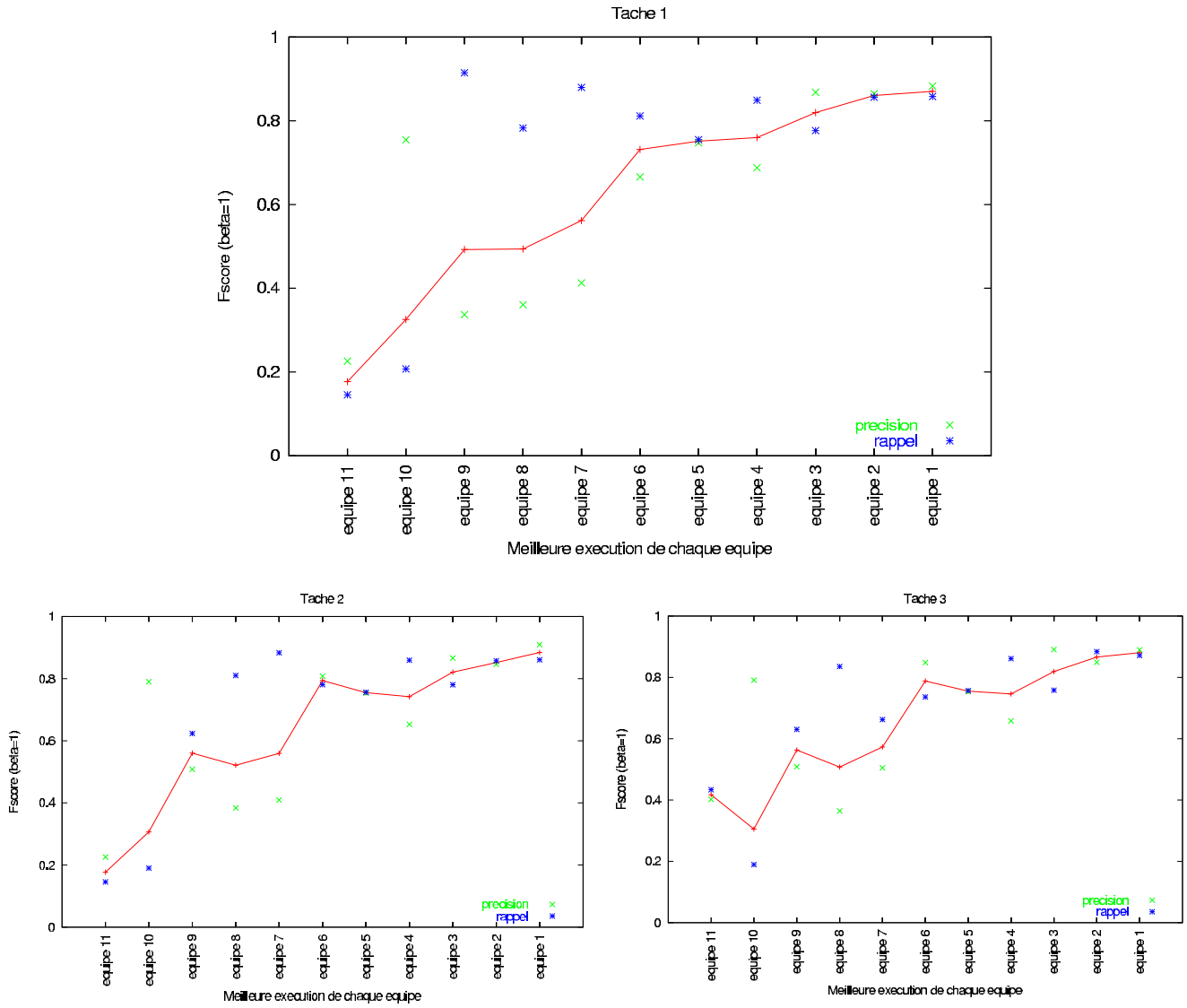


Figure 2: Fscore ($\beta = 1$) pour les meilleurs exécutions.

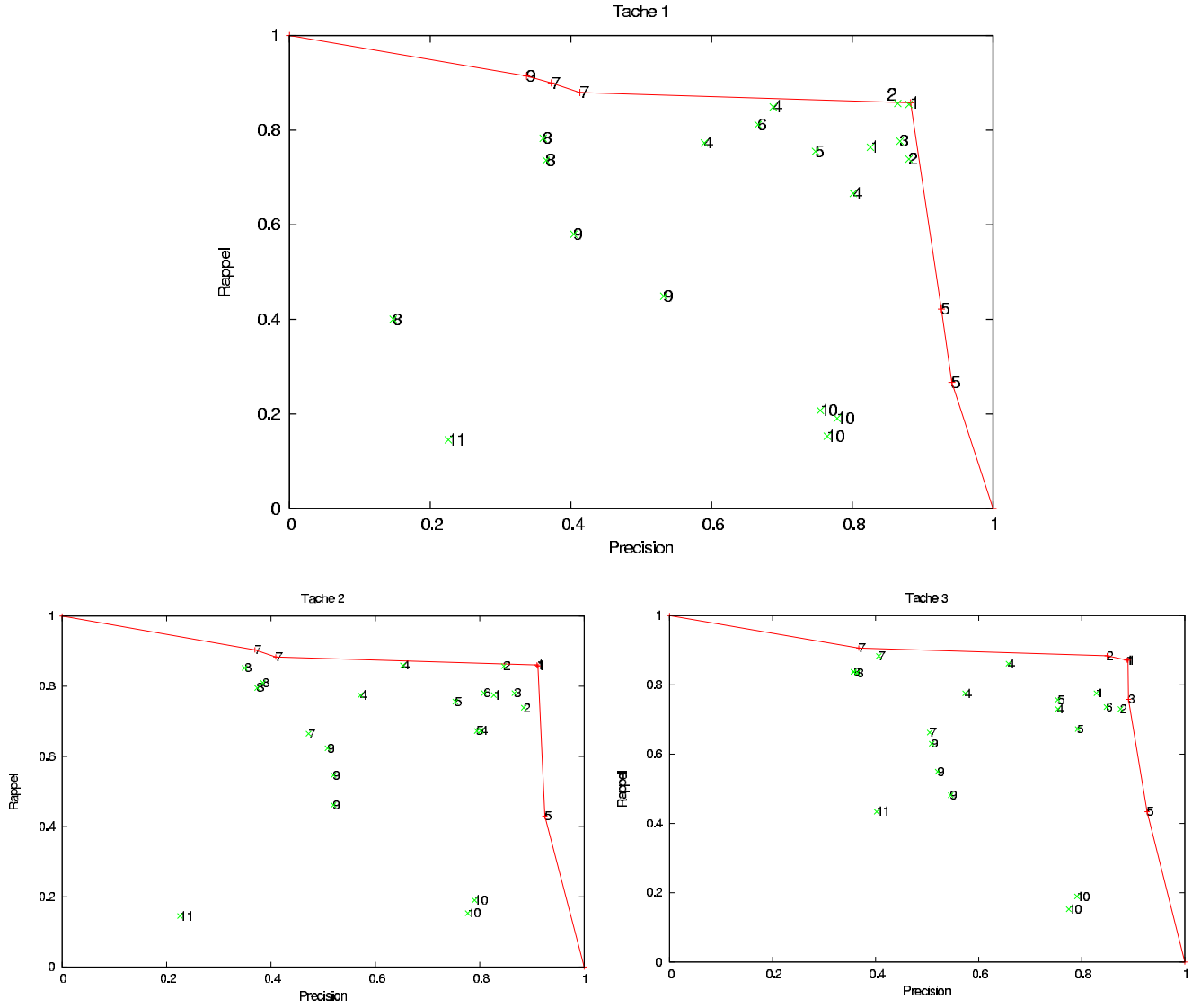


Figure 3: Front de Pareto pour les tâches 1, 2 et 3.

Références

Rudolf M., M. Świdziński (2004), Automatic utterance boundaries recognition in large Polish text corpora, *Proceedings of IIPWM'04 (Intelligent Information Processing and Web Mining), Springer Verlag series "Advances in Soft Computing"*, 247-256.

Smadja F. (1993), Retrieving collocations from text: Xtract, *Computational Linguistics*, Vol. 191, p143-177.

Soboroff I., Harman D. (2003), Overview of the TREC 2003 Novelty Track, *NIST Special Publication: SP 500-255 The Twelfth Text Retrieval Conference (TREC 2003).*

Amrani. A, Azé. J, Heitz. T, Kodratoff. Y and Roche. M (2004), From the texts to the concepts they contain: a chain of linguistic treatments, *Proceedings of TREC'04 (Text REtrieval Conference), National Institute of Standards and Technology, Gaithersburg Maryland USA*, pages 712-722.

Atelier DEFT'05, TALN 2005, Dourdan, 6-10 juin 2005

Application des vecteurs sémantiques à la fouille de texte

Jacques Chauché
LIRMM-CNRS et Université Montpellier 2
161 rue Ada, 34395 Montpellier cedex 5
chauche@lirmm.fr

Mots-clefs : analyse syntaxique, analyse sémantique, similitude sémantique

Keywords: syntactic analysis, semantic analysis, sémantic similitude

Résumé L'approche présentée ici se base sur un traitement du contenu syntaxico-sémantique par un analyseur du Français, le système SYGFRAN, pour retrouver un ensemble de phrases appartenant à différents discours du président François Mitterand plongées dans un ensemble de phrases appartenant à différents discours du président Jacques Chirac. Ce traitement se fait par calcul de vecteurs sémantiques de phrases (méthodologie définie dans l'article) et par la définition d'une relation de similitude décrivant l'inclination de vecteurs dont l'inclinaison, ou distance angulaire, est proche. A l'aide de cette relation, des phrases sont attribuées par le système à l'un ou l'autre des auteurs, et l'article indique des F-mesures obtenues sur le premier corpus, dit d'apprentissage, légèrement supérieures à 80%.

Abstract The approach presented here is based on a treatment of the syntactico-semantics contents by an analyzer of French, system SYGFRAN, to find a group of sentences belonging to various speeches of president François Mitterand mixed with a group of sentences belonging to various speeches of president Chirac. This treatment is done by a calculation of semantic vectors of sentences (methodology defined in the article) and by the definition of a relation of similarity describing the inclination of vectors to which the slope, in angular distance, is close. Using this relation, sentences are allotted by the system to one or the other of the authors, and the article indicates the F-measurements obtained on the first corpus (also called training corpus) slightly higher than 80%.

Le défi 2005 organisé pour le congrès annuel TALN consiste à retrouver un ensemble de phrases appartenant à différents discours du président François Mitterand plongées dans un ensemble de phrases appartenant à différents discours du président Jacques Chirac. Les phrases introduites traitent d'une thématique distincte de la thématique retenue pour les phrases des discours de Jacques Chirac. L'approche présentée ici se fonde sur un traitement du contenu par opposition aux traitements habituels basés sur des approches statistiques. Le vocabulaire utilisé par l'un ou l'autre n'aura d'importance qu'à travers les idées qu'il véhicule.

1 Vecteur sémantique

1.1 Vecteur de terme

Définition
Un vecteur sémantique projette un terme donné dans un espace sémantique dont une famille génératrice correspond à un ensemble d'idées.

L'ensemble des idées nécessaires pour former une famille génératrice peut être définie par un thésaurus.

La procédure est la suivante : on projette la totalité des lexies du dictionnaire sur un espace défini à partir d'une famille de concepts "à la Roget" (Roget 1852). Pour le Français, les lexicologues du Larousse ont défini une famille de 873 concepts hiérarchisés en 4 niveaux (Larousse 1992). Sur un plan vectoriel, cela produit un espace à 873 dimensions que l'on admet comme étant de dimension donnée. Les approches à la "Roget" sont relativement nombreuses depuis quelques années, dans la littérature anglo-saxonne, (Yarowsky, 1992), (Ellman et Tait 1999). En Français, l'indexation automatique à partir du thésaurus a été proposée à l'origine par nous-mêmes, (Chauché 1990), mais on la retrouve aujourd'hui utilisée dans de nombreux travaux (Crestan et al. 2003).

Formellement, on considère que tout terme t du dictionnaire est représenté par un vecteur $\vec{t}$ dans l'espace vectoriel considéré, que l'on nommera $\vec{\mathcal{V}}$. On suppose qu'il existe une application qui plonge l'espace lexical linguistique dans l'espace vectoriel engendré par la famille de concepts du thésaurus. Pour des besoins de calcul, seule une version normée $\vec{t}_{nor}$ de ce vecteur est conservée dans l'espace. Comme on ne traite que de vecteurs normés, par convention, on écrira $\vec{t}$ pour désigner le vecteur normé du terme t. Pour cela, on introduit une norme euclidienne sur l'espace vectoriel sémantique.

La majorité des mots,étant polysémique, renvoie à une multiplicité d'idées, ou concepts du thésaurus.

Exemple
Les idées associées au mot *calcul* sont par exemple : Calcul, Opération arithmétique, Maladie et Intention.

L'emploi de ce mot simplement ne permet donc pas de définir sa signification: par exemple, *calcul arithmétique* ou *calcul biliaire*, ou *Il m'a aidé par calcul.*

Cela signifie que le terme doit être représenté, non seulement par la manière dont il est indexé dans le thésaurus, mais aussi par ses différentes significations, qui elles, ont un sens lorsque le mot est utilisé dans une construction (groupe ou phrase).

Le calcul sémantique sur une phrase doit donc incliner le sens du mot “calcul” vers une des significations possibles.

1.2 Vecteur sémantique d’une phrase

Définition
On dira que l’on représente toute *phrase* construite, par un vecteur produit comme une combinaison linéaire de vecteurs sémantiques des *groupes* qui la composent.
On dira que l’on représente tout *groupe* construit, par un vecteur produit comme une combinaison linéaire de vecteurs sémantiques des *termes* qui le composent.

Pour cela on introduit les opérations suivantes :
Somme normée : Soient deux vecteurs $\vec{t_1}$, et $\vec{t_2}$ représentant les vecteurs (normés) de deux termes t_1 et t_2.

$$\overrightarrow{(t_1 + t_2)_{nor}} = \frac{\vec{t_1}+\vec{t_2}}{\|\vec{t_1}+\vec{t_2}\|} \tag{1}$$

Remarque : la somme normée n’est pas associative :
$\overrightarrow{(t_1 + t_2 + t_3)_{nor}}$ n’est pas égal à $(\overrightarrow{(t_1 + t_2)_{nor}} + \vec{t_3}))_{nor}$. Par convention, on ne retiendra comme opération de somme que la somme normée, et on omettra dorénavant l’indice ’nor’.

Multiplication par un scalaire : Soit un vecteur $\vec{t}$ normé. Soit λ un scalaire. Le vecteur $\vec{\lambda t}$ est égal à $\lambda * \vec{t}$. Cela signifie que toutes les composantes du vecteur sont multipliées par le scalaire.

Remarque : cette multiplication a pour objectif de renforcer la "présence" du vecteur dans une combinaison linéaire, et ne s’utilise en principe jamais isolément.

Produit terme à terme : Soient deux vecteurs $\vec{t_1}$, et $\vec{t_2}$ normés. Le produit terme à terme des deux vecteurs se définit comme :

$$\overrightarrow{(t_1 * t_2)_{nor}} = \frac{\vec{t_1}*\vec{t_2}}{\|\vec{t_1}*\vec{t_2}\|} \tag{2}$$

où si $a_{p,i}$ est la ième composante de $\vec{t_1}*\vec{t_2}$, et $a_{1,i}$ et $a_{2,i}$ respectivement celles de $\vec{t_1}$, et $\vec{t_2}$, on a :

$$\forall i \in [1, 873], a_{p,i} = a_{1,i} * a_{2,i} \tag{3}$$

Par convention, on omettra l’indice nor et on appellera par défaut $\overrightarrow{(t_1 * t_2)}$ le produit terme à terme normé.

Distance "angulaire": La distance selon Salton, servant de mesure de similarité est calculée comme le *cosinus* de l’angle de deux vecteurs.

$$sim(\vec{t_1}, \vec{t_2}) = \cos \widehat{\vec{t_1}, \vec{t_2}} = \frac{\vec{t_1}.\vec{t_2}}{\|\vec{t_1} * \vec{t_2}\|} \tag{4}$$

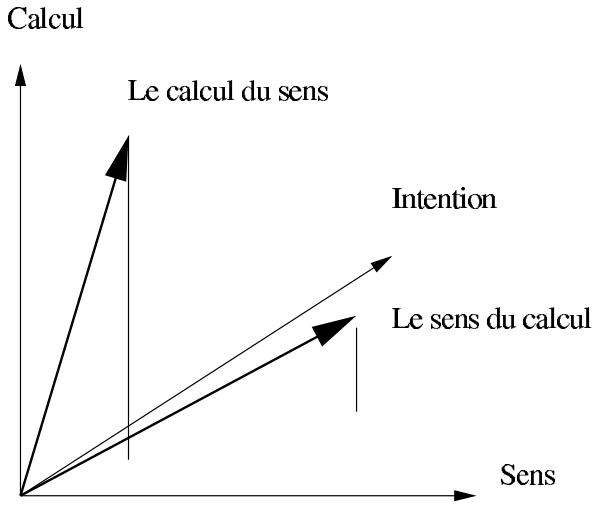
où "." est le produit vectoriel classiquement défini. La distance que nous utilisons correspond à une mesure relative à l’angle $\widehat{\vec{t_1}, \vec{t_2}}$. Comme nous ramenons tous les angles considérés à l’espace $[0, \frac{\pi}{2}]$, alors la mesure que nous proposons se calcule par :

$$\delta(\vec{t_1}, \vec{t_2}) = 1 - \cos \widehat{\vec{t_1}, \vec{t_2}} \quad (5)$$

Remarques: Ramener les valeurs de δ à $[0, 1]$ est plus pratique que de mesurer des valeurs entre 0 et $1,67$ radiants. Lorsque deux vecteurs sont totalement divergents (intersection vide), leur angle est de $\frac{\pi}{2}$, et le cosinus vaut 0 : leur distance est maximale et vaut 1. Lorsque ces vecteurs sont très proches, leur angle tend vers 0 , le cosinus tend vers 1 et la distance, vers 0. Tous les vecteurs ont un angle forcément compris entre 0 et $\frac{\pi}{2}$, par construction, et appartiennent au même espace vectoriel.

1.3 Vecteur de groupe

La deuxième propriété du calcul sémantique correspond à une définition différentiée d'un groupe suivant sa structure. Ainsi le sens du groupe "le calcul du sens" est distinct du sens du groupe "le sens du calcul",ces deux groupes ayant rigoureusement les mêmes éléments (le langage naturel n'étant pas commutatif). Comme le mot "sens" est très riche sémantiquement (une vingtaine de sens justement) nous prendrons pour l'exemple de la représentation l'idée associée : Sens.L'id/'e est différente du terme, selon les lexicologues, en ce qu'elle étiquette un champ sémantique. Le terme peut appartenir ou relever de plusieurs champs, en raison de sa polysémie. Dans le sous-espace ayant comme axe *Calcul*, *Intention* et *Sens* les vecteurs associés aux deux groupes précédents seront :



1.4 Calcul du vecteur de phrase

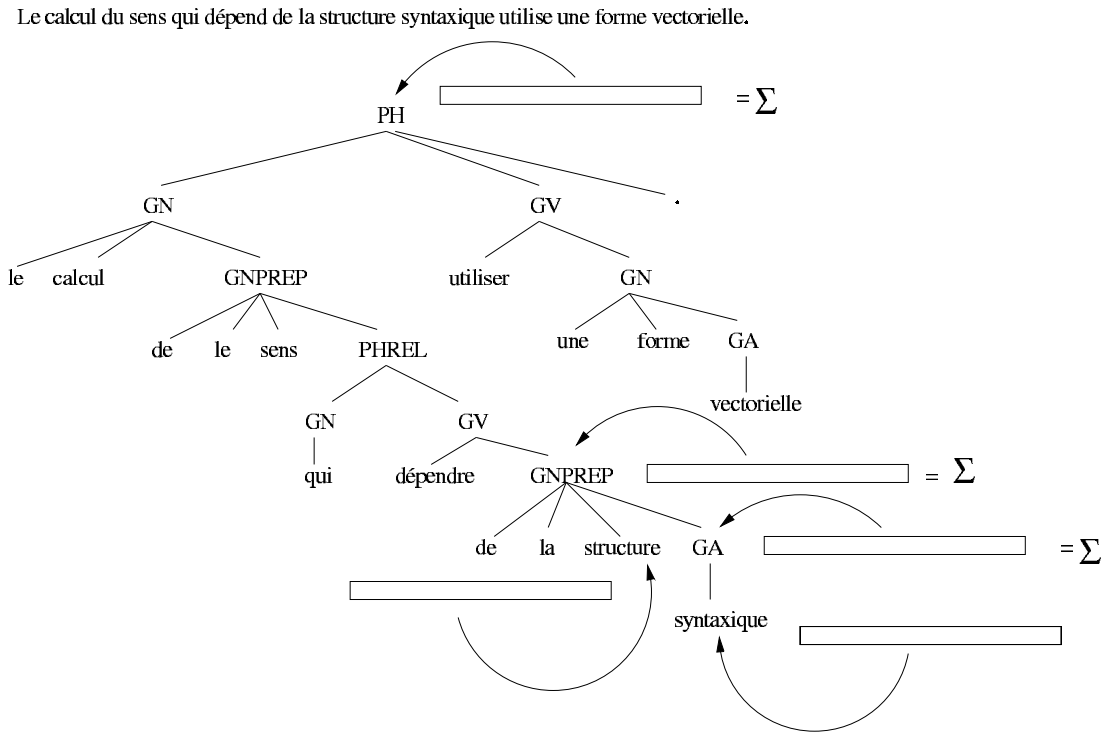
Le calcul d'un vecteur de phrase s'effectue (sur une phrase) en plusieurs étapes à partir de la structure syntaxique :

- La première étape consiste à associer à chaque feuille un vecteur sémantique issu de la lecture d'un dictionnaire (vecteur de terme)

Si un élément à plusieurs sens ou interprétations possibles, le vecteur associé correspond au *centroïde* de l'ensemble des vecteurs associés à chaque interprétation (somme normée de tous les vecteurs indexant ce terme).

- La deuxième étape consiste à calculer récursivement le vecteur associé à chaque groupe.

Le vecteur associé à un groupe est obtenu par une combinaison linéaire des vecteurs associés aux éléments de ce groupe. Les coefficients de cette combinaison linéaire dépendent de la fonction syntaxique de l'élément : gouverneur du groupe, sujet, objet, etc...



- La troisième étape actualise les vecteurs associés aux feuilles. Cette actualisation consiste à effectuer un produit terme à terme du vecteur à actualiser avec le vecteur obtenu du texte.

Cette actualisation terminée un nouveau calcul est effectué. La convergence est très rapide et deux itérations suffisent pour obtenir un vecteur significatif.

1.5 Propriétés du modèle

Le classement s'effectue à partir des vecteurs sémantiques de phrases.

La valeur intrinsèque de la norme d'un vecteur n'est pas significative. Seul compte l'inclinaison de ce vecteur par rapport à une idée ou un autre vecteur donné (la distance angulaire). Aussi tous les calculs se termineront par la normalisation des vecteurs. On ne considérera donc que les points de la shpère unité.

1.5.1 Comparaison d'inclinaison entre vecteurs

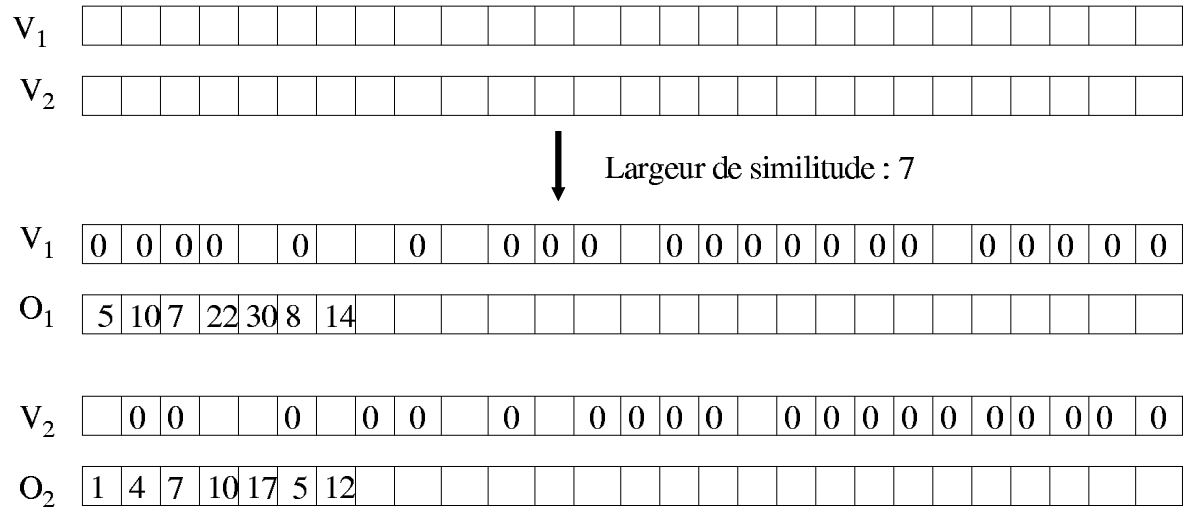
La première mesure de comparaison sera donc la valeur de l'arc séparant deux vecteurs sur cette sphère (Cette mesure sera donc naturellement donnée par la fonction *arcosinus* $(\vec{V_1}, \vec{V_2})$. Comme toutes les composantes de tous les vecteurs sont positives ou nulles nous utiliserons

seulement le produit scalaire. Dans ce contexte un élément sera plus proche d'un autre par rapport à un troisième si le produit scalaire de cet élément avec le troisième a une valeur supérieure au produit scalaire du deuxième avec le troisième.

Le produit scalaire ne rend que très imparfaitement **l'inclination** d'un élément par rapport à l'autre. En effet, l'inclination comprend *l'inclinaison*, mais indique jusqu'à quel point un vecteur "s'assimile" à un autre. Aussi le produit scalaire sera completé par une mesure de *similitude* tenant compte de l'importance relative de chaque idée à l'intérieur de chaque vecteur.

1.5.2 Similitude entre vecteurs d'inclinaison proche, ou mesure d'inclination

Le calcul de la similitude s'effectue sur une largeur donnée. On associe un *vecteur d'indices* à chaque vecteur opérande. Ce vecteur est trié de façon que sa lecture donne un ordre décroissant des composantes du vecteur auquel il est associé. Les composantes du vecteur pour lesquelles l'indice ne se trouve pas dans les premiers éléments du vecteur d'indices sont annulées. Une fois cette opération terminée le nouveau vecteur est renormé. Ensuite la valeur de la similitude correspond à la somme des produits des composantes pondérées par l'écart relatif existant dans les vecteurs d'indices.



$$\alpha_5 = V_1[5] \times V_2[5] \times \frac{1}{1 + \beta \times (1-6) \times (1-6)}$$

$$Sim(V_1, V_2) = \sum \alpha_i$$

Nous avons bien évidemment pour tout vecteur $\vec{V}$ non nul $sim(\vec{V}, \vec{V}) = 1$ et du fait que toutes les composantes sont positives ou nulles :

1.5.3 Propriétés de la similitude

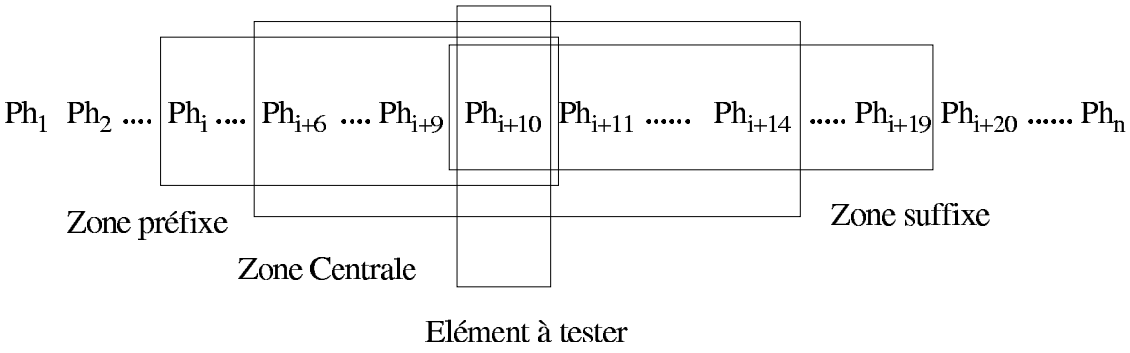
- pour tous vecteurs $\vec{V_1}$ et$\vec{V_2}$ orthogonaux : $sim(\vec{V_1}.\vec{V_2}) = 0$.

- la similitude est également symétrique :

$$sim(\vec{V_1}.\vec{V_2}) = sim(\vec{V_2}.\vec{V_1}) \quad (6)$$

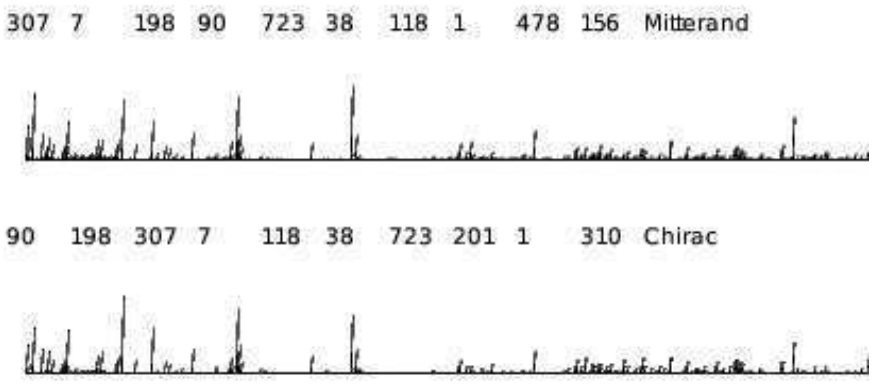
2 Fouille de texte

Le calcul des vecteurs sémantiques s'effectue sur chaque phrase du texte. Le principe de décision pour la sélection d'une phrase est construit sur le calcul moyen des vecteurs associés aux phrase situées à l'intérieur d'une fenêtre. Pour une décision à propos de la phrase Ph_{i+10}les vecteurs concernés seront les centroïdes des trois zones *préfixe,centrale* et *suffixe* telles que définies ci-après.



Le classement des phrases s'effectue par comparaison des différents vecteurs avec des vecteurs spécifiques $\overrightarrow{V_{Chirac}}$et $\overrightarrow{V_{Mitterand}}$, que nous symboliserons par $\vec{V_C}$ et $\vec{V_M}$ respectivement.

> Ces deux vecteurs ont été obtenus en calculant le centroïde des vecteurs des phrases associées à chacun d'eux dans le corpus d'apprentissage. Pour le calcul de ce vecteur, chaque vecteur est affecté d'un poids proportionnel à sa taille (le coefficient utilisé est égal au millième du carré de la longueur en octets).



> Les indices correspondent aux concepts majoritaires dans l'ordre décroisant.

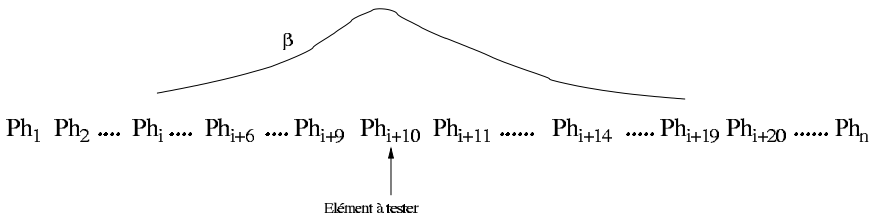
Soient :

- $\vec{V_{ZP}}$ le vecteur de la zone préfixe, que nous représenterons par $\vec{V_{ZP}}$
- $\vec{V_{ZC}}$ le vecteur de la zone centrale, que nous représenterons par $\vec{V_{ZC}}$
- $\vec{V_{ZS}}$ le vecteur de la zone suffixe, que nous représenterons par $\vec{V_{ZS}}$

Le premier filtre compare les produits scalaires $<\vec{V_{ZP}}.\vec{V_C}>$ et $<\vec{V_{ZP}}.\vec{V_M}>$, $<\vec{V_{ZC}}.\vec{V_C}>$ et $<\vec{V_{ZC}}.\vec{V_M}>$ et $<\vec{V_{ZS}}.\vec{V_C}>$ et $<\vec{V_{ZS}}.\vec{V_M}>$.

> Pour qu'une phrase soit candidate au classement comme phrase appartenant au discours de Mitterand il est nécessaire qu'au moins un produit scalaire ($<\vec{V_{ZP}}.\vec{V_M}>$, $<\vec{V_{ZC}}.\vec{V_M}>$ ou $<\vec{V_{ZS}}.\vec{V_M}>$) soit supérieur à son correspondant ($<\vec{V_{ZP}}.\vec{V_C}>$, $<\vec{V_{ZC}}.\vec{V_C}>$ ou $<\vec{V_{ZS}}.\vec{V_C}>$).

Dans le cas ou une phrase est candidate au classement un score d'appartenance est évalué. Ce score correspond au nombre de produits scalaires $<V_{Ph_i}.\vec{V_M}>$ supérieur aux produits scalaires $<V_{Ph_i}.\vec{V_C}>$. Dans le calcul du score on fait intervenir un coefficient indiquant la proximité avec la phrase candidate :



Si le score atteint un certain seuil (dépendant de β) et que deux produits scalaires au moins ($<\vec{V_{ZP}}.\vec{V_M}>$, $<\vec{V_{ZC}}.\vec{V_M}>$ ou $<\vec{V_{ZS}}.\vec{V_M}>$) dont le central sont supérieur aux produits scalaires correspondants ($<\vec{V_{ZP}}.\vec{V_C}>$, $<\vec{V_{ZC}}.\vec{V_C}>$ ou $<\vec{V_{ZS}}.\vec{V_C}>$) on effectue un calcul de similitude :

> Le calcul de similitude compare seulement cinq phrases : les deux précédentes, la phrase sélectionnée et les deux suivantes :
>
> pour la phrase sélectionnée :

$$\text{Sim}(\vec{V_M}, V_{Ph_{i+10}}) \text{ et } \text{Sim}(\vec{V_C}, V_{Ph_{i+10}})$$

Si la similitude de la phrase testée par rapport au vecteur représentant le discours de Mitterand est supérieure à la similitude du vecteur testé par rapport au vecteur représentant le discours de Chirac et qu'il en va de même soit pour les deux phrases précédentes soit pour les deux phrases suivantes la phrase testée est attribuée au discours de Mitterand (Nous supposons donc que les parties insérées du discours de Mitterand comportent au moins quatre phrases consécutives).

La recherche des phrases associées au discours de Mitterand se termine par un petit correctif éventuel pour tenir compte de la propriété : *Il n'y a pas de phrase isolée associée au discours de Mitterand.* Ansi si l'on a une configuration comme CCMCC où C désigne une phrase associée au discours de Chirac et M une phrase associée au discours de Mitterand, la phrase associée au discours de Mitterand est désélectionnée. Il en va de même pour l'inverse. Une configuration comme MMCMM sélectionne la phrase centrale comme appartenant au discours de Mitterand.

3 Résultats sur le corpus d'apprentissage

Le corpus d'apprentissage comprenait 7523 phrases apprtenant au discours de Mitterand. L'extraction a donné 6479 textes dont 5782 correctement trouvés. Soit une précision de 0.89, un rappel de 0.76 et un Fscore de 0.82.

4 Résultats sur le corpus de test

Sur le corpus de test le rappel s'est effondré : 0.15. La précision a faiblie dans une moindre proportion : 0.77. Une partie de ce phénomène vient du fait que la tâche demandée était plus axée sur le rhétorique ou la distribution du vocabulaire que sur la sémantique. Comme cette méthode essaie au contraire de s'affranchir de la rhétorique et du choix du vocabulaire le résultat va de soi. Nous pouvons illustrer ce phénomène dès le départ :

1ere phrase trouvée et suivantes:

> La quantité elle est facilement fournie, car la qualité c'est de l'exigence, et l'exigence par rapport à soi-même. Il n'y a pas de haute société, de progrès, il n'y a pas de grands pays qui n'ait à son propre égard une forte exigence. On ne fait rien dans la mollesse, dans la faiblesse, et en tout cas certainement pas dans l'ignorance ! Les gouvernements antérieurs ont signé ce pacte qu'on appelle par ses initiales - GATT - cet accord commercial qui implique que les produits de substitution américains pénètrent à l'intérieur de l'Europe sans subir de taxe. Ils viennent donc concurrencer indûment nos propres productions. Si l'on ajoute à cela les montants compensatoires monétaires qui font que des pays comme la Hollande et l'Allemagne peuvent vendre en France des produits moins chers, alors que faire ? Et je vois des foules paysannes dressées pour dénoncer, pour maudire, les produits de substitution américains. Vous êtes ici affrontés, dans des conditions plus rigoureuses qu'ailleurs, à tous les problèmes urbains des grandes concentrations humaines. Il y a votre population, celle qui figure sur les documents officiels, 44000 habitants peut-être, puis il y a les autres, souvent les laissés-pour-compte ou les migrants, qui vont d'une ville à l'autre, ne sachant où se loger, où s'arrêter. Normalement attirés par ces lieux où l'on peut se perdre, tant et tant d'hommes cherchent à être reconnus, tant d'autres cherchent à ne pas l'être. Il vous faut assurer la synthèse de ces aspirations, de ces besoins, de ces moyens. Croyez-moi, la France c'est comme cela. La France c'est le pays que l'on sait, on connaît sa beauté, ses attraits, on connaît aussi ses misères. Depuis toujours il a été composé d'alluvions venues d'un peu partout au gré des combats, des conquêtes ou bien de leur reflux, au gré aussi des aventures humaines qui conduisent plus naturellement les hommes à venir là où l'on se sent bien plutôt qu'ailleurs et on se sent généralement bien en France à condition bien entendu que la France sache recevoir et accueillir, qu'elle s'ouvre plutôt que de se fermer. Elle a reçu d'immenses bienfaits. Il m'est agréable de souligner que, grâce à la haute conscience que vous avez de vos devoirs, le Brésil peut vivre dans un Etat de droit. Vous avez derrière vous, un siècle et demi d'existence et vous avez pu dans ce temps-là élargir les compétences, adapter la jurisprudence à

> l'évolution du droit public et de la société. Bref, vous constituez en quelque sorte l'organe régulateur de la machine complexe de l'Etat.

phrase du discours de Mitterand les précédant :

> Je ne m'attarderai pas sur ce problème, mais je dirai un mot quand même de la Communauté européenne dans un instant. Comme vous savez, après avoir été visité cette exploitation, je me suis rendu aux Haras d'Aurillac. J'y ai rencontré les organisations régionales agricoles. Je me suis exprimé selon ma conviction et dit ce que je pensais de l'avenir de l'agriculture française, et auvergnate en particulier, en insistant sur le fait que pour assurer les mutations essentielles et préserver la compétitivité française, il fallait accepter un certain nombre de risques. La Communauté européenne, j'en suis tout à fait partisan. J'ai voté tous les accords européens. En 1957, j'étais un partisan fervent du Marché commun agricole. C'était une bataille difficile et je suis extrêmement touché de voir aujourd'hui avec quel empressement ceux qui ne l'ont pas voté me reprocheraient de ne pas suffisamment réussir l'-entreprise dont ils ne voulaient pas. Ce qui prouve qu'on peut changer d'opinion, ce qui est honorable, et même de devenir, c'est généralement le caractère des néophytes, des zélateurs enthousiastes, même à retardement. Cette Communauté des Dix doit devenir, je l'espère, et je ne négligerai rien pour cela après avoir pris les précautions élémentaires pour un marché loyal, la Communauté des Douze. Elle a décidé en effet de limiter la production laitière. Je pose aux agriculteurs et j'ai posé partout, y compris à ses dirigeants nationaux, la question suivante : est-ce que vous voulez de l'Europe ou est-ce que vous ne la voulez pas ? Si vous ne la voulez pas, vous êtes logiques. Si vous en voulez, on est Dix. On est Dix, il faut l'accord des Dix. Et oui, parce qu'il y a eu une évolution de la Communauté, une mauvaise évolution. On s'est éloigné des dispositions du Traité de Rome qui fixaient la possibilité de voter à l'unanimité dans des cas très stricts de préservation des souverainetés nationales. Et c'est à la demande de la France dans les années 60, que cette loi a été rompue et que l'on a adopté un compromis, dit compromis de Luxembourg qui, en fait, contraint à l'unanimité dans les dispositions qui ne le méritent pas, ce qui bloque le système. A la diète de Pologne, rappelez-vous, il fallait que tous les députés de Pologne votassent à l'unanimité les lois. Vous imaginez si cela était comme cela en France ! Heureusement qu'il y a une majorité ! Mais en demander davantage donnerait vraiment le champ un peu trop libre à ceux qui n'aiment pas les majorités. L'Europe a fait cela à la demande de la France. La France a eu beaucoup d'initiatives dans cette Europe. C'est très bien. J'ai déjà eu le plaisir de vous accueillir dans le passé, pas forcément les mêmes, mais un certain nombre d'entre vous, et je suis très heureux de l'occasion qui m'est donnée de vous recevoir à nouveau et donc de revoir certains d'entre vous et de faire connaissance des autres. Cela a une très grande force symbolique, ce titre de "meilleur ouvrier de France". Il évoque bien des valeurs fondamentales, et d'abord l'amour du métier ; ensuite un grand savoir-faire, une capacité d'expression, j'ai dit tout à l'heure de beauté, d'esthétique, la maîtrise de l'outil et la perfection technique qui ne peuvent se passer de la maîtrise de l'esprit. C'est aussi un concours qui récompense des femmes et des hommes qui travaillent souvent dans des techniques de pointe, dans les métiers nouveaux. Il n'y a pas que les métiers traditionnels, bien qu'il faille aussi les honorer, mais il faut suivre l'évolution de la technique, l'évolution des

> temps ; il faut que la France dispose des meilleurs ouvriers possibles dans tous les domaines : ceux qu'on a coutume de connaître à travers les générations et ceux qui se révèlent comme des techniques et savoir-faire indispensables avec l'évolution de la technologie. Je félicite donc les lauréats, je ne pourrai tous les connaître, mais je suis heureux et flatté qu'ils soient aujourd'hui les hôtes de la Présidence de la République et je veux qu'ils rapportent chez eux, quand ils rentreront à la maison, dans leur famille, le sentiment d'avoir reçu le juste prix qui les honore et nous honore. Vous savez on ne se passera jamais de la qualité. La quantité elle est facilement fournie, car la qualité c'est de l'exigence, et l'exigence par rapport à soi-même.

Conclusion:

> Comme on peut le constater il s'agit avant tout de politique européenne. Par rapport à la séparation de thème politique intérieure / politique étrangère cette partie se situerait plutôt du coté de la politique intérieure.

Mais, dans le corpus d'apprentissage nous trouvons entre autre :

> <107:13:C> Vous avez également consacré beaucoup de temps et d'énergie à la construction de l'Europe de la Défense.
> <107:14:C> C'était et c'est encore un défi essentiel pour l'avenir de notre continent.
> <107:15:C> Cette Europe de la Défense, vous l'avez aidée à naître en surmontant tous les obstacles, en déployant des efforts constants.
> <107:16:C> Vous avez accompagné avec confiance et avec foi l'élan de Saint-Malo.
> <107:17:C> Vous avez su tirer pour nos armées toutes les conséquences du Conseil européen de Nice.
> <107:18:C> Si l'Europe de la Défense est désormais plus qu'une espérance, si elle est aujourd'hui une réalité qui s'enracine, c'est en partie à vous que la France et ses partenaires le doivent.
> <107:19:C> Il me revient enfin, en tant que chef des armées, de rendre hommage à vos mérites personnels.
> <107:20:C> Votre sens du devoir, votre franchise, votre fidélité sans faille à la mission sont d'abord des qualités de soldat.
> <107:21:C> Votre souci constant de la reconnaissance par la Nation du dévouement de nos armées et votre attachement à l'amélioration de la condition militaire vous ont valu le respect de tous.

Donc le fait de retrouver ces phrases associées à la politique européenne associées au discours de Jacques Chirac implique que cette partie ne pas doit être associée au discours de François Mitterand. Une lecture aléatoire de parties du corpus d'apprentissage montre que ce phénomène se reproduit de temps en temps. Les vecteurs de références s'en trouvent nécessairement affectés. Comme dans le corpus d'apprentissage le nombre de phrases associées aux discours de Jacques Chirac est beaucoup plus important, il est normal que les phrases traitant de la politique européenne se trouvent associées à ces discours. Bien sûr il existe peut-être d'autres facteurs que nous essaierons de dégager. La méthode présentée sera mieu adaptée à un traitement de classification en fonction du thème traité.

Références

Chauché J. (1990), Détermination sémantique en analyse structurelle : une expérience basée sur une définition de distance. *TA Information* vol 1/1, p 17-24.

Crestan E. , El-Bèze M. , de Loupy Claude (2003). Peut-on trouver la taille de contexte optimale en désambiguisation sémantique ? *Actes de TALN2003*11-14 juin, Batz-sur-Mer. Vol 1. Pp 85-94.

Ellman J., Tait, J. (1999) Roget's thesaurus: An additional Knowledge Source for Textual CBR? *Proc. of 19th SGES Int. Conf. on Knowledge-Based and Applied AI*. Springer-Verlag.pp 204 Ð 217.

Larousse.(1992) *Thésaurus Larousse - des idées aux mots, des mots aux idées.* Paris.

Roget P.(1852) *Thesaurus of English Words and Phrases* Longman, London.

Yarowsky D. (1992) Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. *Proc. of COLING92.*

Peut-on rendre automatiquement à César ce qui lui appartient ? Application au jeu du Chirand-Mitterrac

Marc El-Bèze, Juan-Manuel Torres-Moreno, Frédéric Béchet

LIA – Université d'Avignon et des Pays de Vaucluse
BP 1 228, F-84 911 Avignon Cedex 09
{ marc.elbeze, juan-manuel.torres, frederic.bechet }@univ-avignon.fr

Mots-clés : Segmentation, Catégorisation thématique, Adaptation, Cohésion

Keywords: Segmentation, Topic Classification, Adaptation, Cohesion

Résumé Nous présentons des modèles d'apprentissage probabilistes appliqués à la tâche de classification binaire telle que définie dans le cadre du défi DEFT05[1]. Au sein de discours de Jacques Chirac, a pu être insérée une séquence de phrases de François Mitterrand. Pour identifier la paternité de ces séquences, nous avons utilisé des chaînes de Markov, des modèles bayesiens, et des procédures d'adaptation de ces modèles. Une comparaison avec diverses approches montre la supériorité des méthodes que nous proposons. Les résultats que nous obtenons, en termes de précision, rappel et Fscore sur le sous corpus de test Mitterrand sont très encourageants.

Abstract We present probabilistic learning models applied to the task of binary classification as defined in the DEFT05[1] challenge (a sequence of François Mitterrand's sentences could have been inserted into a speech of Jacques Chirac). Markov chains, Bayes models and an adaptative process have been used. A comparison with a baseline and perceptron approaches shows the superiority of our methods. In terms of precision, recall and Fscore over the Mitterrand test sub-corpus our results are very promising.

[1] http://www.lri.fr/ia/fdt/DEFT05

1 Introduction

A priori, un travail de classification à 2 classes (ici Chirac et Mitterrand[2]) paraît simple. Or, de nombreuses raisons font que le problème est complexe. Au terme d'une étude portant sur 68 interventions télévisées composées de 305 124 mots, (Labbe, 1990) distingue 4 périodes dans les discours de Mitterrand. L'une d'elles dénommée « *le président et le premier ministre* » (octobre 1986-mars 1988) n'est probablement pas la plus facile à traiter sous l'angle de vue particulier proposé par le défi DEFT05. Dans d'autres conditions, c'eût été loin d'être évident. Ici, on peut s'attendre à des difficultés accrues pour différencier deux orateurs qui se sont exprimés dans maints débats sur les mêmes sujets. Facteur aggravant : on ne dispose que d'un petit corpus déséquilibré : 109 279 mots pleins pour l'un et 582 595 pour le second répartis dans 587 discours (dont la date n'est pas fournie). Notons qu'une classification supervisée binaire avec un perceptron optimal à recuit simulé (TORRES *et al.*, 2002) appliqué sur la catégorie grammaticale de mots (l'utilisation de tous les mots générant une matrice trop volumineuse) donne un taux d'extraction des segments Mitterand décevant Fscore ≈ 0.43 ; la méthode[3] K-means sur les mêmes données conduit à un Fscore ≈ 0.4. Avec des classifieurs à large marge (AdaBoost avec BoosTexter et SVM avec SVM-Torch), on plafonne à 0.5.

Dans cet article, nous décrivons quelques méthodes employées dans le cadre de ce défi. Nous présentons en section 2 une première approche reposant sur des modèles bayésiens, une chaîne de Markov, des adaptations statiques et dynamiques et un réseau sémantique de noms propres. En section 3, nous présentons une deuxième approche probabiliste ne faisant appel à aucun filtrage ou lemmatisation, combinée avec un automate légèrement différent. Des expériences et résultats sont présentés en section 4. Une méthode de fusion de plusieurs approches y est esquissée avant de conclure et d'envisager quelques perspectives.

2 Modélisation I

La chaîne de traitement que nous allons décrire dans les sous-sections suivantes est constituée de 4 composants dont un seulement est totalement dédié à la tâche de DEFT05. On pourrait facilement le modifier ou au pire s'en passer, s'il fallait changer de domaine d'application, Sur un Pentium portable cadencé à 1,7 GHz et doté d'une RAM de 384Mo, l'intégralité de la chaîne s'exécute en 20' qui se décomposent en 5' pour l'apprentissage, et 15' pour le test soit une minute par itération du couple adaptation – étiquetage.

2.1 Modèles Bayésiens

Guidée par une certaine intuition que nous avions des caractéristiques de la langue et du style de chacun des deux orateurs, une analyse des données d'apprentissage nous a poussés à retenir

---

[2] Pour des facilités d'écriture, nous prenons la liberté de désigner les deux derniers présidents de la République, par leur nom de famille, sans les faire précéder d'un titre, ou d'un prénom, et pour plus de concision, il nous arrivera de nous contenter de remplacer Mitterrand et Chirac par les étiquettes *M* et *C*.

[3] En comparaison, avec une méthode de type *base-line (random)* où avec une probabilité de 0.80 pour la classe *C* et 0.20 pour *M*, on assigne la classe d'une phrase du test, on obtient un Fscore ≈ 0.22 .

certaines de leurs caractéristiques plutôt que d'autres. En premier lieu, il était naturel de tabler sur une caractérisation s'appuyant sur les différences de vocabulaire. Des études anciennes comme celles de (COTTERET & MOREAU, 1969) sur le vocabulaire du Général de Gaulle, ou d'autres plus récentes (LABBE, 1990) partent du même présupposé. Pour plusieurs raisons, cette approche est incontournable mais comme on en rencontre tôt ou tard les limites, on est amené naturellement à ne pas s'en contenter. En effet, la couverture des thématiques abordées par les différents présidents est très large. Les trajets politiques de deux présidents consécutifs se recoupent forcément. En conséquence, on observe de nombreux points communs dans leurs interventions, recouvrements auxquels viennent s'ajouter les reproductions conscientes ou inconscientes (citations ou effets de mimétisme).

Pour diversifier les points d'appuis, nous en avons testé d'autres comme la longueur des phrases (LL), le pourcentage de conjonction de subordination (Pcos), d'adverbes (Padv) ou d'adjectifs (Padj). Cinq de ces variables (Pcos, Padv, Padj, LL , et Plm) ont été modélisées par des gaussiennes dont les paramètres ont été estimés sur le seul corpus d'apprentissage. En ce qui concerne, le vocabulaire lui-même, qu'il s'agisse de lemmes ou de mots, nous avons entraîné sur ce même corpus des modèles *n*-grammes et *n*-lemmes (P#M et P#L), avec $n < 3$.

$$P(t) = \lambda_0 \times p_0(t) + (1-\lambda_0)\sum_{i=1}^{n} \lambda_i \times p_i(t) \qquad avec \ \sum_{i=1}^{n} \lambda_i = 1 \quad (1)$$

Les valeurs des coefficients λ_i que nous avons attribuées de façon empirique à chacune de ces 9 variables figurent dans le tableau 1 ci-dessous.

	P1L	P1M	Padj	LL	P2L	P2M	Pcos	Plm	Padv
λ_i	0.39	0.15	0.15	0.14	0.05	0.04	0.05	0.02	0.01

Tableau 1 : Caractères employés pour la modélisation bayésienne et coefficients associés

Lorsqu'on utilise des chaînes de Markov en TALN, on est toujours confronté au problème de la couverture des modèles. Le taux de couverture décroît quand augmente l'ordre du modèle. Le problème est bien connu et des solutions de type lissage ou Back-off (MANNING & SCHÜTZE, 2000) sont une réponse classique au fait que le corpus d'apprentissage ne suffit pas à garantir une estimation fiable des probabilités. Le problème devenant critique lorsqu'il y a un déséquilibre flagrant entre les deux classes, il nous a semblé inutile, voire contre-productif de calculer des trigrammes sur le sous corpus *M*.

En nous inspirant des travaux menés en lexicologie sur les discours de Mitterrand, nous avons essayé de prendre en compte certains des traits qualifiés de dominants chez Mitterrand par (ILLOUZ *et al.,* 2000) : adverbe négatif, pronom personnel première personne singulier, point d'interrogation, ou des expressions comme *c'est, il y a, on peut, il faut* (dans les 4 cas, à l'indicatif présent). Après vérification de la validité statistique de ces traits sur le corpus DEFT05, nous les avons intégrés dans la modélisation mais dans un second temps, nous les avons retirés car même s'ils entraînaient une légère amélioration sur les données de développement, rien ne garantissait qu'il ne s'agissait pas, là, de tics de langage liés à une période potentiellement différente de celle du corpus de test. Par ailleurs, en cas de portage de l'application à un autre domaine ou une autre langue, nous ne voulions pas être dépendants d'études lourdes. En tous les cas, nous avons préféré faire confiance aux modèles de Markov pour capturer automatiquement une grande partie de ces tournures.

2.2 Prise en compte de contraintes générales

Un discours de Chirac peut avoir fait l'objet de l'insertion d'au plus une séquence de phrases. La séquence *M*, si elle existe, est d'une longueur supérieure ou égale à deux. Pour prendre en compte cette contrainte particulière, nous avions, initialement, pensé écrire des règles, même si une telle façon de faire s'accorde généralement peu avec les méthodes probabilistes. Dans le cas présent, que faut-il faire si une phrase détachée de la séquence *M* a été étiquetée *M*, avec une probabilité plus ou moins élevée (certainement au dessus de 0.5, sinon elle aurait reçu l'étiquette *C*) ? Renverser la décision, ou la maintenir ? Si l'on opte pour la seconde solution, il serait logique d'extraire également toutes les phrases qui la séparent de la séquence *M*, bien qu'elles aient été étiquetées *C*. Ne pas se contenter du *statut quo* comporte un piège : un gain aléatoire en rappel risque de se faire au prix d'une chute de précision.

Pour pouvoir trouver, parmi les chemins allant du début à la fin du discours, celui qui optimise la production globale du discours, nous avons exploité un automate probabiliste à cinq états (dont un initial *I* et 3 terminaux, C_1, C_2, et M_2). Comme on peut le voir sur la figure 1, vers les états dénommés C_1 et C_2 (resp. M_1 et M_2), n'aboutissent que des transitions étiquetées *C* (resp. *M*). À une transition étiquetée *C* (resp. *M*), est associée la probabilité d'émission combinant pour *C* (resp. *M*) les modèles probabilistes définis en section 2.1.

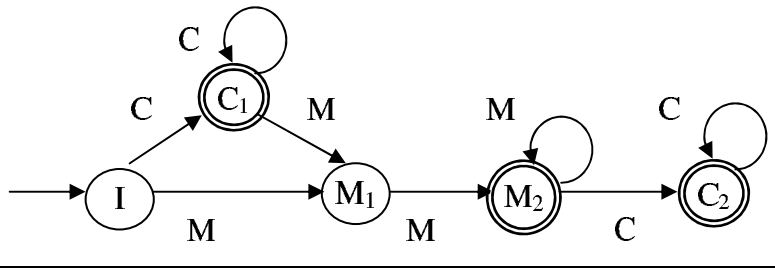


Figure 1 : Machine de Markov exprimant les contraintes générales des trois tâches

Avant de décrire les étapes ultérieures du processus de catégorisation segmentation, notons que c'est ce composant qui a permis de faire un saut conséquent (plus d'une dizaine de points) au niveau des performances et a ouvert ainsi la voie à la mise en place de procédures d'adaptation décrites en section suivante. Remarquons par ailleurs que la question aurait pu être gérée autrement, par exemple en utilisant, pour chaque discours, la partie triangulaire d'une matrice carrée *M*[*n*,*n*] (*n* étant le nombre de phrases contenues dans le discours en question). Dans chaque case *M*[*i*,*j*], on calcule la probabilité que la séquence soit étiquetée *M* entre *i* et *j*, et *C* du début jusque *i-1* et de *j+1* à *n*. Déterminer les bornes optimales de la séquence Mitterrand revient alors à rechercher un maximum sur toutes les valeurs *M*[*i*,*j*] telles que $i > j$. Si celle valeur optimale est inférieure à celle qu'on aurait obtenue en produisant toute la chaîne avec le modèle associé à Chirac, on se doit de supprimer la séquence *M*.

La complexité de cette seconde méthode est supérieure à celle de l'algorithme de Viterbi que nous avons employé. Il nous a paru néanmoins intéressant d'en faire état car elle offre la possibilité de combiner aisément des contraintes globales plus élaborées que celles que nous avons à prendre en compte. Elle peut aussi permettre de mixer des modèles issus de l'apprentissage et d'autres optimisant des variables dédiées à la modélisation de la cohésion interne des séquences qui se trouvent dans le discours traité, et n'ont fait l'objet d'aucun apprentissage préalable.

2.3 Adaptation statique et dynamique

Durant cette étape, ont été mises en œuvre des procédures d'adaptation statique et dynamique qui permettent de gagner entre 3 et 4 points de Fscore. La contrainte de ne pouvoir enrichir le corpus d'apprentissage, sous peine de disqualification, nous a poussé à tirer un parti intégral des données mises à notre disposition. Or, en dehors du corpus d'apprentissage, ne restait plus que les données de test. C'est sur elles, que l'adaptation a été pratiquée. Dériver un modèle à partir de l'intégralité des données de test correspond à ce que nous appelons ici adaptation statique. L'adaptation dynamique, quant à elle, repose sur un modèle découlant seulement du discours en train d'être testé. Bien entendu, il n'est pas interdit de conjuguer les 2 approches.

Dans un premier temps, nous avions envisagé de pratiquer un étiquetage des données de test, l'objectif étant à l'itération $i+1$ de n'adjoindre au corpus d'apprentissage[4] de X que les phrases s ayant reçu au pas i une probabilité $P_i(X|s)$ supérieure à un certain seuil $T_{X,i}$. Un apprentissage de type maximum de vraisemblance effectué sur les données ainsi collectées peut autant rapprocher qu'éloigner du point optimal. Pour pallier cette difficulté, nous avons opté pour un apprentissage EM, consistant à ne compter pour chaque couple {élément=e, X} observé dans les données d'adaptation que la fraction d'unité égale à la probabilité de l'orateur X sachant la phrase qui contient e. La prise de décision repose sur une formule analogue à celle de la formule (1). La variable en position 0 est la probabilité de l'étiquette sachant la phrase qui lui a été attribuée à l'itération i. Nous avons fait décroître le poids λ_0 qui lui est associé, de façon progressive, d'une itération à l'autre par pas de 0.1. Les 4 modèles employés sont, pour les 2 premiers, lemmes et mots issus de l'adaptation locale, pour les 2 derniers, lemmes et mots issus de l'adaptation globale. La pondération entre les différentes probabilités est restée la même durant toutes les itérations : { `0.9, 0.02, 0.003, 0.005` }.

2.4 Réseau de Noms Propres et Cohérence interne des discours

À partir de la tâche 2, l'ensemble des noms propres était dévoilé aux participants. Établir un lien entre différents éléments apparaissant dans des phrases même éloignées d'un discours donné, nous a paru être un bon moyen pour mettre en évidence une sorte de réseau sémantique permettant aux segments de s'auto regrouper autour d'un lieu, de personnes et de façon implicite d'une époque. Dans le cas de données bien séparables, plusieurs ensembles de noms ancrés dans une Histoire et une Géographie commune devraient former des composantes connexes (idéalement deux) sur lesquelles il suffirait ensuite de mettre l'étiquette M ou C. Bien que cela ne soit pas tout à fait la démarche que nous avons adoptée, ces remarques aident à en comprendre l'esprit.

1339 termes ont été regroupés dans 275 « concepts » qui pour épouser la richesse des discours traités dépassent largement un cadre restreint aux seules considérations géopolitiques (le Sport et la Culture sont souvent abordés lors de cérémonies de remises de médailles). Un terme peut se retrouver dans plusieurs classes, comme par exemple Miguel Angel Asturias, qui a été placé aussi bien dans la classe des guatémaltèques que dans celle des écrivains étrangers. Afin de mixer les relations entretenues entre les noms de pays, leurs habitants, les capitales, le

---

[4] X pouvant prendre ici les valeurs M ou C .

pouvoir exécutif, nous avons complété un réseau fourni par le Centre de Recherche de Xerox, en y rajoutant quelques relations issues des Bases de Connaissance que l'équipe TALN du LIA utilise pour faire fonctionner son système de Questions / Réponses (BELLOT *et al.*, 2003). Ci-dessous, figure un petit extrait de ce réseau non structuré :

ARGENTIN
Argentine Alfonsin Carlos_Menem Bioy_Casares Buenos_Aires Alfredo_Arias Jorge_Remes
MEXICAIN
Mexique Mexico Zedillo Zédillo Benito_Juarez Carlos_Fuentes Octavio_Paz FOX Fox Cancun Monterrey

Après 4 itérations, sur 57 301 phrases valides que comptait le corpus de développement (test : 27 120), 6 011 ont été regroupées en 906 groupes (test : 432 groupes de 2 907). Plus de 10% des segments se retrouvent donc dans des groupes, dont le cardinal moyen est d'environ 6,5. Le plus grand groupe contient 50 segments (test : 63). Seuls 16 groupes (test : 12) regroupent, de façon confuse, des étiquettes M et C. C'est le cas du discours 38, où la phrase 30 étiquetée M possède en commun Casablanca MAGHREB (en fait, il s'agissait du sommet de Casablanca) avec la phrase 173 étiquetée *C*, où Chirac fait état de ses récents voyages au Maroc. L'avantage d'un réseau probabiliste est que cette erreur n'est pas rédhibitoire. En effet, dans notre soumission, la phrase 30 a été correctement extraite et non la phrase 173. Cela ne fonctionne pas toujours aussi bien ! Dans le cas du discours 739, la séquence *C* et la séquence *M* ont en commun 2 « termes-concepts » (Espagne-Espagnol et Méditerranée-Méditerranéen). Il se trouve que la seconde confusion aurait pu être évitée si le *TGV Paris-Lyon-Méditerranée* dont parle Mitterrand n'avait pas fait l'objet d'une sur découpe au moment de la tokenisation. Mais cela n'aurait pas suffi, car avec l'aide de l'autre terme (*Espagne*) quatre phrases M 30, 35, 36 et 37 ont été regroupées par transitivité avec 12 phrases étiquetées *C* (1, 3, 6-17, 20-5, 27, 47). De fait, aucun segment du discours 739 n'a été extrait. Il est clair que nous sommes encore loin d'une représentation élaborée des relations entretenues entre des concepts et leur expression au travers de textes, mais le réseau que nous avons élaboré à peu de frais est un premier pas dans cette direction.

3 Modélisation II

Nous nous sommes demandé si la recherche des caractéristiques propres à un auteur pourrait être facilitée par le fait de ne pas filtrer ou éliminer quoi que ce soit des discours. Ainsi, nous avons fait l'hypothèse que l'utilisation répétée, voire exagérée de certains symboles de ponctuation ou l'emploi de termes ne servant qu'à assurer le bâti de la phrase, pouvait prétendre au statut d'indicateur fiable.

3.1 Modèle probabiliste

Pour ce deuxième modèle, nous sommes partis du principe que les techniques de *n*-grammes appliquées à des tâches de classification, pourraient se passer d'une phase préalable de lemmatisation ou de stemming, du rejet des mots-outils et de la ponctuation. Pour les systèmes *n*-grammes, (Jalam & Chauchat, 2002 ; Sahami, 1999) ont montré que les performances ne s'améliorent pas après stemming ou élimination des mots-outils. Dans cet esprit, nous avons laissé les textes dans leur état originel. Aucun prétraitement n'a été effectué, même si cette démarche a ses limites : par exemple, *Gasperi* et *Gaspéri* comptent pour des mots différents, qu'il y ait ou non erreur d'accent ; *premier* et *première* sont aussi comptabilisés séparément en absence de lemmatisation. Malgré cela, nous avons voulu donner

au modèle un maximum de chances de capturer des particularités de style (manies de ponctuation, sur ou sous emploi de subjonctifs, gérondifs, …) qui auraient été gommées après application de prétraitements comme la lemmatisation.

3.2 Adaptation naïve

Une idée naïve nous est venue à l'esprit. Les contraintes DEFT05 indiquent qu'il y a zéro ou au moins 2 phrases de Mitterrand insérées dans tout discours de Chirac. De ce fait, nous avons implanté la méthode simple d'absorption suivante : si une phrase *i* appartenant à la classe *M* est précédée et suivie des phrases *i*-1 et *i*+1 ∈ à la classe *C*, alors on transforme l'étiquette de la phrase *i* en *C*. Le cas opposé a été également pris en compte : des phrases de la classe *C* enveloppées par des phrases du type *M* seront donc absorbées comme *M*. Et cela pour tous les discours. Nous avons étiqueté la première et dernière phrase de chaque discours comme appartenant à la classe *C*. Même si elle est simple, cette méthode fait gagner entre 4 et 5 points de Fscore.

3.3 Adaptation par Viterbi

La méthode naïve fait progresser de quelques points, mais elle présente deux inconvénients majeurs : i) elle est trop dépendante des contraintes fixées par DEFT05, ii) elle a tendance à laisser, de façon indésirable, des îlots de la classe *M* au milieu des discours. Comment les éliminer ? Nous avons procédé de la même manière que dans la section 2.2, avec un automate légèrement différent. Nous avons donc construit l'automate probabiliste à cinq états représenté en figure 2, avec un état initial *I*, final *F* et 4 états intermédiaires C_1, M_1, M_2 et C_2. Comme dans le cas de l'automate de la figure 1, à une transition étiquetée *C* (resp. *M*), est associée la probabilité d'émission combinant pour *C* (resp. *M*) les modèles de *n*-grammes définis en section 3.1. Nous avons alors appliqué l'algorithme de Viterbi (MANNING & SCHÜTZE, 2000) pour trouver la séquence optimale. Nous obtenons, de cette façon, un Fscore de 0.818 sur l'ensemble de développement de la tâche 1.

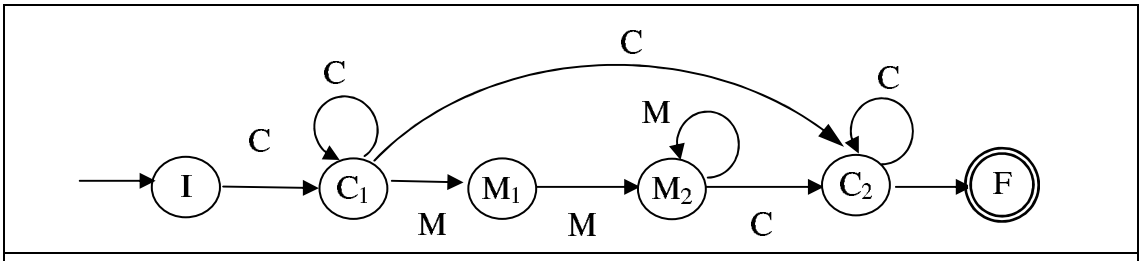


Figure 2 : Machine de Markov exprimant les contraintes générales des trois tâches

4 Expériences

Modélisation I : Tous les corpus (apprentissage et test) ont été traités par l'ensemble d'outils LIA_TAGG (www.lia.univ-avignon.fr), pour effectuer une tokenisation, un étiquetage morpho-syntaxique grâce au tagger dérivé du tagger ECSTA (SPRIET & EL-BÈZE, 1998). Dans la phase de développement, le corpus d'apprentissage a été découpé en 5 sous corpus de telle sorte que pour chacune des 5 partitions, un discours appartient dans son intégralité soit au test soit à l'apprentissage. À tour de rôle, chacun de ces sous corpus est considéré comme corpus

de test tandis que les 4 autres font office de corpus d'apprentissage. Cette répartition a été préférée à un tirage aléatoire des phrases tolérant le morcellement des discours. En effet, un tel tirage au sort présente deux inconvénients majeurs. Le premier provient du fait qu'un tirage aléatoire peut placer dans le corpus de test des segments très proches de segments voisins qui eux ont été placés dans le corpus d'apprentissage. Le second inconvénient (le plus gênant des deux), tient au fait qu'une telle découpe ne permet de respecter le schéma d'insertion défini dans les spécificités de DEFT05.

Modélisation II : Un autre protocole expérimental a été défini. Aucun filtrage, ni étiquetage syntaxique, ni lemmatisation. Nous avons créé des sous corpus d'apprentissage *A* et de développement *D* à partir du corpus d'apprentissage disponible (corpus.tache.learn), gardant la proportion de 80%-20% de discours respectivement. Nous avons éclaté le corpus d'apprentissage *A* en deux sous-ensembles, respectivement {*C*} et {*M*} contenant les phrases de Chirac ou celles de Mitterrand. Puis, nous avons construit les *n*-grammes (n = 1,2,3) de chaque sous-ensemble et nous avons calculé leur entropie moyenne. Nous avons obtenu $E(C) = 3.66$ et $E(M) = 3.84$. La classification des discours de l'ensemble de test *T* se fait comme suit : une phrase *i* d'un discours *j* est décomposée dans ses *n*-grammes, puis on calcule son entropie, celle de Chirac *EC* (sur les *n*-grammes de l'ensemble {*C*}) et celle de Mitterrand *EM* (sur les *n*-grammes de l'ensemble {*M*}). Nous avons défini un seuil $\delta = EC - EM$ et si la quantité $\delta < \varepsilon$ (avec ε suffisamment petit), la phrase sera attribuée à la classe Mitterrand, autrement à Chirac. Nous avons combiné par un lissage analogue à celui présenté en formule (1), avec λ_1=0.625, λ_2=0.166 et λ_3=0.208. Nous obtenons alors un Fscore de 0.83, 0.80.et 0.83 sur l'ensemble de développement des tâches 1, 2 et 3 respectivement.

4.1 Résultats

Pour alléger les graphiques, nous nous sommes limités au tracé de 4 courbes, par figure. Tant pour le Fscore que pour le rappel ou la précision, n'ont été retenus que les résultats obtenus sur les données de la tâche 2 (test et développement) dans 2 conditions expérimentales : avec ou sans les groupes de Noms Propres définis en section 2.4 (T ou D, avec ou sans Noms).

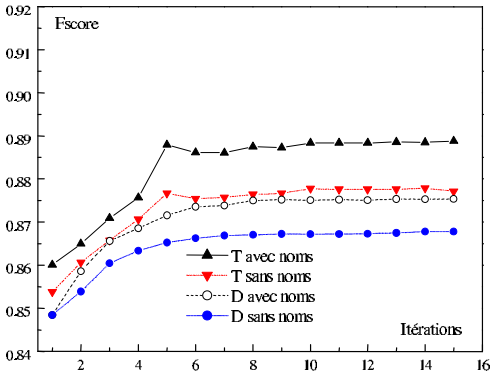


Figure 3 : Courbes de Fscore / tâche 2 / Modèle I / corpus de test (T) et développement (D)

Comme on peut le remarquer sur la figure 3, le Fscore s'améliore de façon notable au cours des 5 premières itérations. Au-delà, il n'y a pas à proprement parler de détérioration mais une stagnation qui peut être vue comme la captation par un maximum local. L'apport des réseaux bâtis autour des noms propres est indéniable, notamment c'est à eux que l'on doit le léger pic observé à la 5^{e} itération.

4.2 Analyse des résultats

Ainsi qu'on peut le remarquer dans le tableau 2, récapitulant les résultats officiels de notre équipe, le dévoilement des dates (tâche T3) permet d'améliorer très légèrement les résultats du modèle **II**, mais entraîne une dégradation sur le modèle **I**. Il est intéressant de voir comment se comportent les courbes de précision et de rappel, au fil des itérations. La figure 4 le montre sur la tâche 2 : sur *T* ainsi que sur *D*, c'est le gain en précision qui explique l'amélioration due aux Noms Propres. Ce gain allant de pair avec un rappel quasi identique (légèrement inférieur pour le test), il apparaît que le composant Noms Propres fonctionne comme un filtre prévenant quelques mauvaises extractions (mais pas toutes).

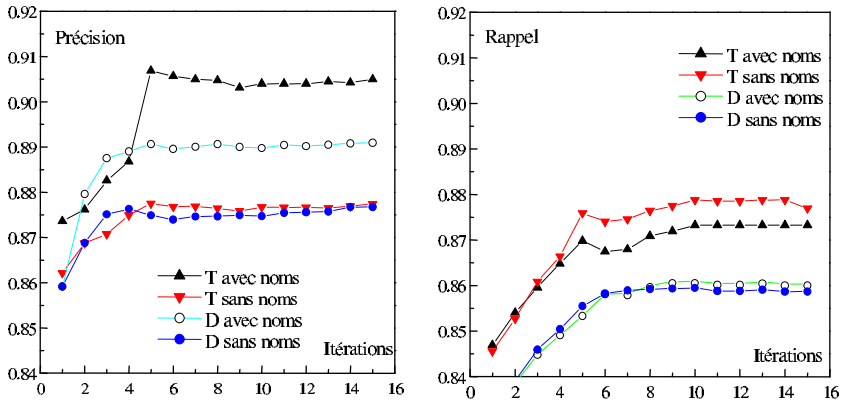


Figure 4 : Courbes de Rappel et Précision / tâche 2 / Modèle **I** / corpus *T* et *D*

4.3 Fusion de méthodes

Le dernier test que nous avons effectué (modèle **III**) repose sur une idée simple : accorder une confiance forte aux segments étiquetés de la même façon dans les différentes soumissions effectuées par les 2 équipes du LIA (junior et senior). Si cette méthode ne permet de gagner qu'entre 0 et 3 millièmes de point, il est clair qu'elle mériterait d'être appliquée sur des approches d'inspiration vraiment différente.

	Modèle I			Modèle II			Modèle III		
	Prec	Rappel	Fscore	Prec	Rappel	Fscore	Prec	Rappel	Fscore
T_1	0.881	0.854	0.867	0.826	0.764	0.794	0.883	0.858	0.870
T_2	0.909	0.861	**0.884**	0.827	0.775	0.8	0.911	0.858	**0.884**
T_3	0.887	0.872	0.879	0.829	0.776	0.801	0.89	0.871	0.880

Tableau 2 : Résultats officiels sur les 3 tâches {T_1, T_2, T_3} pour les trois soumissions

5 Conclusion et Perspectives

Les résultats que nous obtenons, en terme de Fscore (0.884) sont très encourageants**.** Ne pas lemmatiser et ne rien filtrer dégrade un peu les performances (Fscore≈0.84 avec le modèle **I**) mais permet de se passer d'un processus additionnel de prétraitement qui pour certaines langues peut être relativement lourd. Le recours à un réseau de Noms Propres est utile et nous encourage par la suite à employer une ressource lexicale comme EuroWordNet pour tirer parti

de réseaux sur les noms communs. Des frontières thématiques ne coïncident pas forcément avec des débuts de phrase. Les thèmes peuvent s'entremêler et composer un tissu discursif où les fils sont enchevêtrés de façon subtile. Beaucoup reste à faire pour pouvoir différencier plusieurs thèmes comme envisagé dans le cadre du Projet Carmel, plusieurs orateurs, ne serait-ce que trois. Et, si le lecteur veut se faire une petite idée de la difficulté de la tâche, nous l'invitons à deviner ce qui dans le présent article est dû à chacun de ses trois auteurs.

Remerciements

Nous remercions Eric Gaussier de Xerox d'avoir mis à notre disposition un lexique de Noms Propres. Nous sommes également reconnaissants envers Jérome Azé et Mathieu Roche du LRI qui n'ont pas ménagé leurs efforts pour organiser la campagne de DEFT05.

Références

BELLOT P., CRESTAN E., EL-BÈZE M., GILLARD L., DE LOUPY C. (2003), *Coupling Named Entity Recognition, Vector-Space Model and Knowledge Bases for TREC-11, Question-Answering Track*, actes de TREC'02, Gaithersburg, USA, NIST Special publication 500-251.

COTTERET J.-M., MOREAU R. (1969) *Le vocabulaire du Général de Gaulle,* Presses de la fondation nationale des sciences politiques, Armand Colin.

DAMASHEK M. (1995), *Gauging Similarity with N-Grams: Language-Independent Categorization of Text.* Science, 267 pp 843–848.

ILLOUZ G., HABERT B., FLEURY S., FOLCH H., HEIDEN S., LAFON P., PRÉVOST S. (2000), *Profilage de textes : cadre de travail et expérience*, Actes de JADT 2000, 5e Journées Internationales d'Analyse Statistiques des Données Textuelles, 9-11 Mars 2000, Lausanne.

JALAM R., CHAUCHAT J.-H. (2002), *Pourquoi les n-grammes permettent de classer des textes? Recherche de mots-clefs pertinents à l'aide des n-grammes caractéristiques,* actes de JADT, 6e Journées Internationales d'Analyse Statistiques des Données Textuelles, pp 13-15, St-Malo.

LABBE D. (1990) *Le vocabulaire de François Mitterrand,* Paris, Presses de la Fondation Nationale des Sciences Politiques, mars 1990.

MANNING C. D., SCHÜTZE H. (2000) *Foundations of Statistical Natural Language Processing,* The MIT Press.

SAHAMI M. (1999), *Using Machine Learning to Improve Information Access.* PhD thesis, Computer Science Department, Stanford University.

SPRIET T., EL-BEZE M., (1998) *Introduction of Rules into a Stochastic Approach for Language Modelling*, Computational Models of Speech Pattern Processing, NATO ASI Series F, vol. 169, ed. Keith Ponting, pp. 350-355.

TORRES J.M., AGUILAR J.C., GORDON M.B. (2002), *Finding the number minimum of errors in N-dimensional parity problem with a linear perceptron.* Neural Processing Letters, p 201-210.

Modèles de langage n-grammes et segmentation thématique pour une tâche de filtrage de textes

Martine Hurault-Plantet (1), Michèle Jardino (1), Gabriel Illouz (1)

LIMSI-CNRS
Bât.508, Université Paris XI, 91403, Orsay
{Martine.Hurault-Plantet, Michèle.Jardino, Gabriel.Illouz}@limsi.fr

Mots-clés : filtrage de textes, modèle de langage n-grammes, segmentation thématique
Keywords: text filtering, n-gram language model, topic segmentation

Résumé La tâche soumise à évaluation dans l'atelier DEFT'05 consistait à identifier les phrases issues d'allocutions de François Mitterrand, qui avaient été préalablement insérées dans un ensemble d'allocutions de Jacques Chirac. Dans chaque document, le thème des phrases insérées et le thème de l'allocution de Chirac dans laquelle elles s'insèrent, sont différents. Dans cet article, nous présentons les méthodes utilisées au LIMSI pour résoudre cette tâche. Nous avons expérimenté deux méthodes automatiques différentes que nous avons ensuite fait coopérer. L'une de ces deux méthodes s'appuie sur des modèles de langage n-grammes, l'autre est basée sur la segmentation thématique de l'allocution.

Abstract The task for the DEFT'05 evaluation workshop consists in identifying sentences selected from speeches by François Mitterrand, which had been inserted within a set of speeches by Jacques Chirac. The topic of the inserted sentences and the topic of the Chirac speech differ one from another in each document. This paper presents the two methods experimented at LIMSI in order to solve this task. The first one is based on n-gram language models and the second one is based on topic segmentation. Results of both methods are then merged.

1 Introduction

La tâche soumise à évaluation dans l'atelier DEFT'05 consistait à identifier les phrases issues d'allocutions de François Mitterrand, qui avaient été préalablement insérées dans un ensemble d'allocutions de Jacques Chirac. Dans chaque document, le thème des phrases insérées et le thème de l'allocution de Chirac dans laquelle elles s'insèrent, sont différents. Dans cet article, nous présentons les différentes méthodes utilisées au LIMSI pour résoudre cette tâche. Nous avons expérimenté deux méthodes automatiques différentes que nous avons ensuite fait coopérer. L'une de ces deux méthodes s'appuie sur des modèles de langage n-grammes (Jelinek, 1998), l'autre est basée sur la segmentation thématique de l'allocution. Le système

de coopération entre les deux méthodes utilise principalement l'intersection entre les résultats obtenus par chacune des deux méthodes.

Dans la première méthode, nous avons construit des modèles de langage n-grammes appris à partir des corpus d'apprentissage fournis, un pour Chirac et un pour Mitterrand. Chacun des modèles a ensuite été appliqué sur chacune des phrases à tester, l'auteur reconnu est celui dont le modèle donne la plus grande probabilité à la phrase.

La deuxième méthode est basée sur la segmentation thématique : le système détermine, pour chaque allocution, un ensemble de thèmes à partir des mots les plus fréquents. Pour chaque thème, le système calcule les intervalles de phrases de l'allocution dans lesquels on le trouve. Deux pôles thématiques sont ensuite sélectionnés parmi les thèmes : le thème le plus fréquent et son complémentaire, sur le critère de non recouvrement des intervalles associés à chaque thème. Chacun des autres thèmes est ensuite agrégé au pôle avec lequel il a un intervalle commun. Les thèmes qui ont un intervalle commun avec les deux pôles à la fois ne sont pas agrégés. Le système prend pour thème de Chirac celui des deux pôles dont les intervalles sont les plus proches du début et de la fin de l'allocution, et pour thème de Mitterrand le pôle complémentaire.

Dans la suite de l'article, nous présentons successivement les deux méthodes puis la méthode de fusion des résultats. Nous conclurons sur un bilan de ces méthodes et les expérimentations que nous envisageons pour améliorer nos résultats.

2 Les modèles de langage n-grammes

Ces modèles sont généralement employés dans les systèmes de reconnaissance de la parole (Jelinek, 1998). Ce sont des modèles probabilistes qui prédisent un mot connaissant les n-1 mots précédents. Nous les avons utilisés après avoir testé des représentations du type « sac de mots » qui ont donné des résultats peu satisfaisants.

2.1 Prétraitement des phrases et approche « sac de mots »

Nous avons effectué un premier prétraitement simple : nous avons conservé la ponctuation qui est un bon indicateur du style et transformé toutes les majuscules en minuscules.

Ensuite nous avons utilisé un algorithme de classification non supervisée (Jardino, 2000) pour partager toutes les phrases de la tâche 1 en deux sous-ensembles. L'espace de représentation est celui des mots indépendamment de leur ordre. Cette méthode appliquée à un corpus bien constitué (Brown Corpus) avait donné d'excellents résultats (Illouz et al., 2001). Sur le corpus Chirac-Mitterrand les résultats sont nettement moins bons. Le tableau 1 reporte les valeurs de précision et rappel des phrases de Mitterrand obtenues sur le corpus d'entraînement initial de la tâche 1, en utilisant différents ensembles de mots. Dans tous les cas la précision est très mauvaise. Les choix des fréquences permettent d'expérimenter différentes zones de la courbe de Zipf comme les zones de haute fréquence où se regroupent les mots-outils ou les zones de moyenne fréquence, censées représenter les mots sémantiquement significatifs. De manière caricaturale nous avons également expérimenté une représentation réduite au point et à la virgule.

Ensemble de mots	Rappel (%)	Précision (%)
Tous les mots	66	19
Mots de fréquence > 500	50	18
Point et Virgule	50	12
Mots de fréquence < 500	59	17
Mots tels que 10 < fréquence < 500	57	16

Tableau 1 : Rappel et précision des phrases de Mitterrand pour une partition non supervisée en deux classes des phrases de la tâche1

Les meilleurs résultats sont ceux obtenus avec tous les mots. En conséquence, nous les avons conservés et utilisés dans un modèle plus puissant prenant en compte l'ordre des mots dans la phrase.

2.2 Modèles de langage n-grammes pour identifier l'auteur de chaque phrase

Nous avons utilisé le logiciel de CMU (Clarkson, 1997) pour construire les modèles de langage n-grammes P_C et P_M à partir des comptes des successions de n mots respectivement dans les phrases de Chirac et dans les phrases de Mitterrand. Si une phrase de longueur l est représentée par la succession des mots : $m_1...m_i...m_l$, la probabilité de cette phrase calculée à partir d'un modèle n-grammes pour l'auteur A est :

$$P_A(phrase) = \prod_{i=1}^{i=l} p_A(m_i / m_{i-n+1}...m_{i-1})$$

Nous avons utilisé un lissage de type Witten-Bell pour calculer les probabilités des évènements non observés dans le corpus d'apprentissage. Ce type de lissage prend en compte le nombre de contextes dans lesquels ont été observés les n-grammes dans le corpus d'apprentissage. Pour choisir la valeur de n la mieux adaptée aux données, nous avons partitionné le corpus de la tâche1 de la façon suivante : 90% pour le corpus d'apprentissage des modèles Chirac et Mitterrand et 10% pour le corpus de test. Nous avons calculé pour chaque phrase du test les probabilités $P_C(phrase)$ et $P_M(phrase)$ et attribuée la phrase à l'auteur dont le modèle donne la plus grand probabilité. Nous avons ensuite évalué les valeurs de précision et de rappel pour les phrases de Mitterrand en comparant hypothèses et références, ceci pour des valeurs de n allant de 2 à 4.

Le tableau 2 montre que le modèle 3-grammes donne les meilleurs résultats sur nos jeux de données.

Modèles	Rappel (%)	Précision (%)
2-grammes	61	40
3-grammes	58	78
4-grammes	58	44

Tableau 2 : Rappel et précision des phrases de Mitterrand dans le corpus de test pour différents modèles de langage

2.3 Améliorations possibles

Une seule partition du corpus en apprentissage (90%) et test (10%) a été effectuée par manque de temps. D'autres tests ont été réalisés après l'expérimentation initiale, montrant des performances plus faibles en terme de précision pour les modèles 3-grammes. Les très bons résultats obtenus par ces derniers sont à relativiser par la sélection d'un corpus de test qui n'était pas en fait assez représentatif.

Les modèles de prédiction peuvent être améliorer par des méthodes de ré-échantillonnage en faisant varier phrases de test et d'apprentissage par validation croisée ou par la technique de Jackknife (Lebart et al., 2000).

2.4 Lissage pour construire des ensembles continus de phrases d'un même auteur

Notre méthode vote pour chaque phrase, ce qui induit des passages discontinus de phrases de François Mitterrand dans les phrases de Jacques Chirac. Pour pallier cet effet, nous avons utilisé un algorithme simple de lissage : chaque fois qu'un ensemble de 1 à k phrases de Jacques Chirac est détecté entre deux phrases de Mitterrand, ces phrases sont attribuées à François Mitterrand. Cette méthode donne un lissage assez fruste avec un taux de rappel très important. La valeur k = 4 a donné les meilleurs résultats sur le corpus de test.

3 Les pôles thématiques de l'allocution

La deuxième méthode présentée est basée sur une segmentation thématique de chaque allocution. Plus précisément, le système détermine d'abord l'ensemble des thèmes de l'allocution, puis, parmi ces thèmes, le thème dominant et le thème qui s'en éloigne le plus. Les méthodes décrites dans cette section utilisent un certain nombre de seuils qui ont été déterminés empiriquement sur le corpus d'apprentissage.

3.1 Détermination des thèmes

Pour déterminer les thèmes d'une allocution, nous effectuons d'abord un pré-traitement des phrases. Les mots de chaque phrase sont lemmatisés (Schmid, 1999), et nous ne conservons

pour indexer la phrase que les lemmes (ou le mot lorsqu'il est inconnu) des substantifs, des adjectifs, des noms propres et des abréviations. Nous utilisons également une liste classique de mots vides[1] à laquelle nous avons ajouté les treize mots les plus fréquents du vocabulaire du corpus d'apprentissage ainsi que les mots *monsieur* et *président*. Nous effectuons ensuite une reconnaissance automatique des adjectifs dont le substantif est dans l'allocution, et nous remplaçons ces adjectifs par les substantifs correspondants. Pour effectuer la reconnaissance automatique des adjectifs, le système utilise une liste de terminaisons d'adjectifs[2] qui lui permet de déterminer des adjectifs candidats et leurs racines respectives. Le système recherche ensuite les substantifs qui commencent par ces racines. Le bruit généré par cette méthode est limité par deux contraintes : la racine doit avoir une taille minimum (au moins deux caractères), et la taille de l'adjectif doit être supérieure à celle du substantif.

Pour renforcer les thèmes, le système utilise une méthode supplémentaire basée sur les séquences fréquentes maximales (Grahne et al., 2003). Le système recherche d'abord les séquences les plus fréquentes de mots dans une allocution, chaque phrase étant considérée comme une transaction. Le système génère ensuite la règle suivante d'équivalence entre les mots des séquences trouvées : le mot le plus fréquent de chaque ensemble est substituable à chacun des autres mots de la séquence. Le système ré-indexe alors chaque phrase de l'allocution suivant ces règles. Le thème le plus fréquent est donc renforcé par les thèmes qui sont en forte cooccurrence avec lui. Pour limiter le bruit produit, nous avons choisi un support élevé (la cinquième plus forte fréquence) pour générer les séquences fréquentes maximales.

Les thèmes finalement retenus par le système sont les mots les plus fréquents de l'allocution. Nous avons choisi un seuil égal à la vingt-cinquième plus forte fréquence, ou à défaut, un seuil de fréquence égal à 2. Par exemple, pour l'allocution 242 du corpus de test, nous obtenons les phrases indexées suivantes :

C 242:1 présidente flamme cas
C 242:2 émotion vie
C 242:3 présidente mental_handicap madame mental handicap accueil
C 242:4 combat obstacle digne
C 242:5 flamme droit
C 242:6 droit handicap
C 242:7 dignité droit
M 242:8 sujet général mondial attention négociation commerce débat raison
M 242:9 liberté général commerce échange
M 242:10 sujet rencontre mot
M 242:11 général obstacle accord
M 242:12 tiers accord
M 242:13 justice mot traitement
M 242:14
M 242:15 mondial clause accord
M 242:16 clause droit nation accord traité
M 242:17 cas messieurs mesdames

---

[1] Cette liste a été trouvée sur le Web

[2] Idem

M 242:18 traité
M 242:19 nation
M 242:20
M 242:21 justice négociation chemin force
M 242:22 débat
M 242:23 général accord
M 242:24 besoin
M 242:25 mondial
M 242:26
M 242:27 traitement
M 242:28 droit
M 242:29 vrai part
M 242:30 échange
C 242:31 dignité attention droit handicap
C 242:32 regard
C 242:33 handicap
…….

Nous avons trouvé une séquence maximale, *mental_handicap,* ainsi que deux correspondances adjectif-substantif, *handicapé-handicap* et *national-nation*. Le système associe ensuite à chaque thème les intervalles de phrases de l'allocution où il apparaît. Pour lisser les intervalles, le système calcule la densité moyenne du thème sur des groupes de cinq phrases, à partir de la première phrase[3]. Ainsi, si un même thème apparaît deux fois à une distance de trois phrases, le système considère qu'il apparaît dans un intervalle de cinq phrases qui comprend les deux phrases où il apparaît et les trois phrases où il n'apparaît pas. Les limites exactes de chaque intervalle sont ensuite déterminées par la recherche de la première et de la dernière phrase de l'intervalle où le thème apparaît.

Les thèmes les plus fréquents de l'allocution 242 et leurs intervalles sont les suivants (les phrases sont re-numérotées à partir de 0) :

12	handicap	2_5 30_32 41_68
7	dignité	6_6 30_37 47_61
6	droit	4_6 15_15 27_30
6	vie	1_1 36_36 48_64
5	accord	10_22
4	combat3_3 37_39	

3.2 Détermination des pôles thématiques

Le pôle dominant est le thème le plus fréquent. Le système cherche ensuite le pôle complémentaire. Deux méthodes ont été expérimentées pour cette recherche. Dans la première méthode, le système recherche le premier thème dont les intervalles n'ont aucun recouvrement avec les intervalles du pôle dominant. Si le système n'a pas trouvé de thème complémentaire, il applique une deuxième méthode qui consiste à rechercher le thème qui a le moins d'intervalles en commun avec le thème dominant. Pour cela, le système ré-indexe

[3] Pour les allocutions courtes (moins de 35 phrases), nous utilisons des groupes de trois phrases.

d'abord chaque thème par les thèmes qui possède un intervalle commun avec lui. Puis il recherche le thème qui a à la fois le moins de thèmes en commun et le plus de thèmes différents de ceux qui indexent le thème dominant.

Dans l'exemple de l'allocution 242, le pôle dominant est *handicap,* le pôle complémentaire est *accord*, premier thème dont l'intervalle n'a aucune intersection avec le pôle dominant.

3.3 Détermination des phrases de Mitterrand

Une fois les deux pôles thématiques trouvés, le système agrège chacun des autres thèmes avec le pôle avec qui il possède un intervalle en commun. Si un thème a un intervalle en commun avec les deux pôles, il est considéré comme un thème commun aux deux pôles et n'est pas agrégé.

Dans l'exemple de l'allocution 242, le pôle *handicap* est agrégé avec les thèmes *dignité*, *vie*, *combat*, *vrai*, *regard*, *besoin*. Le pôle *accord* est agrégé avec les thèmes *général*, *mondial*, *négociation*, *nation*, *justice*, *échange*. Les thèmes *droit*, *raison*, *obstacle*, *rencontre*, *liberté* sont des thèmes communs aux deux pôles, ils ne sont donc pas agrégés. Le système produit finalement les intervalles suivants :

Pôle *handicap* : intervalles 0_6 23_23 28_28 30_39 41_69
Pôle *accord* : intervalles 7_22 24_24 26_26 29_29

Le système attribue le locuteur Chirac à celui des deux pôles qui débute l'allocution et la termine. Il attribue le locuteur Mitterrand à l'autre pôle. Le système indexe ensuite les phrases des intervalles associés à chacun des deux pôles par leurs interlocuteurs respectifs, à l'exception des intervalles ne comportant qu'une seule phrase. Nous estimons en effet que, si le thème est isolé dans une seule phrase, il peut s'agir d'un thème en partie commun aux deux locuteurs. Par ailleurs, si les intervalles finaux des pôles (pôle et thèmes agrégés) se croisent, c'est-à-dire si l'un débute l'allocution et l'autre la termine, nous considérons que les deux pôles sont des thèmes de Chirac, l'un étant un sous-thème de l'autre, et le système indexe toutes les phrases par le locuteur Chirac.

Les phrases indexées de l'allocution 242 sont donc dans les intervalles suivants :

Phrases attribuées à Chirac : intervalles 0_6 30_39 41_69
Phrases attribuées à Mitterrand : intervalles 7_22

Certaines phrases ne sont pas indexées, soit parce qu'elles ne contiennent aucun des thèmes retenus, soit parce qu'elles ne contiennent qu'un thème commun aux deux pôles. Finalement, le système retient comme phrases de Mitterrand les phrases des intervalles du pôle que le système a étiqueté Mitterrand ainsi que les phrases non indexées qui sont comprises entre les limites d'un intervalle Mitterrand et la limite de l'intervalle Chirac qui suit et de celui qui précède. Ce qui donne l'intervalle 7_29 (c'est-à-dire les phrases <242 :8> à <242 :30>) pour les phrases de Mitterrand dans l'allocution 242.

3.4 Les limites de la méthode

La méthode de détermination des pôles thématiques ne s'applique pas toujours aussi bien que dans l'exemple de l'allocution 242, et cela pour plusieurs raisons. Tout d'abord, la reconnaissance des thèmes repose sur les mots les plus fréquents de l'allocution. Or il se trouve que les deux locuteurs ont souvent, à l'intérieur d'une même allocution, un grand nombre de mots en commun. Il arrive même que les phrases de Mitterrand n'aient aucun mot spécifique fréquent, ou que le thème dominant soit un thème commun entre Chirac et Mitterrand. Par ailleurs, si on arrive en général bien à déterminer le thème dominant, en grande majorité attribué à Chirac, en revanche nous n'avons pas trouvé de méthode sûre nous permettant de distinguer entre le thème de Mitterrand et un sous-thème de Chirac. C'est plus particulièrement le cas lorsque l'allocution de Chirac est longue. L'exemple de l'allocution 111 du corpus de test illustre bien ces problèmes.

Pôle dominant *europe* : intervalles 24_44 78_84 99_104 110_119 127_127
Pôle complémentaire *développement* : intervalles 50_50 69_69 131_139

Les deux thèmes principaux de Mitterrand dans cette allocution sont en réalité *exploitation* (intervalle 94_100) et *agricole* (intervalle 94_105), et le thème *développement* est en réalité un sous-thème de l'allocution de Chirac. Le problème vient de ce que le pôle dominant trouvé est un thème commun aux allocutions de Chirac et de Mitterrand : *europe* apparaît en effet à la fois dans l'allocution de Chirac et dans l'intervalle 99_104 des phrases de Mitterrand.

Parmi les 294 allocutions du corpus de test, on a un grand nombre de pôles qui sont des thèmes communs aux deux locuteurs. Le tableau 3 récapitule les différentes combinaisons de locuteurs réels trouvées dans nos résultats.

Locuteur réel du pôle dominant	Locuteur réel du pôle complémentaire	Nombre d'allocutions
Chirac	Mitterrand	67
Mitterrand	Chirac	9
Chirac	Mitterrand et Chirac	45
Mitterrand *et* Chirac (*ou* Chirac *ou* Mitterrand)	(Chirac *ou* Mitterrand *ou*) Mitterrand *et* Chirac	107
Chirac	Pas de pôle complémentaire trouvé	42
Chirac	Chirac	69

Tableau 3 : Les locuteurs réels des pôles thématiques trouvés

4 La coopération entre les deux méthodes

La méthode des modèles de langage n-grammes et la méthode de segmentation thématique produisent chacune un ensemble de phrases attribuées à Mitterrand. Les deux méthodes obtiennent des Fscores proches mais avec des précisions et rappels différents. La méthode utilisant les modèles de langage obtient un très bon rappel mais une précision médiocre. La méthode utilisant la segmentation thématique obtient une meilleure précision mais un plus mauvais rappel. La stratégie de fusion des résultats adoptée consiste à retenir d'une part les phrases présentes dans les deux résultats afin d'augmenter la précision, et d'autre part les phrases de la méthode utilisant les modèles de langage lorsqu'on obtient aucune phrase par l'autre méthode, afin de garder un bon rappel. Nous avons obtenu une légère amélioration des performances. Par exemple, la fusion des résultats pour la tâche 3 donne les résultats suivants :

Modèles de langage : précision = 0.41, rappel = 0.88, Fscore(beta=1) = 0.56
Segmentation thématique : précision = 0.52, rappel = 0.55, Fscore(beta=1) = 0.53
Fusion des deux méthodes : précision = 0.51, rappel=0.66, Fscore(beta=1) = 0.57

5 Conclusion

Au vu des résultats obtenus, nos méthodes ont incontestablement des faiblesses. L'apprentissage du modèle de langage de chaque auteur doit être amélioré par des méthodes de ré-échantillonnage (voir paragraphe 2.3). Par ailleurs, l'existence de nombreux thèmes communs entre les auteurs rend faiblement efficace la segmentation de l'allocution en deux pôles thématiques (voir paragraphe 3.4). Pour résoudre l'ambiguïté créée par les thèmes communs aux deux auteurs, d'autres exploitations des thèmes sont envisagées. En particulier, nous n'avons pas utilisé le positionnement des thèmes les uns par rapport aux autres dans le fil du discours. En effet, certains thèmes sont entrelacés le long du discours alors que d'autres se séparent plus nettement. La fusion des résultats des deux méthodes produit une amélioration assez faible (voir paragraphe 4). Nous envisageons d'étudier le recoupement entre les thèmes trouvés et les prédictions des modèles de langage pour chaque allocution, afin de trouver un autre type de coopération entre les deux méthodes.

Références

CLARKSON P.R., ROSENFELD R. (1997), Statistical Language Modeling Using the CMU-Cambridge Toolkit, Actes de *ESCA Eurospeech 1997*.

GRAHNE G., ZHU J. (2003), Efficiently Using Prefix-trees in Mining Frequent Itemsets, Actes de *First IEEE ICDM Workshop on Frequent Itemset Mining Implementations (FIMI'03)*, Melbourne, FL.

ILLOUZ G., JARDINO M. (2001), Analyse statistique et géométrique de corpus textuels, *T.A.L., Traitement automatique des langues et linguistique de corpus,* Vol 42:2, pp. 501-516.

JARDINO M. (2000), Unsupervised non-hierarchical entropy-based clustering, *Data Analysis, Classification and Related Methods,* Eds. H.-H.Bock, W.Gaul, M.Schader. Springer.

JELINEK F. (1998), *Statistical Methods for Speech Recognition* , MIT Press.

LEBART L., MORINEAU A., PIRON M. (2000), *Statistique exploratoire multidimensionnelle*, Dunod.

SCHMID H. (1999), Improvements in Part-of-Speech Tagging with an Application To German. In Armstrong, S., Chuch, K. W., Isabelle, P., Tzoukermann, E. & Yarowski, D. (Eds.), *Natural Language Processing Using Very Large Corpora*, Dordrecht, Kluwer Academic Publisher.

Extraction d'information à partir de modèles de Markov cachés

Frédéric Kerloch , Patrick Gallinari

LIP6, Equipe Connexionniste – Université Pierre et Marie Curie
8 rue du Capitaine Scott, 75015 Paris
{kerloch/gallinari}@poleia.lip6.fr

Mots-clés : Extraction d'Information, Modèles de Markov Cachés
Keywords: Information Extraction, Hidden Markov Models

Résumé La quantité gigantesque de données textuelles produites sous forme numérique a créé d'énormes besoins en outils capables d'exploiter l'information contenue dans ces textes. L'hétérogénéité et la taille des corpus condamnent par avance toute approche basée sur des techniques purement manuelles. Il faut donc développer des méthodes automatiques capables de structurer les textes, ainsi que d'extraire l'information pertinente. Dans ce cadre, nous avons développé un système d'extraction basé sur des modèles de Markov cachés (MMC). Nous abordons l'apprentissage de la structure ainsi que celui des paramètres d'émission. Nous présentons ensuite des résultats obtenus sur une instance de problème proche de l'extraction, le défi DEFT05.

Abstract The huge quantity of textual data now available has created a need of tools able to exploit the information present in these texts. Manual approaches are inappropriate, due to the heterogeneity and the size of the corpora. Automatic methods must be developed in order to structure the texts and to extract relevant information. In this context, we have developed an extraction system based on Hidden Markov Models. We consider the structure training as well as the estimation of the parameters. We present some results obtained in DEFT05, a data mining challenge close to extraction.

1 Introduction

L'abondance de données disponibles sous forme numérique a rendu cruciales les techniques permettant de faciliter l'accès à l'information pertinente. La conception de ce type de systèmes doit mettre l'accent sur leur caractère automatique, tout travail purement manuel (sans intervention d'un processus automatique) étant voué à l'échec face à la quantité gigantesque d'informations à traiter.

Plusieurs types de tâches peuvent être définies à l'intérieur de cette problématique, selon le niveau de granularité avec lequel les corpus sont abordés:

- Aide à la navigation : au niveau de granularité le plus élevé, un système d'aide à la recherche d'information peut permettre de simplifier la navigation dans les données. Cette aide peut se faire par exemple en structurant hiérarchiquement le corpus (Njike-Fotzo, 2004), ou encore en apprenant automatiquement à détecter des comportements de navigation face à un corpus, permettant ainsi de proposer un système de liens dynamiques s'adaptant à l'utilisateur (Blanchard, 2004)

- Recherche d'information : se plaçant au niveau du document, de nombreuses techniques permettent l'interrogation de larges corpus de données via une indexation du corpus, puis le calcul d'une mesure de similarité performante permettant renvoi des documents les plus pertinents pour une requête donnée.

- Extraction d'information : enfin, au niveau de granularité le plus fin, il est possible d'envisager des systèmes permettant l'identification précise de l'information pertinente. Ceci peut être fait à l'intérieur de corpus homogènes, où chaque texte est censé contenir des informations dont le type est prédéfini (problématique typique de l'extraction d'information), ou de manière plus large dans des corpus hétérogènes directement interrogés à partir de questions posées par l'utilisateur, le système devant renvoyer un passage de texte répondant précisément à la question (problématique de type question / réponse)

Dans la suite, nous nous intéresserons plus particulièrement aux problèmes d'extraction d'information. Dans la deuxième partie, nous donnons la définition précise d'une tâche d'extraction, et faisons un bref état de l'art. Nous détaillons ensuite l'utilisation de modèles de Markov cachés en extraction d'information. La quatrième partie décrit le modèle retenu, ainsi que les méthodes d'apprentissage des paramètres. La cinquième partie présente quelques résultats obtenus sur le corpus DEFT05. La sixième partie présente les conclusions et perspectives.

2 Extraction d'Information

2.1 Définition d'une tâche d'extraction

La définition précise des objectifs et des termes employés en extraction d'information sont issus des campagnes MUC (Messages Understanding Conferences, Grishman 1996). Ces campagnes avaient pour but l'évaluation de méthodes permettant d'extraire de l'information à partir de textes. Elles ont introduit la notion de patron d'extraction, qui contient une description de l'ensemble des champs à extraire dans chaque texte. Une tâche d'extraction consistait pour chaque texte à remplir un patron d'extraction associé. L'évaluation se faisait ensuite en regardant pour chaque champ du patron si les bons extraits de texte étaient présents.

On peut distinguer plusieurs manières d'aborder une tâche d'extraction. La première consiste à effectuer l'extraction de chaque champ séparément : on parle alors d'extraction single slot. Mais il peut être utile de considérer des dépendances entre les champs, que l'on extraira alors simultanément : on parle alors d'extraction multi-slot.

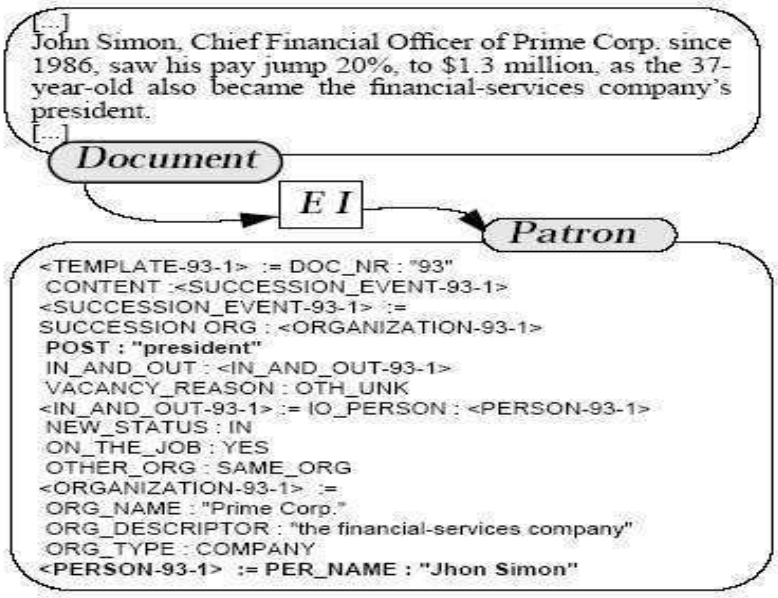


Figure 1 : Un extrait de texte issus de MUC5, et son patron associé

2.2 Etat de l'art

Dans les premières campagnes d'évaluation MUC, la majorité des systèmes était construite à partir d'analyses linguistiques complexes effectuées sur le corpus d'apprentissage. Ces analyses permettaient la construction manuelle de motifs d'extraction. Lors du passage à un corpus différent, tous les motifs étaient à réécrire. Le besoin d'utiliser des méthodes automatiques, ou au moins semi-automatiques, se fit donc rapidement sentir. (Cardie 1993), (Riloff 1993), (Soderland 1995) présentent des systèmes permettant l'acquisition automatique de ressources linguistiques utiles à la construction de motifs d'extraction. Par la suite, trois directions principales de recherche peuvent être identifiées.

La première consiste en l'utilisation de méthodes issues de l'inférence grammaticale pour apprendre des expressions régulières permettant de caractériser les différents champs à extraire. Ces systèmes étaient utilisés initialement pour effectuer de l'extraction dans des pages HTML très structurées qui permettaient une identification simple des champs. L'algorithme BWI (Kushmerick, Freitag 2000) apprend des délimiteurs de champs peu précis, puis améliore l'apprentissage par des techniques de Boosting. (Kosala et al. 2002) construisent des automates d'arbres pour retrouver les noeuds pertinents dans l'arbre de dérivation HTML.

La seconde s'apparente à de l'apprentissage de règles, essentiellement à base d'apprentissage relationnel. (LP)² (Ciravegna 2003) est à ce jour l'algorithme d'extraction le plus performant sur les tâches classiques d'extraction. Il construit ses motifs d'extraction en partant d'une règle vide, puis en lui ajoutant progressivement des contraintes. Seules les meilleures règles générées sont conservées. Deux règles différentes sont générées pour chaque début et fin de champ, ce qui permet d'atteindre un haut niveau de précision. Le rappel est ensuite augmenté en réintroduisant des règles moins bonnes mais qui, en association avec une règle apprise dans l'étape précédente, permettent de finir la délimitation d'un champ.

Le troisième type d'approche utilise des méthodes d'apprentissage statistique pour définir des modèles stochastiques capables d'effectuer l'extraction. On distinguera tout d'abord des méthodes plutôt « génératives » : (Freitag 1999) utilise des Modèles de Markov cachés (MMC) pour effectuer l'extraction. (Peshkin, Pfeffer 2003) étendent ce type de modèle en utilisant des réseaux bayésiens dynamiques (RBD), qui leur permettent d'enrichir simplement leur représentation des token des textes, et d'utiliser des relations entre les différents champs. Il obtient des performances équivalentes à (LP)² sur le corpus classique d'annonce de séminaires [1]. Enfin, (McCallum et al. 2004) introduit un modèle stochastique légèrement plus souple que les MMC, les Conditional Random Fields (CRF), qui leur permettent là aussi d'enrichir leur représentation des tokens. D'autres méthodes utilisent des approches plus « discriminantes », comme (Kushmerick et Finn 2004) qui utilisent des modèles discriminants pour apprendre les frontières de délimitation des champs.

3 Modèles de Markov cachés : application à l'Extraction d'Information

Deux approches sont possibles pour l'utilisation de MMC dans des tâches d'extraction : Une approche single slot, où un HMM est construit par slot. Un état ou plusieurs états sont associés à l'information à extraire. Les tokens générés par ces états seront considérés comme appartenant au type d'information associé. Cette approche est simple et robuste, mais ne permet pas de prendre en compte l'agencement des slots les uns par rapport aux autres. Il peut donc être utile de passer à une approche multi slot, ou un seul MMC est utilisé pour extraire tous les champs. Comme dans l'approche single-slot, on a un ou plusieurs états par champ, plus des états non pertinents, mais ici tous les champs sont représentés. A cause du plus grand nombre de paramètres, l'estimation des probabilités du modèle est moins robuste, mais permet la prise en compte de relations entre les champs.

(Leek, 1997) a le premier introduit les MMC pour effectuer de l'extraction. Il construit à la main une architecture spécifique de MMC pour l'extraction d'information dans des corpus biomédicaux.

(Zaragoza 1998) utilise un modèle de MMC pour de l'extraction dans un corpus de journaux financiers. L'extraction se fait en deux étapes. Dans un premier temps, un classifieur à base de réseau de neurones multicouches est appliqué aux phrases afin de repérer celles susceptibles de contenir de l'information pertinente. Dans la seconde étape, un MMC permet l'identification précise de l'information à l'intérieur des phrases sélectionnées.

(Freitag 1999) utilise des MMC pour de l'extraction dans un corpus d'annonce de séminaires. (Freitag 2000) tente d'apprendre la structure du modèle en partant d'un MMC de base, qu'il complexifie à partir de règles heuristiques. Le MMC obtenant les meilleurs résultats sur le corpus d'apprentissage est conserve.

(Skounakis 2003) utilise des modèles de Markov hiérarchiques. L'idée de base se rapproche de celle de (Zaragoza 1998), la sélection des phrases pertinentes se faisant de manière implicite dans le modèle : le premier niveau hiérarchique du MMC effectue la classification, le deuxième l'extraction dans les phrases pertinentes.

[1] Voir : http://www.isi.edu/info-agents/RISE/

4 Un modèle à base de MMC

4.1 Motivation du choix du modèle

Nous souhaitions utiliser un modèle robuste, dont les probabilités sont simples à estimer, et qui permette une utilisation avec des variables cachées. Notre idée était de pouvoir utiliser notre modèle avec des données non étiquetées, et de pouvoir descendre à un niveau de granularité plus fin dans l'étiquetage (séparer les champs en début / milieu /fin), sans utiliser d'heuristique a priori. Les MMC, RBD et les CRF respectaient les deux premiers critères. Les MMC permettent une utilisation avec des variables cachées et un temps d'apprentissage raisonnable dans le cas de données partiellement étiquetées.

4.2 Description générale du modèle

Le modèle est constitué d'un MMC possédant 3 états par champs à extraire : un état pour les tokens précédant les champs, un état pour les champs, et un état pour les tokens succédant aux champs. Les deux états préfixes et suffixes doivent permettre de capter des régularités dans les séquences annonçant et terminant les champs. Le modèle possède aussi un état « non pertinent » générant tous les autres tokens.

Le modèle d'émission des observations prend en compte d'éventuelles informations venant s'ajouter à la simple donnée du token (tags morpho-syntaxiques, étiquettes sémantiques, capitalisation …). Les symboles d'émission sont donc des vecteurs d'attributs.

4.3 Apprentissage de la structure du modèle

En partant d'un modèle ergodique, la structure est apprise implicitement en estimant directement les probabilités de transition sur la base d'apprentissage. D'autres techniques d'apprentissage de la structure sont envisageables : des techniques de type bottom-up, la structure est construite à partir d'un modèle simple que l'on enrichit progressivement, ou bien top-down, où une structure très complexe est successivement élaguée. Dans les deux cas, une recherche exhaustive dans l'espace des structures est impossible (à cause de l'explosion combinatoire du nombre de modèles à tester). La solution se trouve alors dans des algorithmes de type Hill Climbing, mais qui ne peuvent garantir l'optimalité de la solution trouvée. L'approche proposée ici permet de résoudre simplement le problème de la structure, mais en fixant à priori le nombre d'états et leur sémantique.

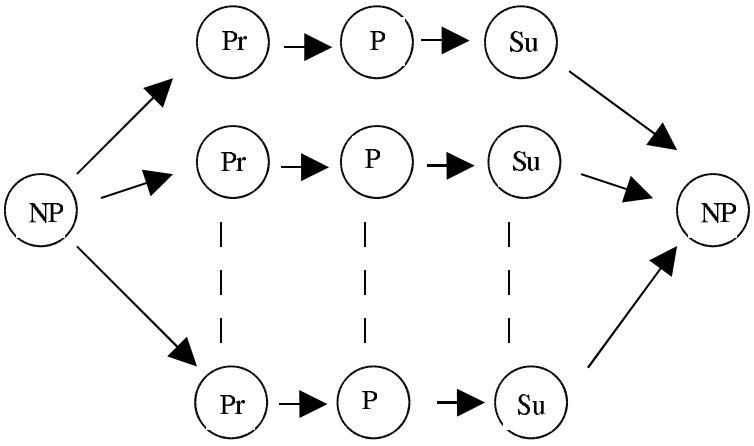


Figure 2 : Modèle après apprentissage de la structure[2], les auto-transitions ayant été supprimées pour des raison de lisibilité

4.4 Apprentissage des paramètres du modèle

Les émissions de notre modèle sont formées par un vecteur contenant les différents attributs de chaque token (informations morpho-syntaxiques, pré-étiquetage d'entités nommées, informations de type majuscule/minuscule, lettre/chiffre).

Dans la phase d'apprentissage, chaque token est associé à un et un seul état : les tokens appartenant à un champ sont associés à l'état correspondant, les tokens dans une fenêtre de *n* mots avant (resp. après) les champs sont associés aux états préfixes (resp. suffixes) respectifs, le reste étant associé à l 'état non-pertinent. Le méta paramètre *n* a été fixé de manière heuristique à 7 pour tous les champs.

Plusieurs types sont d'estimation possibles : on peut estimer directement les probabilités d'émission par maximum de vraisemblance. On se trouve confronté au problème de la taille de l'espace de représentation (on doit estimer un nombre de paramètres égal au produit des tailles des espaces de représentation de chaque attribut). Les estimateurs obtenus ne sont pas assez robustes.

Pour éviter ce problème, il est possible de traiter les attributs comme indépendants sachant l'état. La probabilité d'émission d'un vecteur devient alors le produit des probabilités de chacun de ses attributs, qui sont là aussi estimées par maximum de vraisemblance. Mais cette approche fait apparaître un problème de normalisation dans le produit des probabilités des attributs, ceux ayant le moins de valeurs possibles voyant leur importance relative surévaluée.

Pour remédier à ces problèmes nous avons utilisé un algorithme d'apprentissage discriminant pour estimer les probabilités d'émissions. Nous avons testé plusieurs modèles de classifieurs probabilistes discriminants. Nous avons retenu un modèle à base de machine à vecteurs support à noyau linéaire, appris par la méthode SMO (Platt 1998)

[2] Pour des raisons de lisibilité, les transitions autres que « gauche –droite » n'ont pas été représentées. Elles disparaissant en générale presque toutes lors de la phase d'estimation des probabilités de transition, sauf dans le cas de champs souvent proches les uns des autres.

En l'état, le MMC appris extrait de manière erronée beaucoup de passages introduits par des token fréquents. Il faut donc forcer artificiellement le modèle à ne considérer que des motifs introductifs suffisamment longs. De plus, le modèle a tendance à ne pas rester suffisamment longtemps dans les états pertinents, tronquant très souvent les fins de passages pertinents. Une solution est d'introduire des modèles de durée, afin de maîtriser plus finement le temps passé dans chaque état. Nous avons opté pour un modèle de durée simple consistant à dupliquer tous les états après l'apprentissage des paramètres. Les passages extraits sont ainsi plus longs, et l'apparition dans une séquence d'un token très courant ne suffit plus à basculer dans un état annonciateur (états Pr de la figure 2, correspondant par exmple à des mots introductifs – article…).

5 Evaluation

Le défi DEFT05 ne correspond pas exactement à la tâche d'extraction générique pour laquelle le système a été conçu et utilisé initialement. Les notions de motifs introducteurs et terminaux usuels en extraction n'ont pas de sens ici, de façon plus générale, la notion de séquence et de succession d'évènements n'est pas présente dans la tâche DEFT. Cette dernière est plus proche d'une tâche de classification de phrases ou de segmentation thématique – le style apportant à première vue peu d'information. Nous avons toutefois utilisé une version simplifiée de notre modèle sur les différentes instances de la tâche DEFT pour évaluer si cette classe de méthode était apte à résoudre ce type de problème.

Les phrases de J. Chirac précédant et suivant les discours de F. Mitterrand ne sont ni annonciatrices, ni terminales, car les discours n'ont aucun lien entre eux. Les états préfixes et suffixes de notre modèle n'ont donc pas d'intérêt, et ont été supprimés. Nous obtenons donc un MMC à deux états modélisant les discours de J. Chirac et de F. Mitterrand. Le modèle d'extraction se réduit à un classifieur simple avec des probabilités de transition entre états.

Description du corpus

Le corpus du défi DEFT05 est constitué de discours de J. Chirac, dans lesquels ont été insérées des portions de discours de F. Mitterrand. L'objectif est de retrouver les passages issus d'allocutions de Mitterrand.

Résultats

Les résultats obtenus sont décrits ci-après. Dans la tâche 1, les noms de personnes et les dates ont été remplacés par des balises, dans la tâche 2, seuls les noms ont été remplacés.

	F1 de notre système	F1 moyen	Ecart type
Tâche 1	0.731591	0.629217	0.25852
Tâche 2	0.793889	0.673813	0.22447
Tâche 3	0.788093	0.690224	0.20492

Figure 4 : résultats pour les différentes tâches de DEFT05.

Notre système se situe au dessus de la moyenne, mais nous ne connaissons pas à ce jour les meilleurs résultats obtenus pour cette tâche. Il faut noter que la présence des noms de personnes et des dates améliore les performances de manière significative. Le modèle et la représentation utilisés sont particulièrement simples et devraient pouvoir être facilement améliorés.

6 Conclusion et perspectives

Nous avons développé un modèle général d'extraction d'information à base de modèles de Markov cachés, pouvant prendre en compte des observations de type vectoriel grâce à notre méthode d'apprentissage des probabilités d'émission à partir d'approches discriminantes. Comme (McCallum 2004) et (Peshkin, Pfeffer 2003), il permet l'intégration rapide de plusieurs attributs pouvant être utiles à l'extraction. L'utilisation d'un modèle simple et robuste comme les MMC permet de plus d'envisager l'utilisation de données non étiquetées pour améliorer l'apprentissage, voire de raffiner le niveau de granularité en redécoupant chaque état pour obtenir une modélisation plus fine des champs à extraire.

Nous envisageons plusieurs améliorations du modèle présenté.

- Nous souhaiterions intégrer une approche plus « discriminante » (au sens donné en 2.2) en incorporant dans le modèle des informations plus spécifiques aux frontières entre pertinent / non pertinent. Dans le cas de DEFT par exemple, il serait souhaitable d'intégrer à notre modèle des informations relatives aux transitions entre phrases (les cassures sémantiques semblant par exemple particulièrement pertinentes pour la tâche). Pour l'instant, cela est fait de manière implicite grâce au mécanisme d'états annonciateurs et terminaux. Une première approche possible serait de combiner notre modèle avec un classifieur à fenêtre.

- Beaucoup d'instances de problèmes d'extraction montrent que l'ordre d'apparition des champs est très régulier : même séparés par de longs passages non-pertinents, ils apparaissent dans le même ordre. Pour l'instant, notre modèle peut capter ce type de régularité uniquement pour des champs très proches via le mécanisme d'apprentissage des probabilités de transitions. Il faudrait par exemple introduire dans le modèle un attribut supplémentaire contenant le type du champ précédemment extrait.

Références

BLANCHARD J., PETITJEAN B., ARTIÈRES T., GALLINARI P.(2005), Un système d'aide à la navigation dans les documents hypermédia, Actes de *EGC 2005*

CIRAVEGNA F. (2001), Adaptive Information Extraction from Text by Rule Induction and Generalisation, in Proceedings of *17th International Joint Conference on Artificial Intelligence* (IJCAI 2001)

CARDIE C. (1993), A case-based approach to knowledge acquisition for domain-specific sentence analysis, in Proceedings of the *Eleventh National Conference on Artificial Intelligence*

GRISHMAN R. (1996), Design of the MUC-6 Evaluation. In Proceedings of the *Sixth Message Understanding Conference* (MUC6) , 13-33. Morgan Kauffman

FINN A., KUSHMERICK N. (2004). Information Extraction by Convergent Boundary Classification. In Proceedings of *AAAI-04 Workshop on Adaptive Text Extraction and Mining*

FREITAG D. , MCCALLUM A. (1999), Information extraction using HMMs and shrinkage, In Proceedings of *AAAI-99 Workshop on Machine Learning for Information Extraction*

FREITAG D., MCCALLUM A. (2000), Information extraction with HMM structures learned by stochastic optimisation, In Proceedings of *AAAI-2000*

FREITAG D., KUSHMERICK N. (2000), "Boosted Wrapper Induction". In Proceedings of the *Seventeenth National Conference on Artificial Intelligence* (AAAI-2000)

KOSALA R., V. DEN BUSSCHE J., BRUYNOOGHE M., AND BLOCKEEL H. (2002), Information extraction in structured documents using tree automata induction, In proceedings of *PKDD 02*

LAFFERTY J., MCCALLUM A. AND PEREIRA F. (2001), Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, In Proceedings of *the 18th International Conf. on Machine Learning* (ICML 01)

LEEK, T. R. (1997). Information extraction using hidden Markov models. Master's thesis, UC San Diego

NJIJE-FOTZO H., GALLINARI P. (2004), Apprentissage de relations de spécialisation/généralisation entre concepts – Application à la structuration hierachique automatique de corpus, Actes de *CORIA 04*

PESHKIN L., PFEFER A. (2003), Bayesian Information Extraction Network, In Proceedings of the *Eighteenth International Joint Conf. on Artificial Intelligence* (IJCAI03)

PLATT J. (1998), Fast training of Support Vector Machine using sequencial minimal optimization, *Advances in Kernal methods – Support Vector Learning*, MIT Presse

SODERLAND S., FISHER D., ASELTINE J. AND LEHNERT W. (1995). Crystal: Inducing a conceptual dictionary. In Proceedings of the *Fourteenth International Joint Conference on Artificial Intelligence* (IJCAI95)

SKOUNAKIS M., CRAVEN M. ET RAY S. (2003) Hierarchical Hidden Markov Models for Information Extraction In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, (IJCAI03)

ZARAGOZA H. AND GALLINARI P. (1998), Coupled Hierarchical IR and Stochastic Models for Surface Information Extraction. In proceedings of *The 20th Annual Colloquium on IR Research, British Computer Society's Information Retrieval Specialist Group* (BCG-RSG'98)

Dépendances syntaxiques et méthodes de détection de passages pour une segmentation sur le locuteur et le thème

Loïc Maisonnasse, Caroline Tambellini

Laboratoire CLIPS IMAG – Université Joseph Fourier
385, rue de la bibliothèque - BP 53
38041 Grenoble cedex9
loic.maisonnasse@imag.fr
caroline.tambellini@imag.fr

Mots-clés : découpage thématique, dépendance syntaxique
Keywords: topic segmentation, syntactic dependancy

Résumé Dans cet article, nous présentons les différentes méthodes que nous avons utilisées dans le cadre de la campagne de fouilles de texte DEFT'05. Dans un premier temps, nous présentons les différents aspects du système que nous avons développé pour effectuer la tâche de fouilles de texte. Nous présentons notamment la façon dont nous avons extrait les unités d'indexation, l'apprentissage du poids associé à ces unités, le calcul de la correspondance entre phrase et locuteur et la segmentation thématique. Nous présentons ensuite nos résultats, sur la base desquels nous envisageons d'éventuelles améliorations.

Abstract In this article, we show the methods we used for the DEFT'05 evaluation campaign. In a first time, we describe the different system aspect that we developed in order to complete the text-mining task. More particularly, we present the indexation unit extraction, the learning process used to weight these units, the matching function between sentences and speaker and the topic segmentation. Finally, we present our results at DEFT'05 and we consider some improvements.

1 Introduction

La tâche de DEFT'05 consiste à détecter des phrases de F. Mitterrand dans des allocutions de J. Chirac. Pour ce faire, nous disposons d'un corpus composé d'allocutions de J. Chirac au sein desquelles des portions d'allocutions de F. Mitterrand ont été insérées. Nous proposons ici de déterminer l'auteur d'une phrase à partir d'éléments représentatifs de la syntaxe. Il semble en effet intéressant de savoir s'il est possible de spécifier l'auteur d'une phrase par rapport aux tournures et aux constructions de phrases qu'il utilise. Nous avons également

voulu étudier la piste des changements thématiques pour déterminer les allocutions de F. Mitterrand.

Nous présentons ici notre apprentissage sur les dépendances syntaxiques, puis l'algorithme de détection des ruptures. L'ensemble de nos apprentissages a été effectué sur le corpus de la tâche 3 (celle contenant toutes les informations).

2 Apprentissage sur les dépendances syntaxiques

2.1 Méthode

Pour prendre en compte la syntaxe, nous utilisons des éléments constitutifs de la phrase qui capturent la forme de celle-ci. Le résultat d'un analyseur syntaxique est utilisée, plus particulièrement une analyse en dépendance extraite par l'analyseur 'Xerox Incremental Parser' (XIP) (Aït-Mokhtar et al., 2002). Dans le but de manipuler une structure moins complexe que l'arbre de dépendance, nous considérons le résultat de l'analyse comme un ensemble de dépendances. Chacune de ces dépendances est caractérisée par son type et la liste des lemmes qu'elle relie. Par ce formalisme, la phrase *'le chat mange la souris'* est représentée par : *'{SUBJ(chat,manger), OBJ(souris,manger)}'*

Pour tester l'approche, le corpus fourni pour l'apprentissage a été divisé en deux parties. La première (de l'allocution 100 à l'allocution 5) est utilisée pour l'apprentissage, le reste est utilisé pour l'évaluation. La figure ci-dessous décrit le processus d'apprentissage :

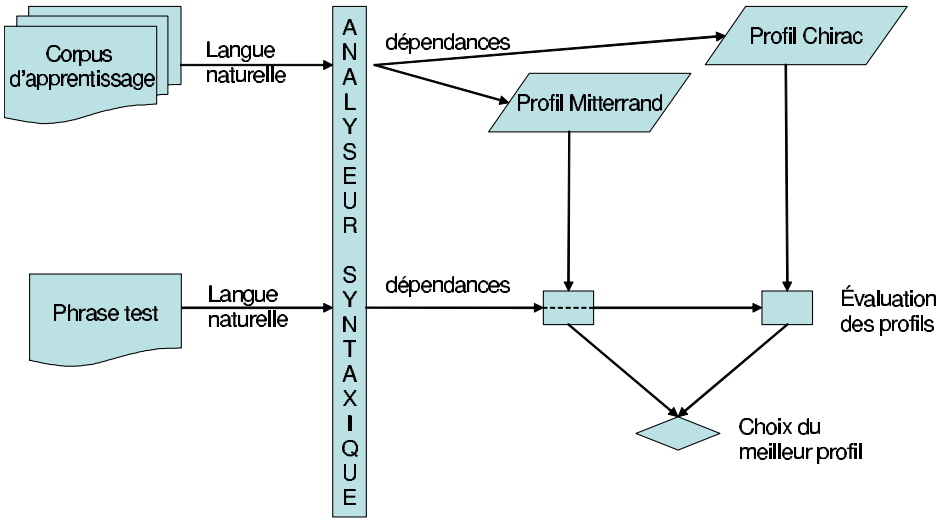


Figure 1 : Processus d'apprentissage

Les dépendances syntaxiques sont utilisées pour la détection de phrase. La première phase de notre approche consiste donc à effectuer une analyse en dépendances sur les phrases du corpus. A partir des résultats de cette analyse, la liste des dépendances syntaxiques de chaque phrase est extraite. Ces dépendances servent de base à un apprentissage permettant d'établir deux profils : un pour F. Mitterrand et un pour J. Chirac. A l'intérieur de ces deux profils, un poids est affecté à chaque dépendance en fonction de sa capacité à distinguer le profil.

Une fois les deux profils créés, chaque phrase du corpus d'évaluation est à son tour analysée. Pour chaque phrase, les dépendances extraites sont stockées sous la forme d'un vecteur et sont pondérées par rapport à leur fréquence. La similarité de ce vecteur par rapport à chaque profil est calculée. Au final, le profil fournissant la meilleure similarité est considéré comme celui correspondant à la phrase.

2.2 L'apprentissage

Nous avons évalué différents types d'apprentissages tout en nous positionnant à différents niveaux de granularité sur le corpus d'apprentissage. Dans un premier temps, nous avons concaténé l'ensemble des phrases de chaque politicien au sein de deux documents. Les dépendances extraites de ces deux documents sont stockées sous la forme de deux vecteurs représentant les profils de chaque président. Le poids de ces dépendances est calculé selon les deux pondérations ltc et lnc présentée ci-dessous :

Ltc	$w_{i,j} = \frac{\ln(f_{i,j}+1) * \log(2/df_i)}{\sqrt{\sum_i f_{i,j}^2}}$
Lnc	$w_{i,j} = \frac{\ln(f_{i,j}+1)}{\sqrt{\sum_i f_{i,j}^2}}$

Figure 2 : Pondération pour l'apprentissage global

où : $w_{i,j}$ est la pondération finale de la dépendance i pour le président j,

$f_{i,j}$ la fréquence de la dépendance i dans le discours de j,

df_i le nombre de documents contenant i (ici 1 ou 2)

Nous avons ensuite calculé les poids pour les dépendances à l'aide d'un apprentissage phrase par phrase. Le poids d'une dépendance résulte de sa répartition dans les phrases pertinentes et non pertinentes d'un profil. Ce calcul est effectué soit par la formule de *Rocchio* soit par la formule utilisée dans (Brouard, 2002) (*N*), ces deux formules sont présentées ci-dessous :

Rocchio	$w_{i,j} = \alpha \frac{\lvert Qi \cap Pj \rvert}{\lvert P_j \rvert} - \beta \frac{\lvert Qi \cap \overline{Pj} \rvert}{\lvert \overline{P_j} \rvert}$, $\alpha = \beta = 1$
N	$w_{i,j} = \frac{\lvert Q_i \cap P_j \rvert}{\lvert P_j \rvert} * \frac{\lvert Q_i \cap P_j \rvert}{\lvert Q_i \rvert}$

Figure 3 : Pondération pour l'apprentissage sur les phrases et les allocutions

Où : P_j Ensemble des documents pertinents pour le président j

Q_i Ensemble des documents contenant la dépendance i

Dans un troisième apprentissage, le poids des dépendances n'est plus calculé phrase par phrase mais par rapport au regroupement par allocution des phrases de chaque président. Les deux mêmes formules d'apprentissage que précédemment sont utilisées.

Une dernière méthode consiste à regrouper les apprentissages de différentes granularités.

2.3 Evaluation

Nous avons appliqué les méthodes précédentes sur notre corpus d'apprentissage. Pour les différentes méthodes, les profils obtenus ont été utilisés pour extraire les phrases de F. Mitterrand du corpus d'évaluation. Lors de cette évaluation les résultats suivants ont été obtenus :

	pondération	fscore
apprentissage global	*ltc*	0,3138
	lnc	0,2400
apprentissage par phrase	*N*	0,1843
	Rocchio	0,1814
apprentissage par allocution	*N*	**0,3286**
	Rochhio	0,2411

Figure 4 : Résultats des différents apprentissages

Le meilleur résultat est celui obtenus avec une pondération N sur les allocutions avec un F-score proche de 0,33. Les moins bons résultats sont ceux obtenus à l'aide de l'apprentissage sur les phrases, cela semble être dû au fait que les phrases prises seules constituent de trop petits éléments d'apprentissage. L'apprentissage global donne de bons résultats notamment par l'utilisation de la pondération *ltc*. En considération de ces résultats, nous avons regroupé l'apprentissage global avec l'apprentissage sur les allocutions, l'apprentissage global est utilisé comme poids initial dans le nouvel apprentissage. Les résultats obtenus sont présentés dans le tableau suivant :

	coef	F-score
Ltc et N	1	0,3195
ltc et Rocchio	1	0,3340

Figure 5 : Regroupement apprentissage global et apprentissage sur les phrases

La pondération globale ltc combinée avec l'apprentissage basé sur la pondération N sur les allocutions fournit des résultats inférieurs à ceux obtenus par la simple formule N. Ces deux formules ne sont donc pas complémentaires. Au contraire, l'apprentissage global ltc combinée avec un apprentissage sur les allocutions de type Rocchio améliore les résultats de base et dépasse les résultats obtenus à l'aide de l'apprentissage N sur les allocutions.

3 Diffusion

Au sein de l'évaluation DEFT'05, les phrases de F. Mitterrand insérées dans les allocutions de J. Chirac sont regroupées. Notre apprentissage ne tient pas compte de cette caractéristique en traitant chaque phrase indépendamment. Nous avons donc mis en place une méthode qui prend en compte le score relatif au locuteur des phrases voisines dans le calcul du score d'une phrase.

3.1 Méthode

Une fois l'ensemble des phrases évaluées par l'une des méthodes d'apprentissage, le score de chaque phrase est recalculé par rapport aux scores des phrases voisines à l'aide de fonctions de diffusion (Huang, 97). Le calcul s'effectue à l'aide de l'équation suivante pour laquelle nous avons testé deux fonctions de diffusion différentes (*A* et *B*) basées sur des cosinus :

$Pt_i = P_i + \sum_{j\in[1,N]} f(j)*(P_{i-j} + P_{i+j})$	
A	$f(j) = \cos(\frac{j*\pi}{2N})$
B	$f(j) = \cos(\frac{j*\pi}{N}) + 0.5$

Figure 6 : Fonctions de diffusion

Où Pt_i Poids final de la i-éme phrase (par ordre de lecture) pour un président

P_i Poids de la i-éme phrase pour un président obtenu par apprentissage

N Taille de la fenêtre des phrases voisines prises en compte

3.2 Evaluation

Pour tester les fonctions de diffusion et la taille de la fenêtre, nous avons utilisé les résultats obtenus lors de l'apprentissage à l'aide de la combinaison Rocchio et ltc. Les résultats de cet apprentissage ont donc été recalculés à l'aide des fonctions précédentes et en faisant varier la taille de la fenêtre utilisée.

	taille de la fenêtre			
fonction de lissage	5	6	7	8
A	0,6664	0,6736	0,6691	-
B	-	0,6696	**0,6742**	0,6731

Figure 7 : Résultats du lissage (F-score)

La fonction de lissage B donne de meilleurs résultats. La taille de fenêtre qui semble la plus adaptée est la taille 7.

4 Découpage thématique

Nous avons voulu étudier la piste des changements thématiques pour déterminer les allocutions de F. Mitterrand. Pour ce faire, nous avons utilisé la méthode du TextTiling de Hearst (Hearst, 1997).

4.1 Principe du TextTiling

Le système de TextTiling (Hearst, 1997) est un système qui recherche les ruptures de thèmes et les identifie lorsqu'un bloc du document présente un moins grand nombre de mots traitant du thème.

Le système de TextTiling (Figure 8) découpe tout d'abord (1) le document en blocs composés d'un nombre fixe de phrases (3 à 5 phrases généralement). Ensuite, (2) toutes les paires des blocs adjacents de texte sont comparées et une valeur de similarité leur est attribuée. (3) La suite résultante des valeurs de similarités, après être mise sous forme de graphe et aplanie, est examinée pour déterminer les pics et les vallées sur le graphique. (4) Des valeurs de similarités élevées, impliquant que les blocs adjacents se suivent de façon logique, sont susceptibles de former des pics, tandis que des valeurs de similarités faibles, indiquant une potentielle limite entre les blocs, créent des vallées. Un pic correspond donc à deux blocs fortement liés thématiquement alors qu'une vallée correspond à une rupture de thème. Chaque vallée est donc considérée comme une rupture de thèmes et correspond à une limite entre deux blocs thématiquement différents.

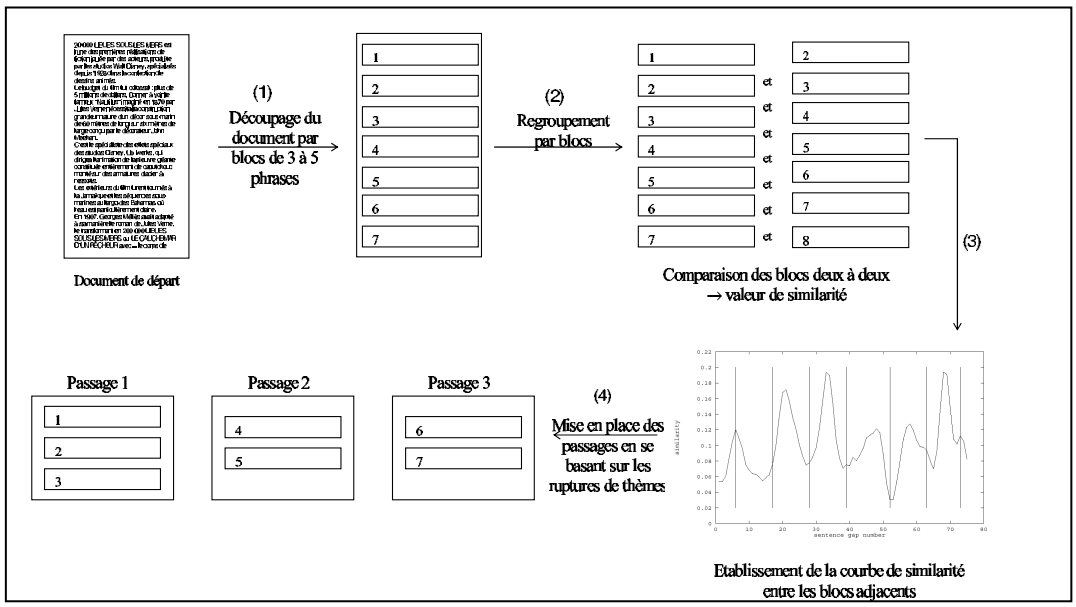


Figure 8 : Méthode du TextTiling

Il a été démontré que ce type d'analyse donne de bons résultats sur des textes dont les termes caractéristiques des thèmes développés ne possèdent pas de synonymes.

4.2 Adaptation au contexte DEFT'05

Afin de déterminer les changements thématiques, nous avons implémenté une adaptation de la méthode du TextTiling. Nous avons gardé le principe du TextTiling et nous l'avons adapté au contexte des tours de parole. Nous détaillons ici les principales étapes du processus :

- Découpage du document en blocs (1)
- Calcul de similarité des blocs pris 2 à 2 (2)
- Détermination du seuil permettant d'identifier les ruptures thématiques (3)
- Détermination des ruptures (4)

Le nombre de phrases constituant les blocs est de 15 tours de parole. Cette valeur a été déterminée suite à différentes expérimentations avec des blocs composés de 10 à 20 tours de parole. L'utilisation de 15 tours de parole pour former un bloc donne les meilleurs résultats. Cette valeur se justifie car les allocutions de F. Mitterrand ont une longueur moyenne de 18 tours de paroles et la majorité des allocutions ont une longueur proche de 15 tours de paroles. Une fois ces blocs de 15 tours de paroles formés (1), la similarité entre les blocs pris deux à deux est calculée (2) :

$$sim(a,b) = \frac{\sum_{t=1}^{n} w_{t,a} w_{t,b}}{\sqrt{\sum_{t=1}^{n} w_{t,a}^2 \sum_{t=1}^{n} w_{t,b}^2}}$$

où t varie pour tous les termes du document et $w_{t,a}$ est le poids tf.idf[1] assigné au terme t dans le bloc a.

Une fois cette similarité calculée, il faut fixer le seuil qui permettra d'identifier les ruptures. Pour ce faire, nous avons effectué plusieurs expérimentations et nous avons obtenus les meilleurs résultats en prenant un seuil correspondant aux 50 % des valeurs de similarités les plus faibles. Plus précisément, nous calculons les valeurs de similarité des blocs pris 2 à 2 (soit 1810 valeurs de similarité dans notre cas), puis nous les ordonnons par ordre décroissant, on se base alors sur la valeur médiane de similarité (soit la valeur de similarité à la position 905, une fois les valeurs ordonnées par ordre décroissant). Cette valeur correspond au seuil (3). Enfin, pour chaque valeur inférieure au seuil, on considère qu'il existe une rupture thématique entre les deux blocs correspondant à cette valeur de similarité (4). Si la valeur de similarité entre un bloc A et un bloc B est inférieure au seuil, la première phrase du bloc B est renvoyée dans un fichier résultat pour être ensuite utilisée dans le processus de détermination des phrases de F. Mitterrand (Figure 9). La première phrase du bloc B sert donc de délimiteur de passages.

---

[1] $\text{Tf.idf} = \frac{\text{nombre d'apparitions du terme dans le bloc}}{\text{nombre de blocs contenant le terme}}$

Les passages ainsi obtenus sont utilisés pour modifier le score des phrases. Le score de chaque passage est calculé en effectuant la moyenne des scores des phrases qu'il contient. Le score final d'une phrase résulte de la somme pondéré entre son score initial et le score du passage dans lequel elle se situe.

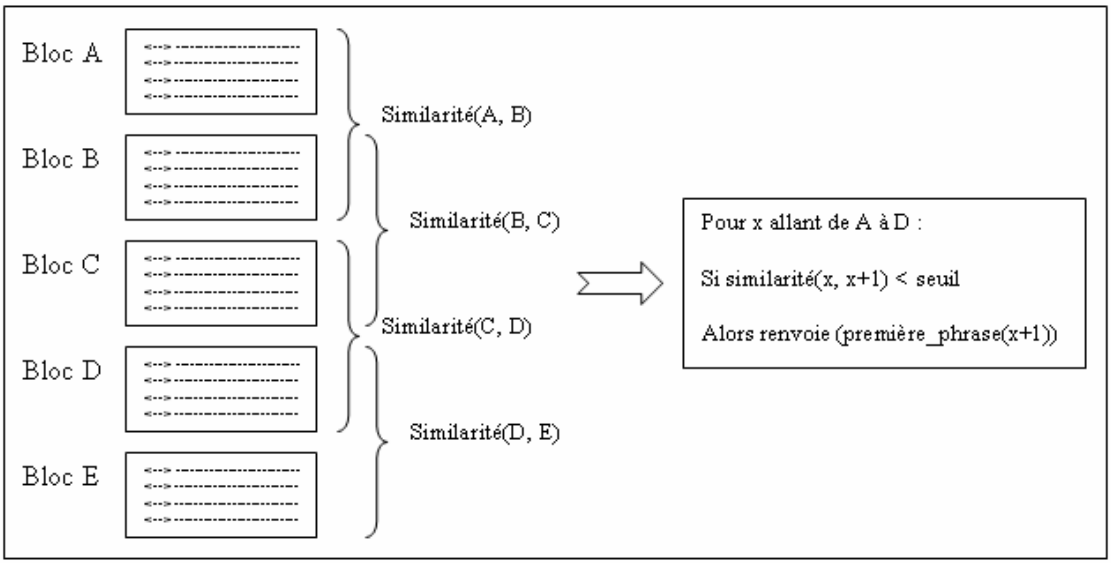


Figure 9 : Principe de détermination des ruptures

4.3 Limites de la méthode

Cette méthode de découpage thématique connaît quelques limites. Tout d'abord, pour avoir de bons résultats, il faut que les allocutions des différents présidents soient thématiquement bien différentes. En effet, pour qu'une rupture thématique soit déterminée il faut que l'on rencontre une différence de similarité entre deux blocs consécutifs. D'autre part, compte tenu du fait que l'on compare des blocs, la rupture thématique ne pourra être déterminée qu'entre deux blocs, or celle-ci peut avoir effectivement lieu au milieu d'un bloc. La méthode du TextTiling ne nous permet pas de le détecter. On constate donc que le choix de la taille du bloc est un problème important dans le bon fonctionnement du système.

5 Soumissions

5.1 Processus

Nos soumissions à l'évaluation DEFT'05 ont consisté à enchaîner les différents modules présentés dans les parties précédentes. Notre approche n'ayant été évaluée que pour la tâche 3, les mêmes exécutions ont été soumises pour les trois tâches (Figure 10).

- L'exécution 1 consiste en un apprentissage global en ltc couplé avec un apprentissage sur les allocutions de type Rocchio (voir partie 2). Sur cet apprentissage, une diffusion basée sur la fonction B avec une taille de fenêtre 7 est appliquée.

- L'exécution 2 est similaire à la première mais l'apprentissage de type Rocchio est effectué sur les phrases.
- L'exécution 3 est similaire à celle de la tâche 1, à la différence prés que les résultats de l'apprentissage sont d'abord modifiés en prenant en compte le poids global des passages détectés. La diffusion n'est effectuée qu'après cette étape.

Exécution 1	Exécution 2	Exécution 3
Ltc + Rocchio (allocutions) + diffusion	Ltc + Rocchio (phrases) + diffusion	Ltc + Rocchio (allocutions) + découpage thématique + diffusion

Figure 10 : Récapitulatif des trois exécutions

5.2 Résultats

Les résultats des trois tâches sont comparés aux moyennes de l'évaluation DEFT'05 comme présentés dans le tableau suivant :

tâche	exécution	précision	rappel	F-score
1	moyenne	-	-	0,6229
	1	0,7477	**0,7549**	**0,7513**
	2	0,9265	0,4216	0,5795
	3	**0,9415**	0,2669	*0,4159*
2	moyenne	-	-	0,6738
	1	0,7533	**0,7563**	**0,7548**
	2	**0,9246**	0,4300	0,5871
	3	0,7943	0,6725	0,7283
3	moyenne	-	-	0,6902
	1	0,7534	**0,7568**	**0,7551**
	2	**0,9268**	0,4337	0,5909
	3	0,7923	0,6725	0,7275

Figure 11 : Résultats d'évaluation DEFT'05

Les meilleurs résultats obtenus dans les trois tâches sont ceux obtenus à l'aide de l'exécution 1. L'utilisation des phrases à la place des allocutions donne un f-score largement inférieur, mais la précision obtenue est améliorée. La troisième exécution donne des résultats intermédiaires sauf pour la première tâche où le résultat est faible. Il est intéressant de remarquer que nos résultats sont stables sur les trois tâches. Cette stabilité peut s'expliquer par le fait que notre système utilise plus la syntaxe que les informations tel que les noms et les dates et que par conséquent celui-ci est peu affecté par leur suppression.

6 Conclusion et perspectives

Notre participation à DEFT'05 nous a permis d'évaluer l'intérêt d'une approche basée sur des éléments représentatifs de la syntaxe pour la détection de phrase ainsi que l'intérêt de la détection de passages dans un contexte de fouilles de texte. Les résultats obtenus montre que dans ce cas les dépendances syntaxiques permettent d'obtenir de bonnes extractions. Filtrer les dépendances extraites dans l'objectif de conserver les plus discriminantes permettrait d'améliorer le résultat. On peut remarquer également que la présence ou non d'informations telles que des dates ou des noms a peu d'impact sur une telle méthode. Toutefois, la prise en compte de telles informations dans un module supplémentaire, permettrait d'améliorer notre système. La détection de passages permet d'améliorer la précision du système. Une détection plus précise des ruptures permettrait d'augmenter cette précision. Une première solution serait l'utilisation d'un pré-traitement du corpus utilisant un anti-dictionnaire et une fonction de stemming.

Références

AÏT-MOKHTAR S., CHANOD J.P., ROUX C. (2002), Robustness beyond shallowness : Incremental Deep Parsing. Actes de *the Natural Language Engineering Journal on Robust Methods in Analysis of Natural Language Data*, Cambridge University Press, pp. 121-144.

BROUARD C. (2002), *RELIEFS : un système d'inspiration cognitive pour le filtrage adaptatif de documents textuels*, Actes de Revue des Sciences et Technologies de l'Information, vol7, no1/2, pp157-182.

HEARST M.A. (1997), TextTiling: Segmenting Text into Multi-paragraphe Subtopic Passage, Actes de *Computational Linguistics*, pp. 33-64.

HUANG C., Principle of information diffusion, Actes de *Fuzzy Sets and Systems*, 91, p. 69-90, 1997.

Classification, combinaison et regroupements pour séparer les discours de Mitterrand de ceux de Chirac

Laurent Pierron (1), Coskun Durkal (2) et Jean-Baptiste Chevalier (3)

(1) Loria – INRIA-Lorraine
Laurent.Pierron@loria.fr
(2) UHP
durka2@etumail.uhp-nancy.fr
(3) ESIAL
Jean-Baptiste.Chevalier@esial.uhp-nancy.fr

Mots-clés : classifieur Bayesien, détection du sujet du discours, structure de phrase, combinaison d'indices, détection de pourriel, fouille de textes, indicateur statistique, algorithme de recherche séquentiel.

Keywords: Bayesian classifier, speech topic detection, sentence structure, clue mixing, spam detection, text miniing, statistical indicator, sequential search algorithm.

Résumé Afin de sélectionner les phrases de Mitterrand parmi les discours de Chirac, nous avons principalement utilisé un classifieur Bayesien, qui a été complété par une détection du meilleur groupe contigu de phrases. Ensuite, nous avons tenté sans succès d'améliorer les résultats du classifieur, une basée sur le thème du discours (national / international), la seconde basée sur la détection de rupture de structure de phrases (nombre de mots).Enfin, une pondération et un seuillage des réponses du classifieur Bayesien ont finalement permis de maximiser le score final.

Abstract To find Mitterrand's sentences inside Chirac's speeches, we mainly used a Bayesian classifier, which was extended with finding the best contiguous block of sentences. After that, we tried to mix the classifier results with two other methods, without improvement, the first one based on the speech topic (national vs international), the second one based on the sentences structure change (identified by number of words). Finally, after Bayesian classifier results weighting and thresholding, we obtained our best score.

1 Introduction

Après avoir analysé rapidement les phrases, nous avons décidé d'utiliser plusieurs approches pour trouver des indices permettant de déterminer les extraits de discours de Mitterrand au sein des

discours de Chirac. Ces divers indices sont ensuite combinés et suivis d'une dernière technique permettant de sélectionner les blocs de texte contigus appartenant aux discours de Mitterrand.

Les trois approches utilisées afin de sélectionner les phrases de Mitterrand sont les suivantes :

1. Classification Bayesienne des phrases sur des groupes lexicaux extraits des phrases et sur le numéro de ligne.

2. Détection des ruptures sémantiques en début et fin des extraits de discours de Mitterrand par mesure de distance entre les phrases.

3. Recherche de l'unité de thème dans le discours, uniquement réalisée sur l'opposition national/international.

La première partie décrira la méthode utilisée pour la classification Bayesienne et les résultats obtenus, la seconde partie décrira la détection des ruptures sémantiques, la troisième l'unité de thème et la dernière partie la combinaison.

Aucun traitement spécifique par tâche n'est envisagé, la méthode choisie doit permettre de s'adapter aux différentes tâches et tenir compte automatiquement des informations supplémentaires.

2 Classification Bayesienne

2.1 Motivation

Dans cette approche, nous avons décidé de travailler sur les phrases plutôt que sur les discours, car le but du challenge est de trouver les phrases de Mitterrand, si le sujet avait été de déterminer l'auteur du discours, notre approche aurait certainement été différente.

Les phrases de Mitterrand insérées au milieu des discours de Chirac font penser aux messages non sollicités (*spam*) insérés au milieu des messages utiles dans les messageries électroniques. Celui qui a une messagerie électronique lourdement encombrée par le *spam*, peut constater que les filtres *anti-spams* à base de classifieurs Bayesiens donnent de bons résultats dans le tri du courrier.

Même si les phrases de Chirac et Mitterrand peuvent dans certains cas être assez similaires étant issues de discours politiques relativement convenus, les différences de sujets et d'époques peuvent certainement fournir des indicateurs discriminants pour les phrases.

2.2 Classifieur Bayesien

La technique utilisée pour le classifieur Bayesien est celle décrite par (Robinson, 2003).

Les phrases sont divisées en deux classes, les phrases de Chirac et les phrases de Mitterrand.

Chaque phrase est divisée en groupes lexicaux ou graphiques, que nous appellerons *mots* pour simplifier le discours. La présence d'un *mot i* dans une phrase, dont l'auteur est inconnu, permet au classifieur de déterminer une probabilité p_{ci} d'appartenir à la classe Chirac et une probabilité p_{mi} d'appartenir à la classe Mitterrand.

Pour chacune des deux classes Mitterrand et Chirac les probabilités individuelles des *mots* sont combinées pour obtenir une probabilité unique pour chaque phrase. La formule utilisée pour calculer la combinaison est la formule de Fisher proposée par (Robinson, 2003).

2.3 Probabilités individuelles

Les probabilités individuelles des mots sont obtenues par entraînement sur un corpus de phrases dont les auteurs sont connus.

Pour chaque mot *m* dans le corpus d'apprentissage on calcule :

- *Mit(m)* = (le nombre total de phrases de Mitterrand contenant le mot *m*) / (le nombre total de phrases de Mitterrand)
- *Chi(m)* = (le nombre total de phrases de Chirac contenant le mot *m*) / (le nombre total de phrases de Chirac)
- $p_M(m) = Mit(m)/(Mit(m)+Chi(m))$

$p_M(m)$ peut grossièrement être interprétée comme la probabilité qu'une phrase choisie au hasard contenant le mot *m* soit une phrase de Mitterand. On calcule de la même manière $p_C(m)$, le probabilité qu'une phrase choisie au hasard contenant le mot m soit une phrase de Chirac.

2.4 Gestion des mots rares

Il y a un problème avec les probabilités calculées ci-dessus, si un mot est très rare. Par exemple si un mot apparaît dans une seule phrase, qui est une phrase de Mitterrand, la probabilité $p_M(m) = 1.0$. Mais il n'est pas du tout sûr que toutes les phrases futures contenant ce mot seront des phrases de Mitterand, nous n'avons simplement pas assez de données pour décider.

Quand une et une seule phrase contient un certain mot et que cette phrase est une phrase de Mitterrand, notre croyance que la prochaine fois que nous verrons ce mot la phrase soit une phrase de Mitterrand n'est pas de 100%. Ceci parce que nous avons une connaissance a priori, qui nous guide. C'est le grand nombre de répétitions de l'evènement qui nous fera penser qu'il y a de fortes que chance que la prochaine fois que nous verrons ce mot ce soit une phrase de Mitterrand.

Pour tenir compte de cette croyance a priori, nous pouvons utiliser la formule suivante détaillée dans (Robinson, 2003) :

$$f_M(m)=\frac{(s\cdot x+n\cdot p_{(M)}(m))}{(s+n)}$$

Où :

- $p_M(m)$ est la probabilité que la phrase soit une phrase de Mitterrand si elle contient le mot *m*.
- *n* est le nombre de phrases contenant le mot *m*.

- s est la force de la croyance, elle peut prendre une valeur positive quelconque, en général entière supérieure ou égale à 1, si elle vaut 0, on retrouve la formule intiale avec $p_M(m)$. Pour le challenge, on a utilisé la valeur s à 1.

- x est la probabilité initiale que la phrase soit une phrase de Mitterrand, pour le challenge nous avons utilisé 0.5.

Il est également possible d'optimiser les valeurs de s et x par apprentissage.

Dans le programme de classification nous avons donc utilisé f_M et f_C en lieu et place de p_M et p_C.

2.5 Combinaison des probabilités

A ce point nous disposons d'une probabilité $f(m)$ qu'un mot donné soit dans une phrase de Mitterand ou de Chirac. Donc à chaque phrase est associé deux ensembles de probabilités, un ensemble pour les probabilités d'appartenir aux phrases de Chirac et un ensemble pour les probabilités d'appartenir aux phrases de Mitterrand.

En sortie du classifieur on souhaite obtenir une seule valeur de probabilité pour chacun des ensembles de phrases de Chirac et de Mitterrand, il est donc nécessaire de combiner les probabilités, pour cela une des techniques les plus éprouvées a été mise au point par R.A. Fisher. If we have a set probabilities, $p1$, $p2$, ..., pn, we can do the following. First, calculate -2ln $p1$ * $p2$ * ... * pn. Then, consider the result to have a chi-square distribution with $2n$ degrees of freedom, and use a chi-square table to compute the probability of getting a result as extreme, or more extreme, than the one calculated. This ``combined" probability meaningfully summarizes all the individual probabilities. Si nous avons un ensemble de probabilités, $p1$, $p2$, ..., pn, nous pouvons effectuer ce qui suit. Premièrement calculer 2ln $p1$ * $p2$ * ... * pn. Puis, nous considérons que le résultat suit une distribution du chi-deux avec $2n$ degrés de liberté, et on utilise une table du chi-deux pour calculer la probabilité d'obtenir un résultat aussi élevé, ou plus élevé, que celui calculé. Cette probabilité « combinée » résume significativement les probabilités individuelles.

2.6 Sélection des ensembles de travail

Le corpus d'entraînement, donné pour le challenge, est divisé en deux parties :

- une pour la création automatique du classifieur, c'est-à-dire la table des probabilités des *mots,*

- la seconde pour tester l'apprentissage en calculant les probabilités d'appartenance des phrases aux discours de Chirac et aux discours de Mitterrand.

Les deux parties ont été obtenues de deux manières différentes pour vérifier que les résultats restent stables. D'abord en coupant le corpus en deux parties égales au milieu de l'ensemble des phrases, puis en utilisant les discours de numéro pair pour l'apprentissage et les discours de numéro impair pour le test. Les résultats obtenus sur ces deux ensembles de travail ont été similaires, nous ne présenterons par la suite que les résultats obtenus sur le second ensemble de travail.

2.7 Sélection des groupes lexicaux

Nous avons utilisé trois ensembles de groupes lexicaux, afin d'essayer de trouver un ensemble optimal ou dans le but de les combiner.

4. D'abord l'ensemble le plus évident, chaque phrase est divisée en mots, un mot est une suite de lettres accentuées ou non et de chiffres. Un mot a une longueur d'une lettre ou plus. Chaque mot est un attribut caractérisant une phrase. Ce premier ensemble d'attributs sera A_{mots}.

5. Ensuite, nous avons choisi de découper le texte en tranche de n caractères non glissants, en basant l'idée sur l'article de (Brunet, 2003), qui montre qu'il n'est pas nécessaire de faire un découpage lexical parfait pour comparer des textes. Nous avons effectué plusieurs essais pour trouver un maximum pour le F-score à 7 caractères. Un des avantages de cette approche est de prendre en compte la ponctuation, qui peut varier entre deux auteurs voire entre deux types de discours. Une expérimentation pourrait être faite en faisant glisser la fenêtre de caractères.

6. Nous avons ensuite généralisé la première méthode, en l'appliquant pour des groupes lexicaux de 1 à 5 mots. Dans les groupes lexicaux de 1 mot, les mots de longueur inférieure à 3 ne sont pas pris. Nous avons effectué plusieurs essais en faisant varier le nombre maximumde mots dans un groupe et nous avons obtenu un F-score maximum pour 4 mots. C'est cette dernière méthode de lemmatisation des phrases, qui a été utilisée pour fournir les résultats du challenge, car bien qu'elle ne donnait pas un F-score très différent en sortie directe du classifieur Bayesien elle s'avérait être la plus performante après la détection des blocs contigus.

Pour calculer les F-scores en sortie du classifieur Bayesien, nous avons attribué une phrase à Mitterrand si et seulement si la probabilité que la phrase appartienne à l'ensemble des phrases de Mitterrand est supérieure à la probabilité que la phrase appartienne à l'ensemble des phrases de Chirac.

2.8 Résultats

Le programme Python Reverend de (Bakhtiar, Delord, 2003), après correction de la formule de combinaison des probabilités individuelles pour être confrome à celle exposée par (Robinson, 2003) et explicitée dans la section 2.2, a été utilisé pour créer le classifieur Bayesien et effectuer la classification.

Les résultats du point de vue du F-Score sont assez similaires de 0.35 sur les mots simples, 0.40 pour les heptagrammes et 0.42 pour les groupes lexicaux de 4 mots. Il faut remarquer quand même que sur l'ensemble d'apprentissage, les heptagrammes obtiennent un score de 0.70 et atteint 0.99 si on retire les phrases mal classées, qui représentent 10% de l'ensemble d'apprentissage, par contre ce sur-apprentissage n'améliore pas le F-score sur l'ensemble de test, il a même tendance à le faire descendre légèrement.

3 Détection et regroupement des blocs contigus

3.1 Méthode

Nous avons créé un programme qui prend en entrée un texte, qui est une suite de phrases dans l'ordre du texte, chaque phrase ayant un et un seul attribut auteur déterminé par une prise de décision après passage dans le classifieur Bayesien ou par une autre méthode.

Ce programme cherche la plus grande suite de phrases attribuées à Mitterrand. Comme il est possible d'avoir pris une mauvaise décision précédemment nous autorisons que des phrases de Chirac soient incluses dans le bloc de phrases de Mitterand. Une phrase de Chirac est autorisée si elle est précédé et suivie par une phrase de Mitterrand, nous avons également essayé avec deux phrases et trois phrases de Mitterrand, dans nos test l'optimum pour le F-score était pour deux phrases avant et après celle de Chirac, mais les résultats sont très voisins. Nous avons également tenté d'accepter deux phrases de Chirac au milieu de celles de Mitterand, mais les résultats se sont dégradés.

Afin de limiter l'apparition de petits blocs de textes, nous n'acceptons finalement un bloc de texte, comme appartenant à Mitterand que si ce dernier a une taille minimale, nous avons fait varier la taille du bloc de 4 à 12 pour trouver un F-score maximal (0,72) pour des blocs de taille 8 sur le jeu d'essai.

Ce programme force également l'attribution des deux premières phrases du texte et des deux dernières à Chirac.

3.2 Seuillage des phrases de Mitterrand

Pour améliorer le score on crée un nouvel indice et un nouveau seuil pour décider qu'une phrase appartienne à l'ensemble des phrases de Mitterrand. Ces calculs sont placés entre la sortie du filtre Bayesien et lla détection des blocs contigus.

Le nouvel indice est égal à 2 fois la probabilité qu'une phrase soit de Mitterrand ajouté à un moins la probabilité que la phrase appartienne aux phrases de Chirac.

Le seuil est appris pour obtenir le meilleur F-score sur l'ensemble de tests à la sortie de la détection des blocs contigus. Un seuil de 1,7 permet d'obtenir le meilleur F-score qui est de 0,80, donc augmentant de'environ 10% le score obtenu sur les blocs contigus détectés directement en sortie du filtre Bayesien.

3.3 Renforcement avec international

Nous avons tenté d'améliorer la détection des blocs de phrases de Mitterrand en travaillant sur un thème : l'international.

Le thème international est défini par une liste de groupes lexicaux relatifs à des discours internationaux, obtenue de manière semi-automatique à partir de noms de pays, de noms d'habitants et de formule de politesse (cher Président, votre Altesse, etc.).

Pour chaque texte, si les deux premières lignes et les deux dernières lignes contiennent des groupes lexicaux parlant de l'international, on suppose que le discours de Chirac est international.

Alors on a des phrases de Mitterrand, quand on trouve des mots se rapportant à des thèmes nationaux.

En phase de chaque phrase, on met donc une probabilité 1 ou 0 d'être une phrase de Mitterrand. On combine le résultat obtenu pour étendre les blocs contigus détectés précédemment.

Cette technique n'a pas permis d'améliorer le F-score, mais ne l'a pas fait baisser.

4 Détection de blocs par nombre de mots

Après avoir fait un calcul sur la moyenne de nombre de mots par phrase de l'un et de l'autre des présidents de la République étudiés, sur le corpus entier, on s'aperçoit que Chirac avait une moyenne de mots par phrase beaucoup plus faible que Mitterrand. Environ 23 pour Chirac alors qu'elle est de près de 29 pour Mitterrand. Nous avons donc eu l'idée de tester un indicateur qui serait le nombre de mots par phrase.

Cependant, considérant qu'il pouvait toujours y avoir une phrase de peu de mots, mais que cette statistique était plus correcte pour un certain nombre de phrases, et sachant que les phrases de Mitterrand étaient regroupés en un bloc à l'intérieur des discours de Chirac, il est apparu nécessaire de travailler sur des fenêtres glissantes.

Nous avons donc utilisé des fenêtres de 7 phrases en glissant par pas de 3 phrases.

Nous effectuons un parcours des fenêtres comme suit : la 1ère fenêtre est celle commençant à la ligne 1, la 2ème est celle commençant à la ligne (dernière ligne – taille fenêtre), la 3ème est celle commençant à la ligne 4 (1 + pas), la 4ème commence à la ligne (dernière ligne – taille fenêtre – pas), et ainsi de suite. On associe la valeur 0 aux 2 premières fenêtres, donc en fait la 1ère et la dernière dans l'ordre du discours.

Ensuite, pour la ième fenêtre calculée, on lui attribue pour valeur :

La somme des différences entre le nombre des mots moyen des phrases de la fenêtre courante, et le nombre de mots moyens des fenêtres calculées précédemment, mais uniquement celles qui lui sont proches, donc on va de 2 fenêtres en 2, et on diminue cette somme par rapport à la distance des 2 fenêtres (plus une fenêtre est loin, moins elle doit avoir d'influence sur le calcul de la fenêtre courante).

On obtient donc des lignes comme celle-ci :

<100:1> 0 C C C C C C C
<100:33> 0 C C C C C C C
<100:4> 2.28571428571 C C C C C C C
<100:30> 1.0 C C C C C C C
<100:7> 2.42857142857 C C C C C M M
<100:27> 11.1428571429 M M C C C C C
<100:10> 2.71428571429 C C M M M M M
<100:24> 47.5714285714 M M M M M C C
<100:13> 14.2857142857 M M M M M M M
<100:21> 58.7142857143 M M M M M M M
<100:16> 18.0 M M M M M M M
<100:18> 13.4285714286 M M M M M M M

<100:19> 24.1428571429 M M M M M M M
<100:15> 32.2857142857 M M M M M M M

Dans le discours 100, les fenêtres commençant à la ligne n°1, 33, 4, 30, dans l'ordre de calcul. Les 7 caractères disent si c'est Chirac ou Mitterrand qui parle à chaque ligne.

Ici, plus la valeur est grande, plus il y a eu rupture de nombre de mots par phrase, donc on est censé détecter des coupures de blocs avec cette méthode, et aussi comme l'heuristique était que Mitterrand avait plus de mots par phrase que Chirac, plus la valeur est grande plus il devrait y avoir de chances que Mitterrand parle.

Ensuite, on on construit un nouveau fichier :

Pour chaque ligne de chaque discours, on fait la somme des valeurs données par le fichier précédent, par exemple pour la ligne 12, on a fait la somme des fenêtres commençant aux lignes 7 et 10 car la ligne 12 est comprise dans les fenêtres [7,14] et [10,17].

Cela donne des valeurs pour chaque ligne de chaque discours.

Exemple :

<3:73:M> 476.285714286
<3:70:M> 405.428571428
<3:76:M> 392.428571429
<3:67:M> 373.714285714
<3:90:C> 344.285714286
<3:91:C> 344.285714286
<3:79:M> 333.857142858
<3:88:M> 325.142857143
<3:74:M> 305.142857143
<3:75:M> 305.142857143
<3:68:M> 296.714285714
<3:69:M> 296.714285714
<3:93:C> 292.714285714
<3:77:M> 283.714285715
<3:78:M> 283.714285715
<3:71:M> 279.857142857
…

On trie par valeur décroissante, ce qui signifie d'après l'heuristique que les lignes parlées par Mitterrand devraient se retrouver en haut de la liste.

Par la suite, on obtient avec un passage sur ce fichier, un nouveau fichier, qui se présente comme suit :

3 [54:86] 27176.4162603
3 [87:101] 7403.68292573
3 [117:134] 3677.4696891
3 [135:152] 2338.71978797
3 [111:116] 1626.0858952
3 [10:27] 1498.41437507
3 [37:45] 1496.41955674

3 [153:174] 1409.26083185
3 [104:110] 1328.16439518
3 [46:51] 846.077319479
3 [28:36] 649.31079412
3 [102:103] 317.477678571
3 [52:53] 205.320408163
…

Cela représente donc des blocs. En effet, on a essayé de reconstruire des blocs à partir des valeurs précédentes. On met la première ligne dans un bloc, puis si la deuxième ligne est assez proche, on la rajoute au bloc, sinon on crée un nouveau bloc, et ainsi de suite. Donc on obtient des petits blocs de phrases avec la somme des différentes lignes dedans. Ainsi, plus un bloc a une grande valeur, plus les éléments dedans avaient déjà une grande valeur, et plus ils étaient regroupés dans le précédent fichier.

Voici les scores que nous avons obtenu en prenant des blocs sur ce fichier :

Les 3 premiers blocs de chaque discours pour fenêtre : Fscore(4270, 24143, 7523) = 0.270

Les 2 premiers blocs de chaque discours pour fenêtre : Fscore(3571, 17471, 7523) = 0.286

Le premier bloc de chaque discours pour fenêtre : Fscore(2458, 9754, 7523) = 0.285

Cette méthode a été combinée à la détection de blocs contigus, mais une fois encore le F-score n'a pas été amélioré.

Remerciements

Langage de programmation Python et son auteur Guido van Rossum, sans lequel nous n'aurions pas pu développer les programmes de ce challenge aussi rapidement.

Martine Cadot pour ses précieux conseils sur l'analyse et la fouille de données et nous avoir fait connaître ce challenge.

Références

ROBINSON G. (2003), A Statistical Approach to the Spam Problem, *Linux Journal*, http://www.linuxjournal.com/article/6467.

BAKHTIAR A., DELORD C. (2003), Divmod Reverend, http://www.divmod.org/Home/Projects/Reverend.

BRUNET E., (2003), PEUT-ON MESURER LA DISTANCE ENTRE DEUX TEXTES ?, CORPUS, NUMERO 2 LA DISTANCE INTERTEXTUELLE -décembre 2003, http://revel.unice.fr/corpus/document.html?id=30.

DEFI DEFT05 : une approche par classifieur de Bayes

Michel Plantié, Gérard Dray, Jacky Montmain,
Alexandre Meimouni, Pascal Poncelet

LGI2P – Ecole des Mines d'Alès
Parc George Besse, 30000 Nîmes
{michel.plantie, gerard.dray, jacky.montmain}@ema.fr
{alexandre.meimouni, pascal.poncelet}@ema.fr

Mots-clés : Modèles vectoriels, Classifieur de Bayes,
Keywords: Vector models, Bayes classifier,

Résumé Cet article présente notre approche de la problématique soulevée par le défi DEFT05. Cette approche est basée sur une représentation vectorielle des textes, et par l'application de méthodes de classification de «Bayes ».

Abstract This paper presents our approach of the DEFT05 problem. This approach is based on a vector representation of texts, and the application of "Bayes" classifier.

1 Introduction

L'objectif du défi fouille de textes 2005 (DEFT'05) était de supprimer des phrases non pertinentes dans un discours politique. Pour cela un corpus d'apprentissage, constitué de phrases issues de discours de François Mitterrand parsemées dans des discours de Jacques Chirac, a été fourni aux participants. Les participants au défi devaient définir une méthode, automatique (ou semi-automatique) et reproductible, permettant de détecter les phrases de F. Mitterrand. L'évaluation des résultats des méthodes proposées a été réalisée par les organisateurs sur un corpus de test.

Nous avons abordé le défi comme un problème de classification de textes. En choisissant cette approche nous avons donc considéré que les phrases étaient indépendantes les unes des autres, la problématique à traiter se traduisant par :

« Quelle méthode de représentation des textes et quelle méthode de classification permettraient de classer au mieux les phrases de François Mitterrand et de Jaques Chirac? »

2 Méthode expérimentale

2.1 Prétraitement et indexation du corpus

Le corpus d'apprentissage initial était composé de 50096 phrases de J. Chirac et de 7183 phrases de F. Mitterrand. Dans une première étape les phrases du corpus ont été étiquetées et lemmatisées. La seconde étape a consisté à indexer les phrases sous la forme de vecteurs. Trois modèles vectoriels ont été utilisés pour les expérimentations ; chaque coordonnée du vecteur correspondant à un mot du vocabulaire du corpus total (les mots vides étant éliminés).

- Modèle binaire : pour chaque mot de la phrase appartenant au vocabulaire du corpus, la coordonnée du vecteur est mise à 1, les autres coordonnées étant à 0.

- Modèle FT (Fréquence des Termes) : pour chaque mot de la phrase appartenant au vocabulaire, la coordonnée correspondante du vecteur est incrémentée. La valeur d'une coordonnée indiquant le nombre d'occurrences du mot dans la phrase. (Remarque : les phrases étant relativement courtes et les mots vides étant éliminés ce modèle est peu différent du modèle précédent)

2.2 Méthode de classification

Comme nous l'avons indiqué dans l'introduction, nous avons abordé le défi comme un problème de classification de textes. La classification de textes consiste à assigner des catégories prédéfinies à des documents textuels. Soit $C = \left(c_1,...,c_j,...,c_n\right)$ l'ensemble des classes d'appartenance possibles pour un document.

Nous avons choisi, pour des raisons de rapidité de calcul, le classifieur naïf de Bayes. Cette méthode peu complexe, a déjà fourni de bons résultats sur des problèmes de classification de textes. Cette approche suppose l'existence d'un modèle stochastique de génération des documents textuels. En inversant ce modèle, nous pouvons prédire, pour un nouveau document, la probabilité d'appartenir à une classe quelconque. La règle de classification de Bayes consistant à attribuer la classe dont la probabilité est la plus élevée.

Un document est représenté par un vecteur $d_i = \left(ft_{i1}...,ft_{it},...,ft_{i|V|}\right)$ dans lequel V représente l'ensemble des mots du vocabulaire retenu du corpus et ft_{it} représente le nombre d'occurrences du mot m_t dans le document d_i.

La règle de classification de Bayes consiste à attribuer à un document d_i la classe dont la probabilité suivante est la plus élevée :

$$P(c_j/d_i)=\frac{P(c_j)P(d_i/c_j)}{P(d_i)},$$

avec c_j une classe et d_i une phrase.

La probabilité $P(d_i)$ étant indépendante des classes, la règle de classification de Bayes peut être exprimée par :

$$\hat{c}(d_i)=\arg\max_j\left(p(c_j)\,p(d_i/c_j)\right),$$

$\hat{c}(d_i)$ représentant la classe estimée pour le document d_i

En considérant l'indépendance des mots :

$$\hat{c}(d_i)=\arg\max_j\left(p(c_j)\prod_{t=1}^{|V|}p(m_t/c_j)^{ft_{ti}}\right)$$

Les probabilités $p(m_t/c_j)$ et $p(c_j)$ peuvent être estimées sur un corpus d'apprentissage par :

$$p(m_t/c_j)=\frac{1+\sum_{d_i\in c_j}ft_{ti}}{|V|+\sum_{t=1}^{|V|}\sum_{d_i\in c_j}ft_{ti}}$$

$p(c_j)=\frac{|c_j|}{\sum_{k=1}^{|C|}|c_k|}$, avec $|c_j|$ représentant le nombre de documents de la classe j.

Cette méthode a été appliquée à la problématique du défi en considérant : deux classes (Mitterrand et Chirac) et les phrases comme des documents.

Dans le cas de la représentation vectorielle binaire, les fréquences des termes sont remplacées par une variable bin_{ti} valant 1 si le mot t apparaît dans la phrase i et 0 sinon.

2.3 Méthodologie d'apprentissage et résultats

2.3.1 Constitution des ensembles d'apprentissage

Le nombre de phrases de F. Mitterrand dans le corpus d'apprentissage étant disproportionné par rapport au nombre de phrases de J. Chirac, nous avons choisi de créer plusieurs ensemble d'apprentissage. Au total 7 ensembles ont été formés, en conservant dans chacun d'entre eux toutes les phrases de F. Mitterrand et en complétant par le même nombre de phrases de J. Chirac par tirage aléatoire.

Sur chacun de ces ensembles un apprentissage par validation croisée à dix ensembles a été réalisé. Pour chaque modèle vectoriel de représentation nous avons donc pu estimer une moyenne des performances envisageables de notre approche sur le jeu de test.

2.3.2 Stratégie de vote

Chaque phrase du jeu de test est analysée par 7 classifieurs et est donc associée à 7 réponses. L'idée de base revient ensuite à considérer chaque classifieur comme un « expert » qui se prononce sur l'appartenance ou non d'une phrase à une classe. Chaque classifieur C_k donne la classe d'appartenance du candidat ainsi que sa probabilité d'appartenance. Pour chaque phrase on construit alors le vecteur à 7 composantes des probabilités d'appartenance à la classe « Chirac ». Pour un classifieur donné, une probabilité supérieure à $\alpha = 0.5$ signifie que le classifieur associe la phrase à la classe « Chirac » ; il s'agit maintenant de combiner les avis des 7 « experts » sur la base de cette règle de choix individuel. Nous nous sommes donc intéressés à la sémantique des règles de choix collectif comme :

- Le vote majoritaire à α % avec $0.5 \leq \alpha \leq 1$. Il fournit des comportements de vote allant de la majorité à l'unanimité.
- Le vote unanime restreint : choisir l'alternative approuvée par la plupart des classifieurs.
- Le veto limité : rejeter une alternative rejetée au moins par q classifieurs.

Dans ces règles, la notion de quantifieur linguistique est implicite. Un quantifieur linguistique modélise une proportion floue et permet d'expliciter des propositions telles que : « q classifieurs parmi n préfèrent la solution S_j ». Le quantifieur Q « au moins q classifieurs parmi n » peut également être représenté par un ensemble flou où $\mu_Q(n)=1$ et $\mu_Q(j) \leq \mu_Q(j+1),\ j=0,n-1$ (voir).

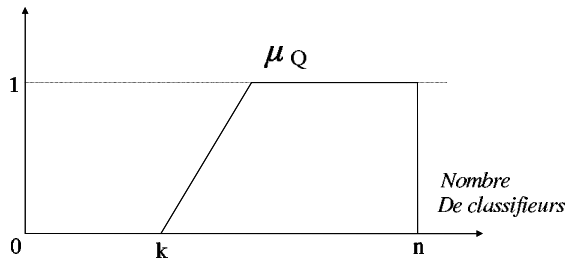


Figure 1

Les règles de majorité peuvent être implémentées en utilisant la notion de Moyenne Pondérée Ordonnée (OWA) introduite par (YAGER, 1988). L'idée est d'utiliser un ensemble de poids $p_1,..., p_n$ qui ne sont pas assignés aux classifieurs mais fonction de l'ordre partiel des scores (les probabilités d'appartenance ici) attribués par ceux-ci : les poids les plus élevés sont assignés aux classifieurs qui expriment les meilleurs scores partiels. Prenons σ une permutation sur (1,2,..,n) telle que $x^{C_{\sigma(1)}} \geq x^{C_{\sigma(2)}} \geq ... \geq x^{C_{\sigma(n)}}$ où $x^{C_{\sigma(k)}}$ est le score associé à une

phrase candidate par le classifieur $C_{\sigma(k)}$. La combinaison convexe est alors définie par :

$$\varphi(x^{C_1}, x^{C_2}, .., x^{C_n}, p_1, p_2, .., p_n) = \sum_{j=1}^{n} p_j x^{C_{\sigma(j)}}$$

La moyenne arithmétique correspond au cas $p_j = 1/n, \forall j$. On a $\varphi = \max$ quand $p_1 = 1, p_j = 0, j \geq 2$; $\varphi = \min$ quand $p_n = 1, p_j = 0, j \leq n-1$. La règle de majorité « q parmi n » est obtenue quand $p_1 = p_2 = .. = p_q = 1/q \; and \; p_{q+1} = p_{q+2} = .. = p_n = 0$.

D'autres modèles de règles de majorité sont proposés dans (ZADEH, 1983 ; KACPRIZYK, 1987).

Les règles d'unanimité restreinte contrairement aux règles de majorité restreinte, n'autorisent pas de compensation par les classifieurs Ck. C'est une altération de la règle du minimum qui stipule qu'une alternative est approuvée collectivement lorsque chaque classifieur C_k séparément approuve cette alternative : $\varphi(x^{C_1}, x^{C_2}, .., x^{C_n}, p_1, p_2, .., p_n) = \min_{j=1,n} \max(p_j, x^{C_{\sigma(j)}})$, où les poids satisfont la condition $\min_{j=1,n} p_j = 0$. L'approbation requise par q classifieurs parmi n est accomplie quand $p_1 = p_2 = .. = p_q = 0 \; et \; p_{q+1} = p_{q+2} = .. = p_n = 1$. Il est à noter que l'approbation partielle ($x^k = q/n$) par tous les Ck n'est pas identique à l'approbation complète par q classifieurs alors que ces approbations sont équivalentes dans le modèle OWA. $\varphi = \max$ quand $p_1 = 0, p_j = 1, \forall j \geq 2$ (KONING, 1990).

Quand $0 = p_1 \leq p_2 \leq .. \leq p_n$, la situation où q est mal défini peut être assimilée à l'approbation « la plupart des classifieurs C_k». « La plupart des » est vu ici comme une proportion absolue décrite par un ensemble flou Q où $\mu_Q(j) = p_j$ (KONING, 1990).

Dans ce qui suit le modèle d'agrégation des classifieurs est assimilé à une règle d'unanimité restreinte « q classifieurs parmi n » : $\varphi(x^{C_1}, x^{C_2}, .., x^{C_n}, p_1, p_2, .., p_n) = \min_{j=1,n} \max(p_j, x^{C_{\sigma(j)}})$ avec $p_1 = p_2 = .. = p_q = 0$ et $p_{q+1} = p_{q+2} = .. = p_n = 1$. Cela correspond à l'idée que si au moins q classifieurs ont donné une réponse favorable (p>0.5) pour l'attribution d'une phrase à la classe « Chirac », alors la phrase sera collectivement classée « Chirac ». L'avis des (n-q) autres classifieurs n'a pas d'importance selon cette règle qui n'autorise pas la compensation.

Nous avons étudié dans le cadre du défi DEFT05 les deux stratégies d'agrégation des votes des classifieurs : l'unanimité restreinte précédemment décrite et la plus classique moyenne arithmétique.

2.3.3 Paramètres des stratégies de vote

Concernant la moyenne arithmétique, un seul paramètre a été réglé : c'est le seuil qui détermine l'appartenance d'un candidat à une classe pour chaque classifieur Ck. Dans notre cas nous avons pris la valeur de 0.5. Ce seuil signifie que chaque classifieur stipulera que si son vote est supérieur à 0.5 le candidat correspondant appartient à la classe « Chirac ». Cette valeur de 0.5 s'explique par le fait que le classifieur lui-même calcule une probabilité d'appartenance d'un candidat à une classe et dès que la valeur de 0.5 est franchie la réponse

est considérée positive pour cette classe. Remettre en cause cette valeur change les données fondamentales du classifieur.

Concernant la règle d'agrégation d'unanimité « q classifieurs parmi n », deux paramètres ont été réglés : le seuil qui détermine l'appartenance d'un candidat à une classe pour chaque classifieur Ck, et le paramètre q. Nous avons choisis comme précédemment dans nos expérimentations un seuil de 0.5 pour tous les classifieurs. Concernant le paramètre q, nous avons choisi la valeur de 4 qui correspond au fait que 4 votes favorables à une classe sur 7 sont suffisants pour associer ce candidat à cette classe. Nous avons effectué quelques essais également avec la valeur 7 pour q, qui correspond à l'unanimité complète, mais la différence n'est pas déterminante.

2.3.4 Résultats

Nous avons appliqués les deux stratégies de vote uniquement sur les classifieurs à base de vecteurs FT. Concernant les vecteurs binaires seule la stratégie d'agrégation d'unanimité de « 4 classifieurs parmi 7 » a été appliquée.

Les performances de classification obtenues (mesure F-score) par validation croisée pour le défi tâche 1 sont :

Pour l'exécution 1, nous avons utilisés les vecteurs binaires, et nous avons appliqué la moyenne arithmétique des classifieurs, ce qui donne les résultats suivants :

Classe « Chirac » : précision 0.807, rappel 0.767, F-Score : 0.786

Classe « Mitterand » : précision 0.789, rappel 0.826, F-Score : 0.807

Pour l'exécution 2, nous avons utilisés les vecteurs FT, nous avons appliqué la moyenne arithmétique des classifieurs, ce qui donne les résultats suivants :

Classe « Chirac » : précision 0.808, rappel 0.770, F-Score : 0.789

Classe « Mitterand » : précision 0.781, rappel 0.817, F-Score : 0.799

Pour l'exécution 3, nous avons utilisés une moyenne arithmétique des 14 classifieurs constitués par les deux jeux d'essais précédents.

Les performances de classification obtenues (mesure F-score) par validation croisée pour le défi tâche 2 sont :

Pour l'exécution 1, nous avons utilisés les vecteurs binaires, nous avons appliqué la moyenne arithmétique des classifieurs, ce qui donne les résultats suivants :

Classe « Chirac » : précision 0.820, rappel 0.761, F-Score : 0.788

Classe « Mitterand » : précision 0.793, rappel 0.841, F-Score : 0.816

Pour l'exécution 2, nous avons utilisés les vecteurs FT, nous avons appliqué la moyenne arithmétique des classifieurs, ce qui donne les résultats suivants :

Classe « Chirac » : précision 0.821, rappel 0.768, F-Score : 0.773

Classe « Mitterand » : précision 0.782, rappel 0.785, F-Score : 0.783

Pour l'exécution 3, nous avons utilisés les vecteurs FT, et nous avons appliqué la stratégie d'unanimité restreinte des classifieurs, ce qui donne les résultats suivants :

Classe « Chirac » : précision 0.768, rappel 0.767, F-Score : 0.789

Classe « Mitterand » : précision 0.796, rappel 0.853, F-Score : 0.824

Les performances de classification obtenues (mesure F-score) par validation croisée pour le défi tâche 3 sont :

Pour l'exécution 1, nous avons utilisés les vecteurs binaires, et nous avons appliqué la stratégie d'unanimité restreinte des classifieurs, ce qui donne les résultats suivants :

Classe « Chirac » : précision 0.842, rappel 0.778, F-Score : 0.809

Classe « Mitterand » : précision 0.783, rappel 0.854, F-Score : 0.823

Pour l'exécution 2, nous avons utilisés les vecteurs FT, nous avons appliqué la moyenne arithmétique des classifieurs, ce qui donne les résultats suivants :

Classe « Chirac » : précision 0.812, rappel 0.752, F-Score : 0.781

Classe « Mitterand » : précision 0.781, rappel 0.835, F-Score : 0.807

Pour l'exécution 3, nous avons utilisés les vecteurs FT, et nous avons appliqué la stratégie d'unanimité restreinte des classifieurs, ce qui donne les résultats suivants :

Classe « Chirac » : précision 0.836, rappel 0.774, F-Score : 0.804

Classe « Mitterand » : précision 0.752, rappel 0.819, F-Score : 0.784

3 Conclusion et perspectives

Les résultats obtenus sur l'ensemble d'apprentissage montrent que les représentations vectorielles (binaire et FT) ne sont pas influentes pour cette problématique. Le modèle binaire donnant même de meilleurs résultats.

Les résultats obtenus par validation croisée, nous laissaient espérer des performances semblables sur le jeu de test. Malheureusement ce ne fut pas le cas. Pour l'instant nous n'avons pas d'explication à cette baisse importante des performances de notre approche sur le corpus de test.

D'autres expérimentations sont actuellement en cours pour essayer de comprendre cette contre performance et améliorer notre approche. En particulier, nous travaillons sur des méthodes permettant de sélectionner les mots les plus discriminants et ainsi diminuer la taille

des vecteurs de représentation. Ainsi, d'autres méthodes de classification pourraient être appliquées (arbre de décision, clustering flou, ...), combinées et évaluées.

Références

ANDREW MCCALLUM, KAMAL NIGAM (1998), A Comparison of Event Models for Naive Bayes Text Classification, AAAI-98 Workshop on Learning for Text Categorization

KACPRZYK, J. (1987). TOWARDS "HUMAN-CONSISTENT" DECISION SUPPORT SYSTEMS THROUGH COMMONSENSE KNOWLEDGE-BASED DECISION MAKING AND CONTROL MODELS: A FUZZY APPROACH. COMPUTERS AND ARTIFICIAL INTELLIGENCE, 6 (2), 97-122.

KONING, J-L. (1990). UN MÉCANISME DE GESTION DE RÈGLES DE DÉCISION ANTAGONISTES POUR LES SYSTÈMES À BASE DE CONNAISSANCES. THÈSE DE L'UNIVERSITÉ PAUL SABATIER DE TOULOUSE.

(SALTON ET AL., 1983) SALTON, G., INTRODUCTION TO MODERN INFORMATION RETRIEVAL », MCGRAW-HILL BOOK COMPANY, NEW YORK 1983

KARL-MICHAEL SCHNEIDER (2004) : A NEW FEATURE SELECTION SCORE FOR NAIVE BAYES TEXT CLASSIFICATION BASED ON KL-DIVERGENCE. OMPANION VOLUME TO THE PROCEEDINGS OF THE 42ND MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL 2004) <HTTP://WWW.ACL2004.ORG/>, BARCELONA, SPAIN, PP. 186-189, 2004

YAGER, R. (1988). ON ORDERED WEIGHTED AVERAGING AGGREGATION OPERATORS IN MULTICRITERIA DECISION-MAKING. IEEE TRANSACTION SYSTEMS, MAN AND CYBERNETICS, 18, 183-190.

ZADEH, L. (1983). A COMPUTATIONAL APPROACH TO FUZZY QUANTIFIERS IN NATURAL LANGUAGES. COMPUTERS AND MATHEMATICS WITH APPLICATIONS, 9, 149-184

Segmentation et classification : deux politiques complémentaires

Alexandre Labadié (1), Yann Romero (1), Laurianne Sitbon (2)
(1){alexandre.labadie,yann.romero}@iup.univ-avignon.fr
(2)laurianne.sitbon@univ-avignon.fr
LIA - Université d'Avignon et des pays du Vaucluse
Agroparc, BP 1228, 84911, Avignon, France

Mots-clefs : méthode d'apprentissage, segmentation thématique, chaînes lexicales, n-grammes, modèles de Bayes, algorithme de Viterbi

Keywords: learning methods, topic segmentation, lexical chains n-grams, Bayes' model, Viterbi's algorithm

Résumé Dans cet article, nous nous sommes attachés à étudier le problème de l'extraction de segments de texte non pertinents dans un texte selon deux approches distinctes. La première consiste à appliquer au texte un algorithme de segmentation sans apprentissage afin de pouvoir ensuite le classifier, bloc par bloc, à l'aide d'un modèle bayésien. La seconde approche consiste à déterminer la probabilité de chacun des orateurs à partir d'un modèle tri-gramme. Dans un second temps, nous avons appliqué un certain nombre d'optimisations relatives aux contraintes imposées.

Abstract In this paper, we tried to study the problem of extracting irrelevant text segments from a text using two different methods. The first one applies a segmentation algorithm to the text in order to classify it block by block with a BayesŠs model. The second one determines the probability of each speaker with a 3-gram model. Furthermore, we applied different optimization.

Introduction

Le problème posé par la détection de phrases de François Mitterand au sein de discours de Jacques Chirac tel que le propose DEFT05[1] peut être décomposé en deux problèmes distincts. Le premier est la détection de ruptures au sein du discours. En effet, l'insertion de segments de texte issus d'un orateur différent, qui de plus traitent d'un sujet différent, provoquent une certaine incohérence dans le discours. Nous sommes partis du principe que cette incohérence pouvait être considérée de la même manière qu'un changement de thème au sein d'un texte. Aussi nous nous sommes attachés à détecter ces "ruptures thématiques" avec les algorithmes sans apprentissage que sont C99 ou encore celui de la méthode LIA_seg. Le deuxième problème posé par cet atelier était de classifier les portions de texte obtenues par la segmentation. Ici encore, nous avons exploré deux champs de possibilité, à savoir une classification via un modèle bayésien et une classification grâce à un apprentissage par n-grammes. Tout l'enjeu de notre démarche aura été de démontrer l'efficacité de la combinaison de méthodes avec et sans apprentissage.
Nous verons donc dans cet article les résultats de ces différentes méthodes. Dans la section 1, nous présentons la combinaison d'algorithmes de segmentation avec une méthode de classification. La section 2 est consacrée à une approche par n-grammes de la classification qui ne fait appel à aucun filtrage. Nos résultats sont exposés et analysés dans la section 3. Enfin, nous tentons de dégager les enseignements de notre démarche dans la section 4.

1 Combinaison de méthodes de segmentation sans apprentissage avec une classification bayésienne

Nous avons choisi de segmenter le texte dans un premier temps afin de permettre une classification plus robuste. En effet la segmentation thématique fournit des unités de traitement plus riches et tout aussi cohérentes que les phrases. On notera que le texte en entrée des méthodes décrites dans cette section aura été au préalable traité avec l'étiqueteur LIA_tagger[2], pour ne conserver que les lemmes d'un nombre limité de catégories (noms, verbes, adjectifs et noms propres).

1.1 Une segmentation sans apprentissage : Variante de C99

Nous avons utilisé ici une variante sur l'algorithme C99 (Choi F., 2000). Nous nous appuyons donc sur l'idée de base de la méthode, à savoir que les mesures de similarité entre des segments de textes courts sont statistiquement insignifiantes, et que donc seul des classements locaux sont à considérer pour ensuite appliquer un algorithme de catégorisation sur la matrice de similarité. Dans un premier temps, une matrice de similarité est donc construite, représentant la similarité entre toutes les phrases du texte à l'aide d'une mesure de similarité classique : le cosinus. On le calcule donc pour chaque paire de phrases du texte, en utilisant chaque mot commun entre les phrases, après avoir éliminé du texte les mots inutiles et la ponctuation et avoir lemmatisé les mots conservés. Ainsi en considérant $f_{i,j}$ la fréquence du mot j dans le segment j, la similarité

[1] http://www.lri.fr/ia/fdt/DEFT05

[2] version uniquement probabiliste de ECSta (Spriet T. and El-Beze M., 1998) distribué par le LIA sous licence GPL

entre les segments x et y est :

$$sim(x,y) = \frac{\sum_j f_{x,j} \times f_{y,j}}{\sqrt{\sum_j f_{x,j}^2 \times \sum_j f_{y,j}^2}} \quad (1)$$

On effectue ensuite un « classement local », en déterminant pour chaque paire d'unités textuelles, le rang de sa mesure de similarité par rapport à ses $m \times n - 1$ voisins, m et n étant les dimensions du masque de classement choisi. Le rang est le nombre d'éléments voisins ayant une mesure de similarité plus faible, conservé sous la forme d'un ratio r afin de prendre en compte les effets de bord.

$$r = \frac{\text{rang}}{\text{nombre de voisins dans le masque}} \quad (2)$$

Une matrice de rang est ainsi créée et vient remplacer la matrice de similarité. Comme dans l'algorithme original, nous considérons la densité D des carrés, le long de la diagonale de la matrice de rang :

$$D = \frac{\sum_{k=1}^{m} s_k}{\sum_{k=1}^{m} a_k} \quad (3)$$

avec a_k l'aire du carré et s_k son poids qui est la somme des tous les rangs des phrases qu'il contient.

Là où notre méthode varie par rapport à l'algorithme original dans l'exploitation de la matrice de rang et des densités calculées. Nous calculons cette densité pour chaque carré de la diagonale ayant pour coin inférieur la limite inférieure de la diagonale.

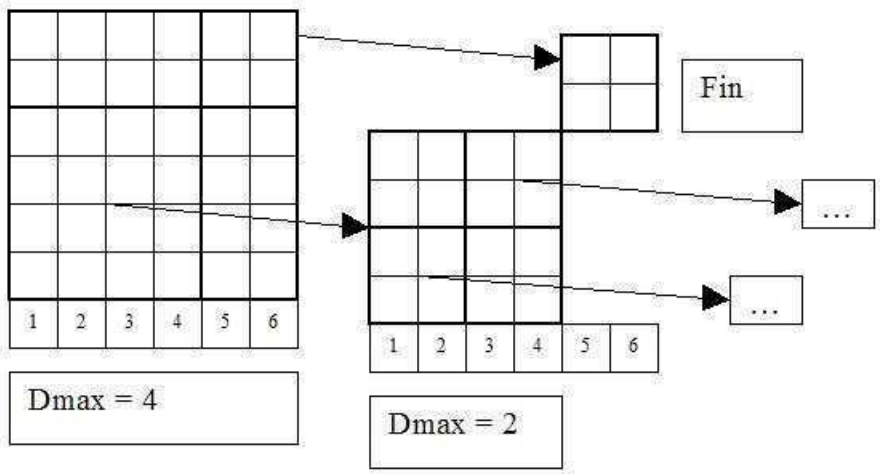


Figure 1: algorithme récursif appliqué à la matrice de rang

La plus forte densité obtenue indique le segment où il y a le plus de chance qu'une rupture dans la logique du discours se soit produite. Nous pouvons donc considérer les deux sous matrices de rang ainsi obtenues le long de la diagonale observée et répéter la même opération de manière récursive sur les matrices suivantes jusqu'à ce que le segment frontière donné par le calcul des densités soit la limite supérieure de la diagonale de la matrice courante ou jusqu'à ce que le nombre de segments contenus dans la matrice soit égal à deux.

1.2 Segmentation avec LIA_seg

L'outil LIA_seg utilisé pour effectuer la segmentation est détaillé et évalué dans (Sitbon L. and Bellot P., 2005). La segmentation s'effectue à l'aide de chaînes lexicales pondérées.

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. La figure 2 illustre le processus de création des chaînes. On peut alors recenser pour chaque phrase les chaînes actives.

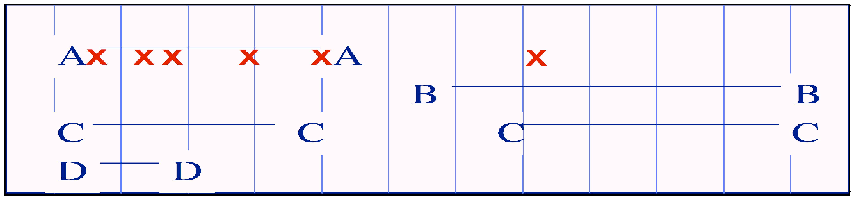


Figure 2: Construction des chaînes lexicales

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). (M. Galley et al., 2003) proposent une pondération des chaînes en fonction de la compacité et de la fréquence du terme considéré, avec un hiatus d'une distance de 11 phrases défini empiriquement. Le poids d'une chaîne associée à un terme m est défini par :

$$score(Chaine, m) = freq(Chaine, m) \times log(\frac{L_{texte}}{L_{chaine}}) \quad (4)$$

Où $freq(Chaine, m)$ est le nombre d'occurrences du terme m dans la chaîne, L_{texte} la longueur du texte, et L_{chaine} la longueur de la chaîne (on ne peut pas dépendre de la taille des textes à segmenter).

Puis on calcule les similarités à chaque fin de phrase, considérée comme une rupture thématique potentielle. La similarité est calculée avec :

$$sim(A, B) = \frac{\sum_m score(A, m) \times score(B, m)}{\sqrt{\sum_m score(A, m) \times \sum_m score(B, m)}} \quad (5)$$

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 dont la similarité est localement la plus basse (inférieure aux 3 fins de phrases précédentes et suivantes) et en dessous d'un seuil déterminé par :

$$sim_{limit} = \mu + \frac{\sigma}{2} \quad (6)$$

où μ et σ sont la moyenne et la variance de toutes les similarités calculées (M. Galley et al., 2003).

1.3 Une classification à partir d'un modèle bayésien simple

L'apprentissage se fait seulement sur les mots jugés utiles à savoir les noms communs, les noms propres, les verbes et les adjectifs.

Afin de pouvoir attribuer les segments de texte à l'un ou l'autre des orateurs, nous nous sommes appuyés sur un modèle probabiliste de type bayésien. Ainsi, la probabilité d'un mot pour un orateur O est la fréquence de ce mot pour O sur le nombre total de mots de O (ici w_O est le nombre d'apparitions du mot w pour l'orateur O et N_O le nombre total de mots observés pour l'orateur O).

$$P(w|O) = \frac{w_O}{N_O} \tag{7}$$

De fait, la probabilité qu'un segment ait été émis par l'orateur O est le produit des probabilités de chacun des mots du segment pour O (on fait l'hypothèse que les $P(w_i|O)$ sont indépendants les uns des autres).

$$P(s|O) \approx \prod_i P(w_i|O) \tag{8}$$

Cette probabilité n'est toutefois pas utilisée telle quelle. En effet, on intègre à cette probabilité la notion de longueur de phrase au travers d'une loi normale calculée sur la moyenne et l'écart type de la longueur des segments pour chacun des présidents. Ainsi $N_O(x)$ est la probabilité qu'un segment de longueur x ait été émis par l'orateur O. On a donc :

$$P'(s|O) = P(s|O) \times N_O(x) \tag{9}$$

Ensuite, elle est normalisée de telle sorte que $\sum_i P(O_i|s) = 1$. Au cours de l'apprentissage, on apprend également les probabilités de trigrammes de mots pour chaque orateur que l'on emploiera plus loin lors des post-traitements.

$$P(w_1, w_2, w_3|O) = \frac{P(w_1, w_2, w_3, O)}{P(w_1, w_2, w_3)} \tag{10}$$

1.4 La jonction des deux méthodes

L'objectif de notre démarche était de combiner une méthode sans apprentissage de segmentation avec une méthode de classification sujette à apprentissage. Aussi plutôt que d'étiqueter chaque segment indépendamment, nous avons tiré parti de la segmentation en utilisant le classifieur sur les « groupes de segments contigus » obtenus lors de la segmentation. Ainsi, on calcule la probabilité qu'un groupe G soit issu d'un orateur O pour chacun des orateurs.

$$P(G|O) \approx \prod_i P'(s_i|O) \tag{11}$$

On effectue ensuite une normalisation afin que $\sum_i P(O_i|G) = 1$ et l'orateur qui obtient ainsi la meilleure probabilité donne son étiquette à l'ensemble du groupe.

1.5 Viterbi en post-traitement

La dernière étape de cette approche est un post-traitement à partir d'un automate de Markov avec lequel on applique l'algorithme de Viterbi. Pour cet automate, on prend pour probabilités d'émission et de transition non seulement les probabilités de chacun des orateurs pour un bloc, mais aussi les probabilités de continuité d'un bloc à l'autre. Si s_1 est le segment frontière du bloc supérieur et s_2 celui du bloc inférieur et si s_1 contient n mots pleins alors la probabilité de continuité pour l'orateur O entre s_1 et s_2 (et donc entre les deux blocs de segments) est :

$$P(\textit{continuit}|O) = \max\{P(w_{s_1,n-1}, w_{s_1,n}, w_{s_2,1}|O), P(w_{s_1,n}, w_{s_2,1}, w_{s_2,2}|O)\} \tag{12}$$

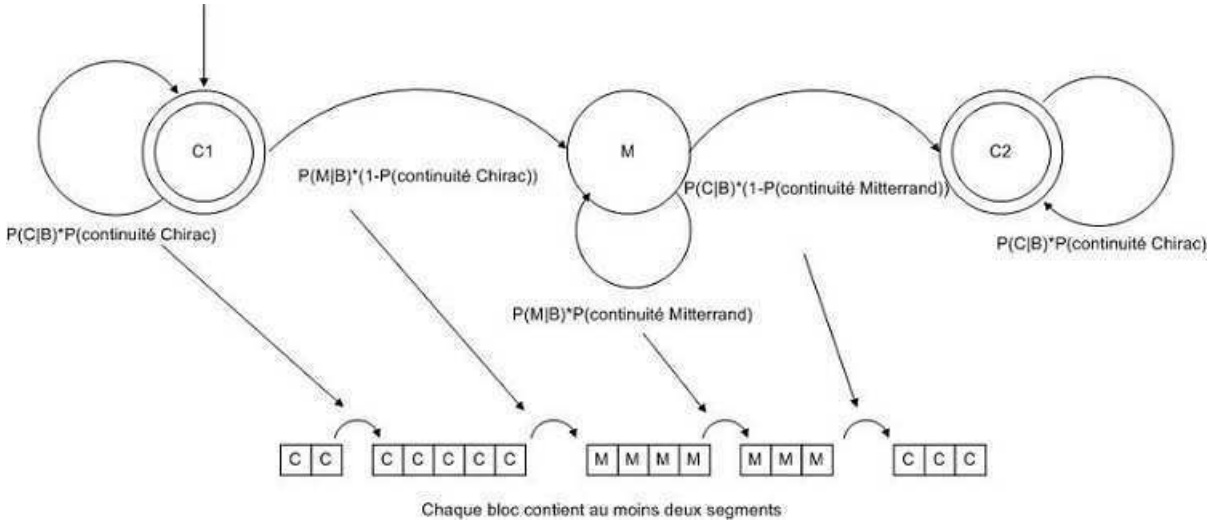


Figure 3: Automate de Markov utilisé dans le post-traitement

On obtient ainsi l'automate représenté en figure 3.

L'automate que nous utilisons devant permettre de trouver le meilleur chemin parmi des groupes de segments d'au moins deux segments chacun, sa forme est relativement simple. Ainsi il ne comporte que trois états et suppose que l'état initial et l'état final soient tous les deux des « blocs Chirac ». Ce postulat peut sembler faux, en effet rien indique que l'on commence ou finisse obligatoirement par un ensemble de segments de Chirac dans l'énoncé de l'atelier. Pourtant c'est la forme de l'automate qui a donné les meilleurs résultats, du fait probablement, qu'elle correspond plus à la réalité du corpus.

2 Apprentissage et optimisation

Cette méthode consiste à comparer les probabilités des n-grammes trouvés en apprentissage avec ceux détectés lors du test. Nous n'effectuons aucun filtrage, seuls les signes de ponctuations sont supprimés. En effet, le but de la méthode est de faire la distinction entre des discours de Chirac et de Mitterrand, donc on se focalise seulement sur les mots.

2.1 L'apprentissage par n-gramme

Nous utilisons le corpus d'apprentissage remis par DEFT. Le corpus est divisé en deux parties : 80 % pour l'apprentissage et 20 % pour le test. Nous résultats sont présentés en section.

Sur le corpus d'apprentissage, nous calculons les probabilités d'avoir un n-gramme. Ces probabilités sont évaluées pour Chirac et Mitterrand. Puis, pour chaque type de n-gramme (unigramme, bigramme, ...) on détermine une probabilité minimale entre Chirac et Mitterand. Cette probabilité minimale sera utilisée lors du processus de test.

La classification se fera sur chaque phrase du corpus de test. Premièrement, nous déterminons les n-grammes de la phrase. Puis nous recherchons si ces n-grammes ont été trouvés en phase d'apprentissage. Trois cas peuvent se produire :

- Aucun n-gramme n'a été trouvé. Dans ce cas, rien n'est fait.
- Un n-gramme a été trouvé chez Chirac ou Mitterrand. On ajoute au score du vainqueur, la probabilité du n-gramme calculé en phase d'apprentissage. On ajoute au score du perdant la probabilité minimale. Ceci évite que le perdant soit mis à l'écart.
- Le n-gramme a été trouvé chez Chirac et Mitterrand. Dans ce cas, on ajoute au score des deux classes les probabilités respectives.

Une fois la détection achevée, les scores sont normalisés par le maximum des deux. Puis, on récupère l'exponentielle de chaque score. On effectue encore une normalisation par la somme des n-grammes entre Chirac et Mitterrand. Puis on applique la formule suivante :

$$f_n = Score_{ngrammeChirac} - Score_{ngrammeMitterand} \; avec \; n = \{1,2,3\} \tag{13}$$

$$f = \sum_{n=1}^{3} \lambda_n * f_n \tag{14}$$

Si $f < \varepsilon$, alors la phrase appartient à Chirac Sinon elle appartient à Mitterand. ε vaut -0.025. λ_n varie en fonction des n-grammes. λ_1 correspond au modèle unigramme, λ_2 correspond au modèle bigramme et λ_3 correspond au modèle trigramme.

Unigramme λ_1	Bigramme λ_2	Trigramme λ_3
0.56	0.4	0.1

Table 1: Valeurs des lambdas en fonctions de n-grams

Les lambdas ont été déterminés par approches successives.

2.2 Optimisation par alternance et par contrainte

L'optimisation par alternance est une méthode triviale qui consiste à transformer les phrases étiquetées Chirac en Mitterrand si elles se trouvent entre deux phrases de Mitterrand. Le cas inverse est également traité.

L'optimisation par contrainte est une méthode qui consiste à étiqueter les deux premières phrases et les deux dernières phrases d'un discours comme étant du Chirac. Cette méthode résulte des observations effectuées sur le corpus d'apprentissage. Le corpus débute et se termine en moyenne par deux phrases de Chirac.

3 Expériences

Les trois exécutions présentées à l'atelier correspondent aux trois approches décrites précédement. La table 2 récapitule les résultats obtenus pour chacune des tâches de l'atelier avec chacune des méthodes.

	Tache 1	Tache 2	Tache 3
C99 + Bayes	0.759816	0.741895	0.745863
LIA_seg + Bayes	0.668947	0.657491	0.65957
3-grammes	0.727665	0.731987	0.741702

Table 2: Récapitulatif des résultats officiels de l'équipe junior du LIA

3.1 C99 + Bayes

La combinaison de notre variante de C99 et d'une classification bayésienne nous a donné les meilleurs résultats avec des valeurs autour de 0.75 lorsqu'un filtrage est appliqué sur l'apprentissage bayésien. Cette valeur se dégrade nettement sans filtrage (table 3).

	Développement	Développement sans filtrage	Test	Test sans filtrage
tâche 1	0.74697	0.715526	0.759816	0.740328

Table 3: Les résultats de C99 avec une classification bayésienne sur la tâche 1

Même si ces résultats ne sont pas présentés ici, on notera que les meilleures performances de la segmentation sont obtenues avec un filtrage quasi inexistant. En effet lors de nos expériences nous avons constaté une baisse de performance de la segmentation lorsque l'on réduisait le champ des mots utilisés. On peut donc déduire que contrairement au cas où l'on souhaite trouver des frontières entre thèmes, pour différencier deux orateurs les mots outils et la ponctuation sont utiles. Cela peut facilement s'expliquer par le fait que les deux orateurs ont des habitudes, des « manies » dans leur façon de s'exprimer qui se manifestent par l'emploi de certaines formes riches en mots outils. On remarquera que les résultats sont tirés vers le bas par la méthode de classification. En effet, si une classification idéale[3] avait été appliquée sur les résultats de la segmentation du corpus de test le F_{score} obtenu aurait été de 0.931596 sur la tâche 1 avec en moyenne 3.9 phrases par segment.

3.2 LIA_seg + Bayes

Afin de pouvoir généraliser nos résultats, nous avons procédé aux expérimentations avec LIA_seg sans optimisations contextuelles. Elles auraient pu consister à ajouter des frontières systématiquement aux changements de discours par exemple. De même nous avons utilisé les paramètres standards de LIA_seg, au lieu de les déterminer de manière empirique. Ceci explique en partie les meilleures performances de la méthode précédente. Nous nous attacheront ici à montrer comment nous aurions pu optimiser nos résultats, et quelle est la part de réussite liée à l'utilisation de la segmentation.

Nous avons évalué la pertinence de la méthode de segmentation en calculant le F_{score} "en cas idéal", c'est à dire si la classification des segments est optimale. Pour cela, on établit la liste des changements de locuteur dans le corpus contenant les index. On les compare aux frontières calculées par le segmenteur : pour chaque changement, on recherche la frontière calculée la plus proche. Si elle est avant un passage de Chirac vers Mitterand (C->M) ou après un passage M -> C, on considère que les phrases seront ajoutées, et on ajoute la différence dans M_{aj}. Si elle est après un passage C -> M ou avant un passage M ->, on considère que les phrases seront

[3]Classification qui donnerait la bonne étiquette à chaque segment avec une probabilité de 100 %

oubliées à l'extraction, et on ajoute la différence dans M_{oub}. Le nombre total de phrases de Mitterand dans l'index M_{tot} est compté dans l'index directement.

le F_{score} défini par :

$$F_{score} = \frac{2 \times M_{corrects}}{M_{tot} + Nb_{segments}} \tag{15}$$

devient alors :

$$F_{score} = \frac{2 \times (M_{tot} - M_{oublis})}{(M_{tot} + M_{ajouts} - M_{oublis}) + M_{tot}} \tag{16}$$

Les résultats de ce calcul du F_{score} en cas idéal sur DEFT05 sont présentés dans la table 4.

	tâche1	tâche2	tache3
F_{score}	0,909762	0.909473	0,908757

Table 4: Calcul du F_{score} en cas de classification idéale des segments thématiques.

Les résultats en cas idéal montrent que le classifieur n'est peut-être pas tout à fait adapté aux informations qu'on lui fournit. On pourrait améliorer la classification en proposant des segments plus grands avec LIA_seg, c'est à dire en augmentant le seuil de décision de rupture 6. Mais alors la fiabilité des frontières trouvées et le F_{score} idéal baisseraient également. Une étude empirique permettrait de déterminer le meilleur ratio.

Nous avons essayé une classification avec les SVM (Support Vector Machines), qui donnait de bons résultats sur les segments de référence (en apprenant sur la moitié du corpus d'apprentissage, avec une matrice creuse contenant l'index des lemmes présents dans chaque segment et leur nombre d'occurrences, avec un noyau gausssien (std = 2), et les paramètres par défaut de SVM-Torch, on avoisinait les 92% de bonne classification pour les segments de la deuxième moitié du corpus d'apprentissage). Cependant la méthode ne parvenait pas à classifier les segments issus de notre méthode, peut être parce qu'ils sont trop petits (6,7 phrases en moyenne) .

3.3 n-grammes, probabilités et optimisations

	Appr. par n-grammes	Opt. par alternance	Opt. par contrainte
tâche 1	0.6152	0.7253	0.727663
tâche 2	0.6230	0.7298	0.731987
tâche 3	0.6341	0.7384	0.741702

Table 5: Apprentissage par n-gramme et optimisations

La méthode d'apprentissage par alternance donne d'assez bon résultats. Couplée à l'optimiseur, nous nous retrouvons avec des résultats satisfaisants. Elle est d'autant plus représentatif qu'il y a beaucoup de termes dans la phrase à analyser. En effet, le score de la phrase contiendra une somme de probabilité. Ce score sera d'autant plus représentatif s'il y a beaucoup de termes.

$$ScoreHommePolitique_{k-gramme} = \sum_{n=1}^{n} p_k \; avec \; k = 1, 2, 3 \tag{17}$$

Le problème des phrases courtes dans les discours est résolu avec les méthodes d'optimisation et plus particulièrement celle d'alternance. La méthode n'est pas tres efficace sur la détection

de rupture entre du Chirac et du Mitterrand. En effet, les résultats montrent que la méthode d'apprentissage détecte des blocs de discours de Mitterrand. Or, d'après la règle du concours, il n'y a qu'un bloc de Mitterand dans un discours.

La méthode d'apprentissage est fiable pour une détection de phrases mais un prétraitement doit être effectué pour juger la crédibilité de la détection.

4 Conclusion

Les méthodes que nous avons testées pour le filtrage se sont révélées plutôt efficaces, si on les compare à la moyenne des résultats de DEFT. Ce sont des méthodes stochastiques fondées sur la cohésion lexicale et les modèles de langage. Cependant, la tâche proposée par DEFT contenait des changements thématiques en même temps que les changements de locuteur, il serait intéressant maintenant de tester la robustesse de nos méthodes sur d'autres types de corpus ou pour des traitements plus « fins ».
Il serait util de pouvoir tenir compte de la syntaxe des phrases, dans la segmentation comme dans la classification. En effet, la structure des phrases est sans aucun doute caractéristique d'un orateur, surtout dans un domaine tel que la politique. La possibilité d'intégrer cette notion devrait permettre d'améliorer encore les performances des méthodes employées. Enfin, une analyse fonctionnelle ainsi que la gestion des anaphores enrichirait le composant dédié au calcul de continuité.

Remerciements

Nous remercions Marc El-Bèze et Juan Manuel Torres pour leurs conseils et leur disponibilité, Frédéric Bechet qui aura consacré un peu de son temps à nous préparer les corpus et enfin Laurent Gillard qui nous aura supporté avec un grand stoïcisme dans son espace de travail.

Références

L. Sitbon and P. Bellot. (2005), Segmentation thématique par chaînes lexicales pondérées, Actes de *TALN'05* Dourdan, France.

L. Sitbon and P. Bellot. (2004), Adapting and comparig linear segmentation methods for french., Actes de *RIAO'04* Avignon, France.

M. Galley and K. McKeown and E. Folser-Lussier and H. Jing. (2003), Discourse Segmentation of Multi-Party Conversation., Actes de *ACL'03* Sapporo, Japan.

Freddy Y. Y. Choi (2000), Advances in domain independent linear text segmentation, *Proceedings of NAACL-00.*

Christopher D. Manning and Hinrich Schültz (1999), *Foundations of Statistical Natural Language Processing* The MIT Press, Cambridge, Massachusetts, London, England.

Thierry Spriet and Marc El-Beze (1998), *Introduction of Rules into a Stochastic Approach for Language Modelling* NATO ASI Series F, Keith Ponting.

Modèle de mélange multi-thématique pour la Fouille de Textes

Loïs Rigouste, Olivier Cappé et François Yvon

Ecole Nationale Supérieure des Télécommunications
(GET /CNRS UMR 5141)
46 rue Barrault, 75634 Paris Cedex 13, France
(e-mails: rigouste, cappe, yvon at enst.fr)

Mots-clefs : modèle de mélange, distributions multinomiales, algorithme de Viterbi

Keywords: mixture model, multinomial distributions, Viterbi algorithm

Résumé Dans cet article, nous montrons comment des techniques probabilistes d'analyse exploratoire peuvent être utilisées pour résoudre une tâche d'apprentissage supervisée: le DÉfi Fouille de Textes 2005. Ainsi, nous présentons une modélisation des textes par un mélange de distributions multinomiales sur les comptes de mots, chaque composante correspondant à un thème particulier. Les paramètres des distributions thématiques sont estimés grâce à l'algorithme EM. Les thèmes étant appris séparement pour chacun des deux auteurs, un produit dérivé de l'identification des thèmes est l'attribution des documents à leur auteur putatif. Dans la phase de test, nous affectons à chaque phrase une variable latente, en prenant soin de lier les thèmes à l'intérieur d'un document par un modèle de Markov caché, dont les paramètres sont fixés a priori. La détermination de la séquence thématique la plus probable pour un texte permet d'attribuer un auteur à chaque phrase.

Abstract In this contribution, we show how we used probabilistic methods from unsupervised text mining to solve a supervised task: the DÉfi Fouille de Textes 2005. Our model consists of a mixture of multinomial distributions over the word counts, each component corresponding to a different theme. Each theme is associated with one author. We apply the EM algorithm to estimate the parameters of these thematic distributions. In the testing phase, we define one latent variable per sentence and link the themes within a document with a hidden Markov model with fixed parameters. Finding the likeliest state sequence finally enables us to attribute every sentence to its most likely author.

1 Introduction

Aux côtés des méthodes classiques de classification non-supervisée, telles que l'algorithme des K-moyennes ou l'Analyse en Composantes Principales (ou une variante proche : l'Analyse Sémantique Latente (LSA, Deerwester *et al.*, 90)), des méthodes probabilistes ont trouvé leur place pour l'analyse exploratoire de données textuelles, les modèles plus populaires étant probablement *Probalistic latent semantic analysis* (PLSA, Hofmann, 01) et *Latent Dirichlet Allocation* (LDA, Blei *et al.*, 02). À l'instar de ces auteurs, nous plaçons ici dans le domaine des statistiques paramétriques et utilisons un modèle de mélange dont les variables latentes ont une interprétation *thématique*. Les paramètres de ces modèles ont une ainsi une interprétation simple, et l'on peut associer à chaque thème une distribution sur le vocabulaire qui identifie les mots les plus représentatifs pour ce thème.

Pour contourner les inconvénients de ces modèles, essentiellement liés à leur complexité, nous considérons ici un modèle plus simple (Nigam *et al.*, 00; Clérot *et al.*, 04), dans lequel chaque document est supposé monothématique. Après avoir présenté ce modèle, nous donnons les équations d'estimation du Maximum A Posteriori, via l'algorithme Expectation Maximization (EM). Nous évoquons ensuite quelques résultats qui montrent l'importance de l'initialisation et suggérons une méthode heuristique pour l'inférence des paramètres, qui est celle que nous avons utilisée pour le "DÉfi Fouille de Textes" (DEFT).

Dans la deuxième partie de l'article, nous expliquons comment utiliser ce modèle pour DEFT, en identifiant des thèmes dans les discours de chaque locuteur. Nous introduisons une variable latente de thème par *phrase*. Le lien entre la variable indicatrice du thème d'une phrase et celle de la phrase suivante est réalisé par un modèle de Markov caché, dont les paramètres sont supposés connus. L'algorithme de Viterbi permet ensuite de proposer une séquence d'états la plus vraisemblable et, par suite, l'auteur probable de chaque phrase.

Finalement, nous étudierons les résultats obtenus par le modèle proposé et ses variantes avant de conclure sur les améliorations possibles et travaux à venir.

2 Modèle de mélange de multinomiales

2.1 Préliminaires et notations

Pour représenter les textes, nous adoptons le modèle du sac-de-mots, c'est-à-dire que le vocabulaire est connu et fini et que chaque document est représenté par un vecteur de comptes sur cet ensemble. On note n_D, n_D^*, n_T et n_W respectivement les nombres de documents dans les corpus d'apprentissage et de test, le nombre de thèmes (i.e. le nombre de composantes du modèle de mélange dans la section 2.2) et la taille du vocabulaire.

Pour $d \in \{1, \dots, n_D\}$, $d^* \in \{1, \dots, n_D^*\}$ et $w \in \{1, \dots, n_W\}$, on note $C_d(w)$ et $C_{d^*}^*(w)$ les termes généraux des matrices de comptes d'entraînement et de test, c'est-à-dire les nombres d'occurrences du mot w dans les documents numéros d et d^*. On note également $l_d = \sum_{w=1}^{n_W} C_d(w)$ le nombre de mots dans le texte d et $l = \sum_{d=1}^{n_D} l_d$ le nombre total de mots dans le corpus d'entraînement, somme de tous les termes de la matrice de comptes. On définit similairement $l_{d^*}^*$ et l^* pour le corpus de test.

2.2 Modèle génératif

Contrairement au cadre de la catégorisation supervisée, on considère ici que l'on ne dispose d'aucune information sur les classes. Le modèle de génération du corpus que nous présentons nous permet, après estimation des paramètres, de proposer un classement des documents suivant les différentes composantes du mélange.

On suppose que les textes sont indépendants. Chaque document (numéroté $d \in \{1, \ldots, n_D\}$) résulte de l_d tirages indépendants sur le vocabulaire selon une distribution dépendant du thème; ce dernier étant défini par une variable cachée tirée une fois par texte. D'où le modèle génératif pour un document:

- Tirer un thème $t \sim \text{Mult}(1, (\alpha_1, \ldots, \alpha_{n_T}))$[1] où les $\alpha_{t'}$ sont des paramètres tels que $\sum_{t'=1}^{n_T} \alpha_{t'} = 1$.
- Tirer l_d mots $C_d = (C_{d1}, \ldots, C_{dn_W}) \sim \text{Mult}(l_d, (\beta_{1t}, \ldots, \beta_{n_W t}))$, β étant une matrice $n_W \times n_T$ de paramètres telle que $\forall t' \in \{1, \ldots, n_T\}, \sum_{w=1}^{n_W} \beta_{wt'} = 1$.

La probabilité d'un document est alors, en notant T_d la variable indicatrice du thème latent:

$$\begin{aligned} p(C_d; \alpha, \beta) &= \sum_{t=1}^{n_T} p(T_d = t; \alpha, \beta) p(C_{d1}, \ldots, C_{dn_W} | T_d = t; \alpha, \beta) \\ &= \sum_{t=1}^{n_T} p(T_d = t; \alpha, \beta) N_d! \prod_{w=1}^{n_W} \frac{p(w|T_d = t; \alpha, \beta)^{C_d(w)}}{C_d(w)!} \\ &= \frac{N_d!}{\prod_{w=1}^{n_W} C_d(w)!} \sum_{t=1}^{n_T} \alpha_t \prod_{w=1}^{n_W} \beta_{wt}^{C_d(w)} \end{aligned}$$

La probabilité du corpus, ou vraisemblance des observations, est obtenue en réalisant le produit de l'expression ci-dessus pour l'ensemble des documents étudiés. Cependant, il n'est pas possible d'établir directement une expression d'un estimateur de maximum de vraisemblance. On fait appel à l'algorithme EM (Expectation Maximization) dans lequel on s'intéresse à l'espérance, conditionnellement aux observations, de la log-vraisemblance *complète* $\mathcal{L}^c$, c'est-à-dire la log-vraisemblance des couples (vecteurs de comptes, thème) en supposant que le thème correspondant au texte d est t_d.

$$\begin{aligned} \mathcal{L}^c &= \sum_{d=1}^{n_D} \log p(C_d, T_d = t_d) \\ &= \sum_{d=1}^{n_D} \Big((\log p(T_d = t_d) + \log p(C_d | T_d = t_d) \Big) \\ &= \sum_{d=1}^{n_D} \left(\log \alpha_{t_d} + \sum_{w=1}^{n_W} \log \beta_{wt_d}^{C_d(w)} + K \right) \\ &= \sum_{d=1}^{n_D} \sum_{t=1}^{n_T} 1_{\{t_d = t\}} \left(\log \alpha_t + \sum_{w=1}^{n_W} C_d(w) \log \beta_{wt} + K \right) \end{aligned}$$

[1] On note $\text{Mult}(k, (\alpha_1, \ldots, \alpha_{n_T}))$ l'opération consistant à tirer k fois suivant une multinomiale de probabilités $(\alpha_1, \ldots, \alpha_{n_T})$.

où K est une constante indépendante des paramètres (que nous oublierons par la suite). La notation 1_A désigne la fonction indicatrice définie par:

$$1_A = \begin{cases} 1 & \text{si } A \text{ est vrai ;} \\ 0 & \text{sinon.} \end{cases}$$

L'espérance, conditionnellement aux observations, et tenant compte des paramètres α', β' issus de l'itération précédente, s'écrit:

$$E[\mathcal{L}^c] = \sum_{d=1}^{n_D} \sum_{t=1}^{n_T} p(T_d = t|C_d; \alpha', \beta') \times \left(\log \alpha_t + \sum_{w=1}^{n_W} C_d(w) \log \beta_{wt} \right)$$

Les probabilités *a posteriori* sont données par la formule de Bayes, conduisant, pour $t \in \{1, \ldots, n_T\}, d \in \{1, \ldots, n_D\}$, à:

$$\begin{aligned} p(T_d = t|C_d; \alpha', \beta') &= \frac{p(C_d|T_d = t; \alpha', \beta')p(T_d = t; \alpha', \beta')}{p(C_d; \alpha', \beta')} \\ &= \frac{p(C_d|T_d = t; \alpha', \beta')p(T_d = t; \alpha', \beta')}{\sum_{t'=1}^{n_T} p(C_d|T_d = t'; \alpha', \beta')p(T_d = t'; \alpha', \beta')} \\ &= \frac{\alpha'_t \prod_{w=1}^{n_W} \beta'_{wt}{}^{C_d(w)}}{\sum_{t'=1}^{n_T} \alpha'_{t'} \prod_{w=1}^{n_W} \beta'_{wt'}{}^{C_d(w)}} \end{aligned} \quad (1)$$

Il est alors possible de déterminer les équations de ré-estimation des paramètres en maximisant la quantité de l'EM, avec la technique des multiplicateurs de Lagrange pour normaliser de façon appropriée. Donc, pour $t \in \{1, \ldots, n_T\}$ et $w \in \{1, \ldots, n_W\}$:

$$\alpha_t = \frac{1}{n_D} \sum_{d=1}^{n_D} p(T_d = t|C_d; \alpha', \beta') \quad (2)$$

$$\beta_{wt} = \frac{\sum_{d=1}^{n_D} C_d(w) p(T_d = t|C_d; \alpha', \beta')}{\sum_{w=1}^{n_W} \sum_{d=1}^{n_D} C_d(w) p(T_d = t|C_d; \alpha', \beta')} \quad (3)$$

Ces formules sont appliquées de façon itérative jusqu'à convergence. Lorsqu'un mot w n'est jamais observé dans un thème t, ces formules conduisent à une estimation nulle pour β_{wt}. Il est alors nécessaire de recourir à des techniques de lissage des estimateurs. Dans la suite, nous utilisons un lissage de Laplace, consistant à augmenter tous les comptes de 0.1, ce qui revient à mettre sur les paramètres β une distribution *a priori* suivant une loi Dirichlet de paramètre 1.1.

2.3 Méthode de construction itérative par ajout de mots rares

Les équations de ré-estimation posées, il reste encore une marge de manœuvre importante pour un expérimentateur désirant inférer les paramètres du modèle. Des questions pertinentes concernent notamment:

- le choix du vocabulaire: faut-il considérer le vocabulaire en entier ou retirer les mots trop rares ou trop fréquents ?
- l'initialisation du modèle

Ces interrogations sont étudiées en détail dans (Rigouste *et al.*, 05), dont nous résumons ici les conclusions qui nous semblent pertinentes pour DEFT. Le corpus qui a servi de base à ces expérimentations est un corpus raisonnablement simple, issu de Reuters 2000[2] et composé de 5000 textes équirépartis dans 5 catégories (arts, sports, emploi, catastrophes, santé). En plus de la log-vraisemblance, nous considérons deux mesures des performances:

- La *perplexité*, qui quantifie la capacité du modèle à prédire de nouvelles données.
- *L'information mutuelle* entre le classement produit par le modèle et les catégories Reuters pré-existantes. Ce critère mesure plus directement la faculté de l'algorithme à retrouver les regroupements d'origine.

Nos expériences nous ont permis de mettre en évidence le fait que la phase d'initialisation de l'algorithme EM est cruciale pour l'obtention de regroupements pertinents des documents. Elles ont également confirmé l'intuition suivante: en l'absence d'information *a priori* sur les thèmes à trouver, la meilleure initialisation consiste à construire à partir de regroupements qui se recoupent largement, l'apprentissage se chargeant en général de les séparer. Dans cet esprit, l'algorithme est initialiser en fixant les probabilités *a posteriori* pour un document d'appartenir à un thème, équation (1), très proches de l'équiprobabilité entre tous les thèmes. Pour chaque essai, on tire donc ces valeurs selon une distribution Dirichlet de variance faible.

Afin d'avoir une idée de la meilleure performance possible, nous avons également essayé d'introduire l'information de supervision disponible, consistant à fonder l'initialisation sur les catégories Reuters. Pour ce faire, l'étape d'initialisation donne à un document d de catégorie Reuteurs t une valeur de 1 à la probabilité *a posteriori* d'appartenir au thème t, et une valeur 0 pour tous les autres thèmes.

Ces expériences nous ont conduit à établir les constats suivants:

- La variabilité entre les deux initialisations est très forte pour toutes les mesures, log-vraisemblance, perplexité et information mutuelle.
- La log-vraisemblance est un indicateur raisonnable de la qualité finale du regroupement produit,; c'est le seul indicateur que l'on puisse obtenir dès la phase d'apprentissage.
- À moins de pouvoir les initialiser correctement (ce qui est impossible sans information de supervision), les mots rares nuisent en général à l'apprentissage et l'écart entre les deux initialisations diminue lorsque l'on réduit la taille du vocabulaire en ne conservant que les mots les plus fréquents.

Sur la base de ces observations, l'idée de la méthode d'initialisation finalement retenue est la suivante: partant d'un vocabulaire extrêmement réduit (environ 1000 mots, soit 2% du vocabulaire total) avec l'initialisation "Dirichlet", une première estimation des paramètres du modèle est obtenue. Ce procédé est répété plusieurs fois et seul le meilleur ensemble de paramètres (au sens de la log-vraisemblance finale) est conservé. La taille du vocabulaire est ensuite progressivement augmentée, en réinitialisant à chaque étape le modèle sur les probabilités *a posteriori* issues de l'étape précédente. Cette procédure est itérée jusqu'à ce que le vocabulaire complet soit finalement pris en compte.

[2]Le corpus est en anglais. Savoir si les conclusions de notre étude se transposent à un autre corpus, en français, comme nous l'avons supposé, reste une question ouverte.

Les résultats présentés dans (Rigouste *et al.*, 05) montrent que l'algorithme d'initialisation itératif parvient au final à atteindre les mêmes valeurs de vraisemblance que celles obtenues en initialisant avec les informations de supervision. L'information mutuelle est un peu moins bonne, montrant que la corrélation entre les deux indicateurs n'est pas absolue, mais se situe dans des valeurs beaucoup plus satisfaisantes qu'avec l'initalisation "Dirichlet" simple.

3 Utilisation du segmenteur en thèmes pour DEFT

L'idée directrice de notre méthode est qu'il devrait être plus facile d'identifier les ruptures thématiques entre les phrases prononcées par J. Chirac et celles de F. Mitterrand si l'on connaît précisément les différents sujets abordés par chaque locuteur. Nous pensons (et cela se confirme dans la dernière section) que le résultat sera meilleur en modélisant les discours de chaque président par plusieurs thèmes, qui lui sont propres, plutôt qu'en utilisant seulement un thème pour chaque personne.

Ainsi, nous utilisons les données d'apprentissage pour estimer les paramètres relatifs aux thèmes abordés par J. Chirac et à ceux abordés par F. Mitterrand. Une fois ces paramètres identifiés, nous utilisons l'algorithme de Viterbi sur les phrases du corpus de test pour déterminer le thème (et donc l'auteur) le plus vraisemblable pour chaque phrase.

3.1 Prétraitements

Pour chacune des trois tâches, la même série de pré-traitements des corpus a été utilisée, consistant à segmenter chaque phrase en mot, à normaliser les chiffres, à mettre tous les mots en minuscule et à supprimer toutes les marques de ponctuation.

À l'issue de ces traitements, le vocabulaire utilisé dans le modèle statistique peut être identifié: il contient toutes les formes qui apparaissent dans le corpus, y compris les mots-outils et les mots rares, soit environ 30 000 formes graphiques. Lorsqu'un document du corpus de test contient un mot qui n'apparaît pas dans le corpus d'entraînement, ce mot est simplement ignoré.

On suppose que, dans un fichier de l'ensemble d'entraînement, toutes les phrases prononcées par un président donné font partie du même thème. Par conséquent, le corpus d'entraînement pour J. Chirac est constitué en supprimant les insertions de F. Mitterrand et en agrégeant les parties de texte séparées par ces insertions. Deux passages qui appartiennent à deux documents différents dans le corpus original ne sont jamais concaténés dans le même texte. De la même manière, chaque fragment attribué à F. Mitterrand constitue un document distinct.

3.2 Description de l'algorithme

L'algorithme itératif d'estimation des paramètres décrit en section 2.3 est utilisé pour obtenir les coefficients β_{wt} correspondant aux thèmes récurrents des discours de J. Chirac. Le nombre de thèmes n_{T_C} est fixé *a priori*. On procède de même pour F. Mitterrand, avec un nombre de thèmes n_{T_M}, permettant d'obtenir au total $n_{T_C} + n_{T_M}$ distributions sur le vocabulaire, qui sont représentatives des différents sujets abordés dans les discours de J. Chirac et F. Mitterrand.

Sur les textes du corpus de test, nous n'avons d'autre choix que d'affecter une variable latente à chaque phrase puisque nous n'avons pas d'information a priori sur les ruptures thématiques. Le problème est alors d'évaluer la séquence thématique la plus probable pour chaque nouveau texte. Pour cela, on utilise un modèle de Markov caché dont les probabilités de transition sont supposées connues. Pour chaque document du corpus de test, l'"état" (c'est-à-dire le thème) le plus probable de chaque phrase est déterminé par application de l'algorithme de Viterbi.

3.3 Prise en compte des contraintes

L'algorithme précédent ne permet pas de respecter différentes contraintes qui constituent pourtant des informations intéressantes:

1. Un texte commence toujours par un thème "J. Chirac".
2. Toute transition directe entre deux thèmes du même locuteur est interdite (à l'exception des insertions, chaque texte est donc supposé monothématique).
3. Chaque fragment contient toujours au moins deux phrases (pas de phrases de F. Mitterrand isolées et au moins deux phrases de J. Chirac en début de texte).
4. L'insertion d'un fragment "F. Mitterrand" sépare deux fragments "J.Chirac" qui appartiennent au même thème.
5. Il n'y a qu'une seule insertion "F. Mitterrand" par document.

Les deux dernières conditions ne sont pas explicites dans les règles de DEFT. Après examen rapide du corpus, il nous a cependant semblé que ces idées simples devraient permettre d'obtenir de meilleurs scores. Pour le vérifier, nous testons donc 3 modèles: le modèle 1 ne tient pas compte des deux dernières contraintes ; le modèle 2 respecte la contrainte 4 mais pas 5; le modèle 3 suit toutes les contraintes ci-dessus.

L'incorporation de ces contraintes dans le modèle s'effectue en dupliquant les états de la chaîne de Markov et en adaptant au besoin les probabilités de transition. Ainsi le modèle 1 correspond à la machine à états finis représentée en figure 1, à gauche, dans l'hypothèse où $n_{T_C} = n_{T_M} = 2$. Pour obtenir le modèle 2, il faut dupliquer tous les états "F. Mitterrand" pour chaque thème de "J. Chirac", comme indiqué à droite sur la même figure.

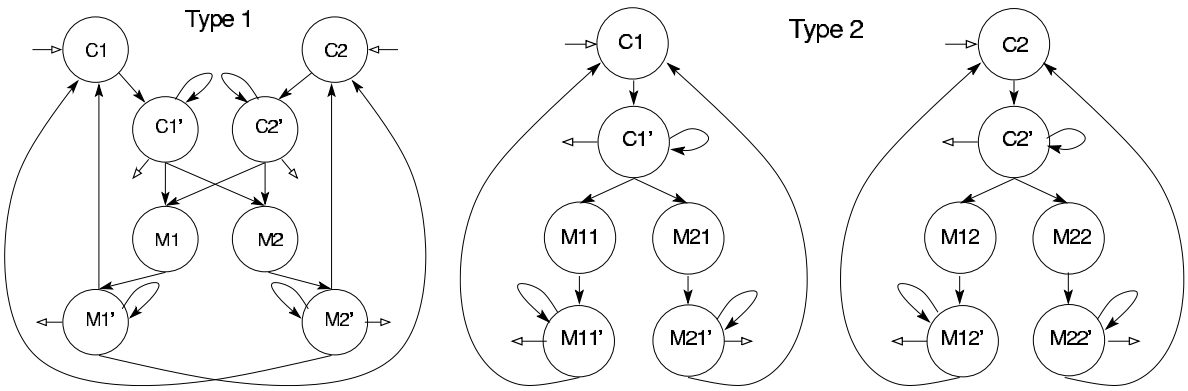


Figure 1: Modèles 1 et 2.

Enfin, dans le cas du modèle 3 (figure 2), on doit également dupliquer les thèmes "J. Chirac" pour avoir un état "pré-insertion" et un état "post-insertion".

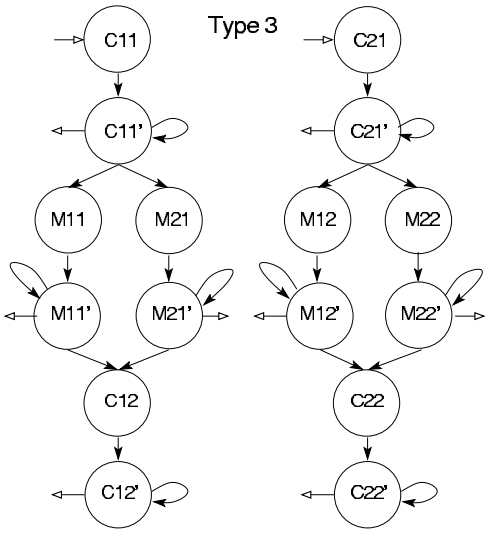


Figure 2: Modèle 3.

Les probabilités de transition et les probabilités de sortie, non figurées sur les graphes précédents, sont calculées comme suit. Pour les probabilités de transition, on multiplie les constantes de changement d'auteur (p_{C2M} et p_{M2C}, fixées à 0.3) par le paramètre α_t correspondant au thème dans lequel on entre. Les probabilités de sortie ont été en général fixées à 0, à l'exception notable du modèle 3 où les probabilités de sortie des états "J. Chirac" post-insertion doivent être augmentées et fixées à p_{C2M}. En effet, dans le cas contraire, on favorise les états post-insertion par rapport aux états pré-insertion et la vraisemblance est alors presque toujours maximisée en affectant les deux premières phrases à J. Chirac, les deux suivantes à F. Mitterrand et toutes les autres à J. Chirac (seule configuration admissible qui maximise le nombre de paragraphes dans les états post-insertion). Les probabilités de rester dans le même état (boucle) sont calculées pour que la somme des probabilités de transition pour un état donné soit égale à 1.

3.4 Résultats

Nous étudions ici uniquement les résultats sur la tâche 1 de DEFT[3]. Pour la campagne officielle de test, nous avons soumis les modèles 1 et 2 avec $n_{T_C} = 10$ et $n_{T_M} = 4$, le modèle 2 obtenant les meilleures performances. La figure 3 montre qu'il est possible d'atteindre des performances légèrement meilleures en utilisant le modèle 3. Le meilleur résultat obtenu à ce jour, de 0.88, est obtenu avec le modèle 3 en fixant $n_{T_C} = 10$ et $n_{T_M} = 3$.

Dans tous les cas, il semble que les nombres optimaux de thèmes soient de 3 pour "F. Mitterrand" et de 8 ou 10 pour "J. Chirac". Cette différence s'explique par la quantité de données d'apprentissage, bien plus importante pour un locuteur que pour l'autre, et qui permet par conséquent d'estimer de façon fiable un plus grand nombre de paramètres.

Enfin, pour évaluer l'apport de l'algorithme itératif d'initialisation utilisé pour apprendre les paramètres des thèmes, nous l'avons comparé avec les performances obtenues en utilisant une

[3]N'ayant pas cherché à tirer profit des informations spécifiques liées aux noms et aux dates, nos performances sur les tâches 2 et 3 sont quasiment les mêmes que sur la tâche 1.

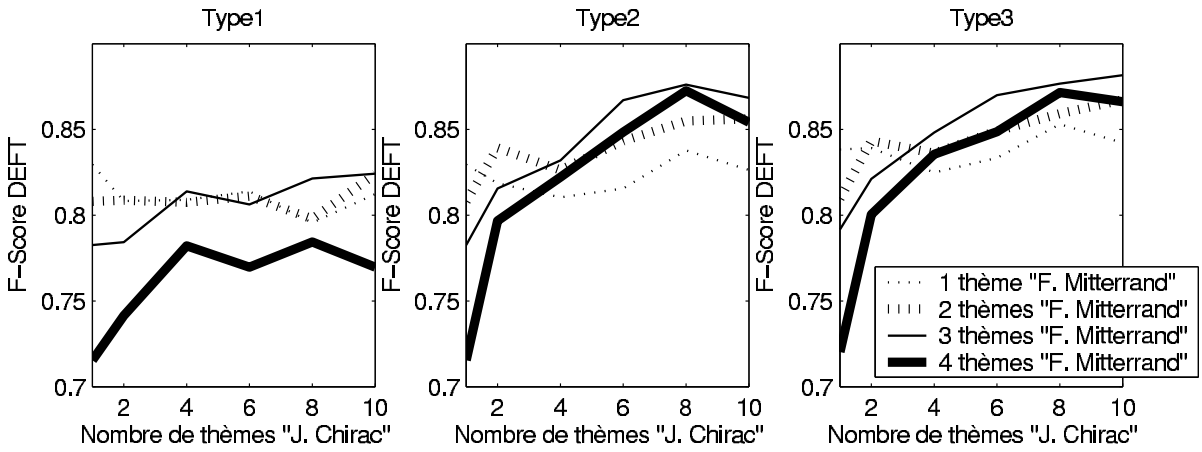


Figure 3: F-Score obtenu sur la tâche 1 de DEFT en fonction de n_{T_C}

procédure d'initialisation plus simple (initialisation "Dirichlet"), qui considère d'emblée tout le vocabulaire. Comme le montrent les résultats de la table 1, moyennés sur 50 tirages, les performances nettement meilleures obtenues par la méthode itérative sur la mesure "perplexité" se traduisent également un gain, d'ampleur plus modeste, sur la tâche d'évaluation extrinsèque de DEFT.

Méthode	Perplexité - corpus C	Perplexité - corpus M	DEFT F-Score
Init. Dirichlet	755.7±3.4	775.5±2.2	0.83±0.01
Init. Itérative	733.3±2.2	760.8±2.2	0.85±0.01

Table 1: Résultats comparés pour deux méthodes d'inférence des paramètres

4 Conclusion

Nous avons présenté dans cet article la méthode utilisée pour répondre au problème posé dans le cadre du DÉfi Fouille de Textes 2005. Il s'agit d'utiliser un modèle non supervisée d'analyse exploratoire pour une tâche de fouille de textes supervisée. En identifiant les distributions thématiques qui sous-tendent les discours de J. Chirac et F. Mitterrand dans le corpus d'entraînement, nous sommes mieux à même de catégoriser les phrases du corpus de test, en déterminant l'enchaînement thématique le plus probable.

Le fait que notre modèle n'ait pas été spécifiquement conçu pour ce travail mais y obtienne malgré tout des résultats satisfaisants démontre son efficacité à segmenter un discours en thèmes, quand bien même les thèmes obtenus sont parfois difficile à analyser. Sur la base des résultats disponibles à ce jour, il semble que comparativement aux autres méthodes, notre modèle est plus efficace pour la tâche 1 que dans les deux autres. Pour obtenir de meilleures performances sur les tâches 2 et 3, il aurait fallu ajuster les poids des dates et des noms de personnes dans le calcul de la vraisemblance de chaque phrase.

Dans le cadre des travaux à venir, nous prévoyons de tester sur d'autres tâches la combinaison de ce modèle de mélange thématique et d'un modèle de Markov caché sur l'enchaînement des variables latentes associées aux phrases ou aux paragraphes. Une autre direction intéressante de

recherche consiste à comparer la méthode itérative heuristique d'inférence présentée dans cet article avec des algorithmes plus évolués de type échantillonneur de Gibbs.

Remerciements

Ce travail est financé par France Télécom, Division R&D, sous le contrat n°42541441.

Références

Blei D., Ng A., Jordan M. (2002), Latent Dirichlet Allocation, *Advances in Neural Information Processing Systems (NIPS)*, Vol. 14, 601-608.

Clérot F., Collin O., Cappé O., Moulines E. (2004), Le Modèle "Monomaniaque" : un Modèle Statistique Simple pour l'Analyse Exploratoire d'un Corpus de Textes, *Colloque International sur la Fouille de Texte (CIFT'04).*

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, Numb. 6, 391-407.

Hofmann T. (2001), Unsupervised Learning by Probabilistic Latent Semantic Analysis, *Machine Learning Journal*, Vol. 42, Numb. 1, 177-196.

Nigam K., McCallum A., Thrun S., Mitchell T. (2000), Text Classification from Labeled and Unlabeled Documents using EM, *Machine Learning*, Vol. 39, Numb. 2/3, 103-134.

Rigouste L., Cappé O., Yvon F. (2005), Inference for Probabilistic Unsupervised Text Clustering, soumis à SSP 2005, IEEE Workshop on Statistical Signal Processing.

TALN 2005 - RECITAL 2005

12ème conférence annuelle sur le Traitement Automatique des Langues Naturelles

9ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues Naturelles

---

ATELIER

LANGUES PEU DOTÉES

TAL et Langues peu dotées

Chantal Enguehard
Laboratoire d'Informatique de Nantes Atlantique – Université de Nantes
2, rue de la Houssinière
BP 92208 44322 Nantes Cedex 03 France
chantal.enguehard@univ-nantes.

Ces dix dernière années ont vu de grands bouleversements dans l'accès aux Nouvelles Technologies. La Toile s'est considérablement étendue et recouvre maintenant la planète,touchant la quasi-totalité des cultures. Cette extension spatiale a été accompagnée par une capacité accrue à représenter les différentes langues. Alors que les caractères étaient codés grâce à un seul octet dans la représentation ASCII, le standard Unicode, apparu en 1992,définit une représentation sur 4 octets qui permet de représenter de manière unique chacun des caractères de chacune des langues. Désormais, le stockage des documents sous une forme électronique qui permet leur traitement analytique autorise de nombreuses langues à franchir la première étape de l'informatisation.

Ce progrès considérable doit être soutenu. Même si le standard Unicode tend à représenter les caractères de toutes les langues, l'inventaire de ces caractères n'est pas entièrement complété,ou bien ce standard de représentation est encore peu connu, et donc peu respecté. Deux articles détaillent ces difficultés :

- Grégory Kourilsky présente la cas de l'écriture tham du Laos qui n'a pas de représentation dans Unicode. Il émet des propositions détaillées pour remédier à cette situation.

- Wunna Ko Ko et Mikami Yoshiki inventorient huit langues du Myanmar, certaines écrites depuis des siècles. Bien que ces langues utilisent majoritairement des caractères représentés dans Unicode, leur informatisation souffre du manque de standardisation (le développement de polices de caractères non normalisées gène le partage des textes) et de coopération entre les chercheurs.

Les travaux visant à constituer des ressources linguistiques sont souvent insuffisants : soit il y a eu peu de recherches en linguistique, soit celles-ci n'ont pas produit de ressources électroniques utilisables. La constitution de telles ressources est une première étape cruciale pour fonder des travaux en Traitement Automatique des langues. Il s'agit de constituer des corpus de textes de taille importante afin d'en extraire des ressources linguistiques électroniques.

- Hubert Naets s'appuie sur un échantillon de langue pour constituer automatiquement un corpus à partir de la Toile. Cette procédure statistique permet de distinguer finement les langues voisines.

- Daniel Yacob présente les travaux menés pour constituer un corpus de l'Amharic et en déduire un lexique avec une représentation normalisée. Il détaille les aspects légaux rencontrés lors de la collecte de textes.

- Emmanuel Schang expose les apports d'un corpus oral transcrit pour saisir une langue dans son expression spontanée.

- Claudia Soria et Monica Monachini s'appuient sur des corpus en différentes langues (dont une langue linguistiquement bien dotée), et traitant d'un même domaine, pour extraire automatiquement la terminologie de ce domaine dans chacune des langues.

L'adaptation de travaux existants, la mise au point de stratégies facilement adaptables à différentes langues constituent en enjeu important pour équiper les langues en outils automatiques.

- Laurent Besacier, Viet-Bac Le, Eric Castelli, Sethserey Sam et Ludovic Protin cherchent à adapter un système de reconnaissance automatique de la parole continue à deux langues peu dotées : le vietnamien et le khmère.

- Johannes Heinecke vérifie qu'il est possible d'adapter un étiqueteur morphosyntaxique au gallois même si cette langue présente des caractéristiques supplémentaires par rapport à la langue initiale visée par l'étiqueteur.

- Frédérick Houben et François Rioult s'appuient sur des propriétés très générales des langues et des méthodes de fouilles de données pour effectuer automatiquement l'étiquetage de textes.

Il apparaît cependant que l'étude linguistique fine d'une langue est indispensable lors de certaines étapes.

- Les travaux de Bali Ranaivo-Malançon visant à construire un étiqueteur morphosyntaxique du malais en s'appuyant sur la morphologie des mots de cette langue se heurtent au manque de consensus sur le jeu d'étiquettes adéquats à utiliser pour cette langue.

Enfin, nous avons inclus un article ne traitant pas directement du Traitement Automatique des Langues mais de la localisation des logiciels. En effet, si les utilisateurs souhaitent s'exprimer dans leurs langues et bénéficier de l'apport d'outils automatiques, comme les correcteurs orthographiques par exemple, ils apprécient également que les outils électroniques s'adressent à eux en respectant leurs langue et culture.

- Dawit Bekele explique l'importance de la localisation des logiciels pour les pays du tiers-monde, puis il détaille les différents aspects techniques de la localisation.

Les recherches présentées lors de cet atelier abordent différents points de la chaîne des traitements appliqués aux langues. Elles constituent également un apport plus général pour le Traitement Automatique des Langues puisqu'elles soulignent des difficultés que peuvent rencontrer toutes les langues, y compris les langues largement dotées comme l'anglais, le français ou l'espagnol. En effet, les langues évoluent, des expressions singulières apparaissent Atelier TAL et Langues peu dotées et il faut trouver des stratégies pratiques pour faire évoluer conjointement les ressources linguistiques sur lesquelles s'appuient les traitements automatiques.

Reconnaissance Automatique de la Parole pour des Langues peu Dotées : Application au Vietnamien et au Khmer

L. Besacier (1), V.-B. Le (1), E. Castelli (2), S. Sethserey (3), L. Protin (3)

(1) Laboratoire CLIPS-IMAG, UMR CNRS 5524, BP 53,
38041 Grenoble Cedex 9, FRANCE
(Laurent.Besacier,Viet-Bac.Le)@imag.fr
(2) International Research Center MICA, 1 Dai Co Viet,
Hanoi, VIETNAM
Eric.Castelli@mica.edu.vn
(3) Institut de Technologie du Cambodge, Bd de Pochentong
BP 86 – Phnom Penh, CAMBODGE
Sam.Sethserey@itc.edu.kh , Ludovic.Protin@online.com.kh

Mots-clés : reconnaissance automatique de la parole, langues mal dotées, modèles multilingues, ressources écrites et orales multilingues.

Keywords: automatic speech recognition, under-resourced languages, multilingual models, multilingual text and speech resources.

Résumé Nous présentons dans cet article une méthodologie qui vise à développer et adapter le plus rapidement possible un système de reconnaissance automatique de la parole continue pour une nouvelle langue peu dotée. Les ressources collectées et les résultats expérimentaux obtenus pour le vietnamien sont présentés. Notre meilleur système pour cette langue obtient actuellement un taux de reconnaissance de mots de 64% environ. Les travaux en cours sur la langue khmère sont également décrits à la fin de cet article.

Abstract We present here a methodology for fast development of ASR systems for new under-resourced languages. The resources collected for vietnamese, and the experimental results of our first vietnamese ASR system are presented. Our best system obtains 64% of word accuracy rate. The current validation of our methodology for khmer language is also described at the end of this paper.

1 Introduction

L'informatique est désormais un instrument pour écrire et communiquer. Traitements de textes, courriers électroniques, voire des systèmes plus avancés comme la dictée ou la synthèse vocale sont des outils largement répandus. La possibilité d'offrir ou pas ces services

pour une langue donnée peut permettre de définir son « niveau d'informatisation » (Berment 2004).

Avant de soumettre un article à un atelier intitulé « TAL et langues peu dotées », il convient d'abord de vérifier que notre travail correspond bien au thème dudit atelier, à savoir : a) est-ce du TAL ? b) les langues traitées sont elles peu dotées ? Pour cela, la réponse peut venir de (Berment 2004) qui définit dans sa thèse la notion de langue peu dotée, en introduisant un indice σ mesurant la satisfaction des utilisateurs de logiciels et, incidemment, le niveau d'informatisation de la langue pour une classe de services donnée (si cet indice est inférieur à 10, la langue peut être considérée comme peu dotée informatiquement) :

- a) première réponse à notre interrogation, parmi les services mentionnés par (Berment, 2004) intervient le *traitement de l'oral* qui regroupe les technologies de synthèse et reconnaissance vocale. Il est peut être bon de repréciser ici, même si cela est évident, que ces technologies, bien que faisant intervenir des problèmes liés à l'acoustique et au traitement du signal, nécessitent également de résoudre des problèmes de TAL (conversion graphème-phonème pour la synthèse, et modélisation du langage pour la reconnaissance automatique de la parole, par exemple).

- b) seconde réponse à notre interrogation, les langues que nous abordons font effectivement partie de la classe des *langues mal dotées* (bien que le vietnamien soit à la limite), tel que définie par (Berment 2004). Le khmer, évalué dans sa thèse obtient un indice σ d'environ 6/20, tandis que notre évaluation du vietnamien, faite par E. Castelli et Nguyen Quoc Cuong, tous deux chercheurs du centre MICA à Hanoï, donne un indice σ de 10/20 environ. Une partie de ces différences s'explique notamment par le fait que le vietnamien, contrairement au khmer, utilise une écriture latine accentuée, rendant plus « faciles » les tâches de reconnaissance automatique de caractères et de tri par exemple.

Par ailleurs, il est important de noter que pour ces deux langues, les services liés au traitement de l'oral sont inexistants. C'est ici que se situe notre problématique. Face à la critique qui pourrait être faite concernant notre choix d'aborder des technologies peut être moins importantes, en terme de développement, que d'autres liées au traitement de texte et aux dictionnaires, nous argumenterons que développer des systèmes de traitement de la parole dans une langue « peu dotée » peut permettre de collecter des ressources utiles pouvant être ensuite remises au « pot commun » d'une langue donnée (dictionnaire phonétique, corpus oral, transcriptions de conversations spontanées par exemple).

Le présent article détaillera d'abord dans la *section 2* le contexte de nos travaux, qui a conduit au choix des deux langues abordées dans ce travail. Notre méthodologie permettant de développer et d'adapter le plus rapidement possible un système de reconnaissance automatique dans une nouvelle langue, sera décrite dans la *section 3*. La *section 4* montrera une première application de cette méthodologie au vietnamien et les quelques résultats associés (ressources obtenues et expériences de reconnaissance vocale). La section 5 décrit quand à elle nos travaux en cours sur le khmer. Nous précisons toutefois dès cette introduction que nous ne disposons pas encore d'un système de reconnaissance automatique de la parole pour la langue khmère. Nous décrirons cependant dans cette partie les ressources déjà collectées pour cette langue, ainsi que les travaux prévus dans un futur proche. Finalement, la section 6 conclura ce travail.

2 Contexte

Le Centre de recherche international MICA a été créé en 2002 pour participer au développement des technologies de l'information au Vietnam et pour répondre aux préoccupations relatives à leur évolution. Les travaux de recherche menés à Mica visent à étudier et à développer des résultats théoriques et des applications dans les thèmes du traitement des signaux complexes (parole et image), des applications multimédia et de l'instrumentation. Une collaboration existe déjà depuis la création de MICA en 2002 avec le laboratoire CLIPS sur le traitement de la langue vietnamienne.

Plus récemment, dans le cadre d'un projet financé par l'AUF[1], le Département de Génie Informatique et Communication de l'Institut de Technologie du Cambodge (ITC) s'est associé à cette collaboration. Le but de ce projet est notamment d'initier au Cambodge, la création d'un nouveau groupe de recherche spécialisé en traitement de la parole en langue khmère. Pour cela, le Centre MICA, qui a démarré ses activités en 2002 et possède désormais un groupe de recherche en parole pour la langue vietnamienne, s'est proposé d'aider l'ITC à finaliser cet objectif.

2.1 La langue vietnamienne

Elle est parlée par environ 70 millions de personnes dans le monde (source : MSN-Encarta). Son origine est toujours sujette à débat parmi les linguistes. Il est cependant généralement admis qu'elle a des racines communes et fortes avec le môn-khmer qui fait partie de la branche austro asiatique et qui comprend le mon parlé en Birmanie et le khmer, la langue cambodgienne, aussi bien que les khmu, bahnar et bru, d'autres langues parlées par les habitants des îles du nord du Vietnam. C'est une langue tonale qui possède six tons. L'orthographe est latine depuis le XVIIè siècle, avec des caractères accentués pour les tons.

2.2 La langue khmère

Elle est parlée par une dizaine de millions de personnes dans le monde (source : MSN-Encarta). Elle appartient également au groupe des langues môn-khmères. La langue khmère est une langue atonale – contrairement aux langues chinoises, thaïes ou vietnamiennes. Cependant, le khmer possède comme ses cousines austro-asiatiques plusieurs registres vocaliques : les voyelles peuvent être allongées (dites voyelles longues), raccourcies (dites voyelles brèves), diphtonguées, reposer sur des consonnes aspirées ou non aspirées, ce qui en modifie complètement le sens. (Ex. slap = mourir; slaaap = aile d'un oiseau). Cette particularité fait du cambodgien un des plus riches systèmes vocaliques au monde. Au niveau de l'écriture, pour adapter les fontes informatiques, il a fallu gérer un ordonnancement, sur plusieurs niveaux, de 33 consonnes, 32 consonnes souscrites, 28 voyelles, 14 voyelles indépendantes et 10 ligatures, sans compter les chiffres et la ponctuation.

---

[1] Projet AUF/TALK : Traitement Automatique de la Langue Khmère

3 Méthodologie

3.1 Rappel des principes généraux de la reconnaissance automatique de la parole

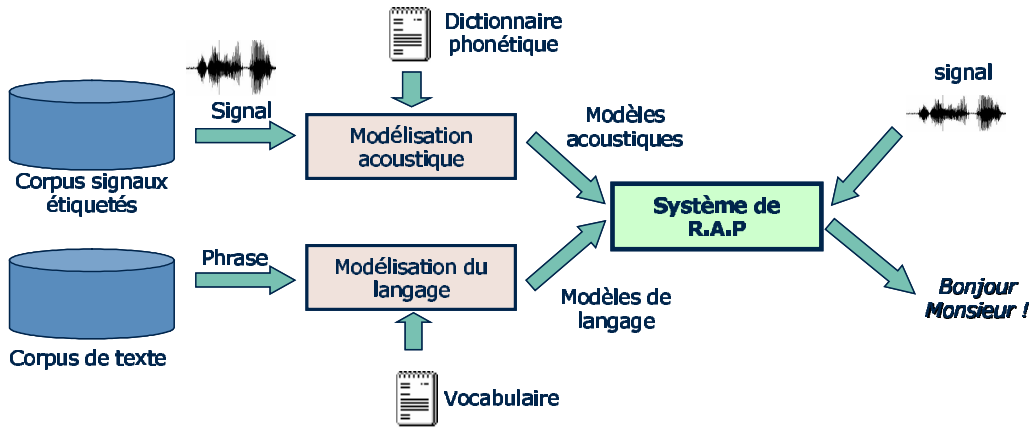


Figure 1. Schéma d'ensemble de la reconnaissance automatique de la parole

Comme l'illustre la *figure 1* ci-dessus, pour développer un système de reconnaissance automatique de la parole continue dans une nouvelle langue, il est souvent nécessaire de rassembler une grande quantité de corpus, contenant à la fois des signaux de parole (pour l'apprentissage des modèles acoustiques du système) mais également des données textuelles (pour l'apprentissage des modèles de langage du système). De tels corpus et systèmes sont désormais disponibles pour la plupart des langues occidentales (anglais, français, espagnol, etc) et pour quelques langues asiatiques (chinois, japonais, etc). Porter un système de reconnaissance vers une nouvelle langue est donc une tâche très fastidieuse si aucun corpus de grande envergure n'existe dans la langue cible, puisqu'il faut alors collecter soi-même les ressources nécessaires : signal de parole, lexique, corpus textuels, etc. Précisons aussi qu'étant donnée la nature statistique des modèles généralement utilisés en reconnaissance automatique de la parole (modèles acoustiques de phonèmes correspondant à des chaînes de Markov où chaque état est une distribution multigaussienne ; et modèles de langage N-grammes), ces ressources doivent être disponibles en quantité importante.

3.2 Collecte de ressources

Une première façon d'accélérer la portabilité des systèmes de reconnaissance automatique de parole continue grand vocabulaire vers une nouvelle langue, est de développer une méthodologie permettant une collecte rapide et/ou facilitée de ressources textuelles et acoustiques. Cette approche a l'avantage de ne pas modifier fondamentalement le cœur des techniques de reconnaissance utilisées.

Recueil de données textuelles

Concernant le recueil de données textuelles en grande quantité (*figure 2*), une approche intéressante consiste à « aspirer » un grand nombre de sites Web dans la langue donnée et à

filtrer les données récupérées pour les rendre exploitables. Ces données textuelles peuvent servir d'une part à calculer des modèles de langages statistiques, et d'autre part à obtenir un corpus pouvant ensuite être prononcé par des locuteurs en vue de la constitution d'une base de signaux conséquente.

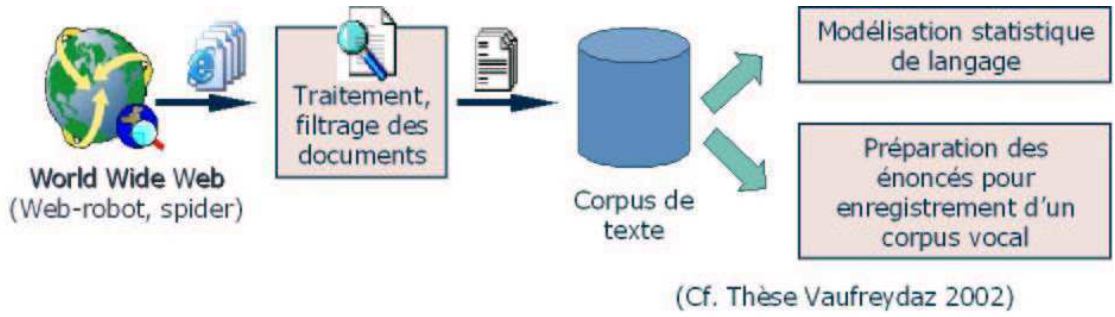


Figure 2 : récupération de données textuelles en utilisant le Web

Une telle approche a déjà été relativement bien validée pour une langue bien dotée telle que le français (Vaufreydaz, 2002). Les problèmes spécifiques pour les langues mal dotées concernent le nombre de sites web qui peut être peu important, la vitesse de transmission et la qualité des documents qui nécessitera alors plus d'outils de traitement. On préfèrera par exemple des sites de nouvelles, au fort contenu rédactionnel tels que VNexpress[2] par exemple, pour le vietnamien. Afin de rendre les données exploitables, un certain nombre de traitements sont nécessaires tels que : 1) transformation html vers texte, 2) normalisation des tags, 3) conversion des encodages (nous avons choisi de tout convertir vers une représentation interne unique utilisant l'encodage UTF-8 d'Unicode), 4) séparation en phrases et 5) en mots, 6) groupement de mots composés, 7) transcription des symboles et 8) filtrage en fonction d'un vocabulaire donné. Alors que certains traitement peuvent être considérés comme relativement indépendants de la langue cible (1-2-6-8), d'autre doivent être repensés (3-4-5-7) pour chaque nouvelle langue cible : par exemple la séparation en mots est triviale pour les écritures latines mais problématique pour d'autres systèmes d'écriture comme le khmer, surtout si l'on ne dispose pas d'un vocabulaire (i.e. une liste de mots) au départ. Une boîte à outils *open source* rassemblant quelques uns de ces outils de traitement a été développée au CLIPS[3]. Une description plus détaillée des traitements réalisés et des expérimentations associées pour la modélisation du langage en vietnamien peut être trouvée dans (Le, 2003).

Recueil de signaux de parole

Pour le recueil de signaux de parole, le CLIPS a développé un outil logiciel ne mettant en œuvre que du matériel standard : EMACOP (Environnement Multimédia pour l'Acquisition et la gestion de Corpus Parole) (Vaufreydaz, 98). La plupart du temps, les campagnes d'enregistrement mobilisent d'importantes ressources humaines pour guider ou assister les locuteurs dans leur tâche de diction, pour organiser l'enregistrement, pour préparer les scénarios et les données, etc. Il faut pouvoir contrôler les différents scénarios pour varier les

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

[3] http://www-clips.imag.fr/geod/User/viet-bac.le/outils/

conditions de capture : la lecture d'un texte ou d'une suite de mots ou de mots isolés, la répétition après écoute d'une phrase, le dialogue en réponse à des questions, etc. Les méthodes d'acquisition rigoureusement contrôlées sont donc lourdes et les difficultés sont amplifiées dans le cas des langues mal dotées où les locuteurs ne sont pas forcément rôdés à l'utilisation de moyens informatiques par exemple. C'est pourquoi, le développement d'un utilitaire portable de gestion et d'acquisition de grands corpus sur un matériel standard, nous est d'un grand bénéfice. Le logiciel respecte le format SAM de définition de bases de signaux. Les interfaces ont été adaptées pour manipuler respectivement les caractères vietnamiens et khmers. La figure 3 montre un exemple de l'interface d'EMACOP adaptée pour la langue khmère.

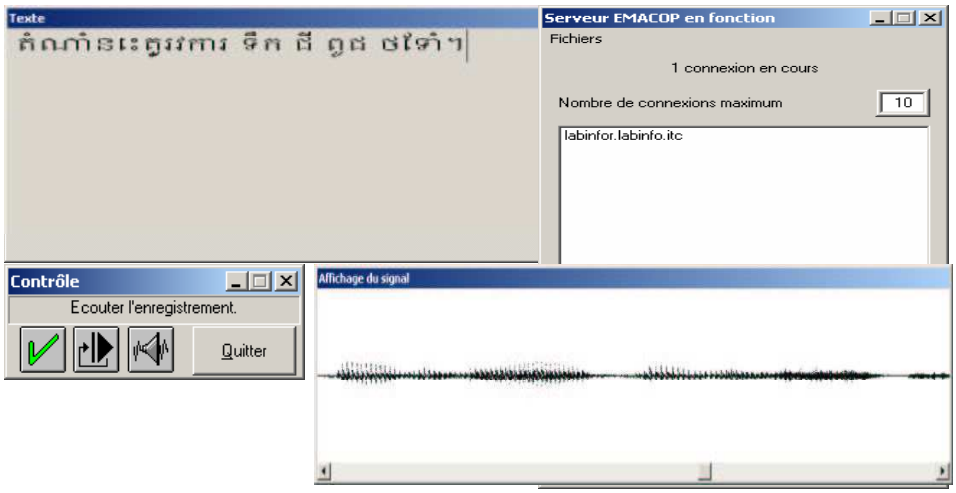


Figure 3 : Interface EMACOP adaptée pour le Khmer

3.3 Dictionnaire de prononciation

Un dictionnaire de prononciation (ou dictionnaire phonétique) est une ressource essentielle aux tâches de synthèse et de reconnaissance de la parole, ou tout simplement pour enrichir un dictionnaire bilingue, permettant au locuteur étranger de connaître la prononciation du mot en langue cible. Cette tâche est cependant difficile pour des langues mal dotées dont le système phonologique est parfois méconnu, ou sujet à débats (langues peu ou mal décrites). Si nous mettons de côté les méthodes manuelles de phonétisation qui, bien que donnant les dictionnaires de prononciation de meilleure qualité, ne nous semblent pas entrer dans le cadre de notre méthodologie, on peut distinguer deux types d'approches automatiques pour constituer un dictionnaire phonétique dans une nouvelle langue :

- Des approches à base de règles, qui nécessitent une bonne connaissance de la langue et de ses règles de phonétisation (qui par ailleurs ne doivent pas contenir trop d'exceptions). Ce type d'approche est assez coûteux en temps (écriture d'un analyseur phonétique), mais donnera des dictionnaires de prononciation de qualité très correcte pouvant ensuite être révisés manuellement relativement rapidement.

- Des approches utilisant un système de reconnaissance phonémique appliqué sur des enregistrements des mots à phonétiser, permettant un premier étiquetage automatique en phonèmes d'une liste de mots qui peut être alors révisé par un opérateur humain.

L'avantage d'une telle approche est bien sûr sa rapidité. Ses inconvénients sont qu'elle nécessite l'emploi d'un système de reconnaissance automatique de phonèmes qui sera généralement celui d'une langue source bien dotée (par exemple un système de reconnaissance des phonèmes du français) ; par ailleurs, l'autre défaut est que les unités phonémiques décrivant les mots en langue cible seront seulement celles pouvant être reconnues par le décodeur en langue source, d'où la nécessité d'employer si possible des décodeurs phonémiques multilingues pour augmenter au maximum la couverture phonémique dans l'alphabet phonétique international (API). La figure 4 illustre ce problème : la langue source utilisée est le français tandis que la langue cible est le vietnamien. Il est évident que la couverture du vietnamien par le français n'est pas du tout optimale. Une telle méthode reste cependant intéressante, notamment lorsqu'on passe d'une langue source à une langue cible qui possèdent un système phonologique proche.

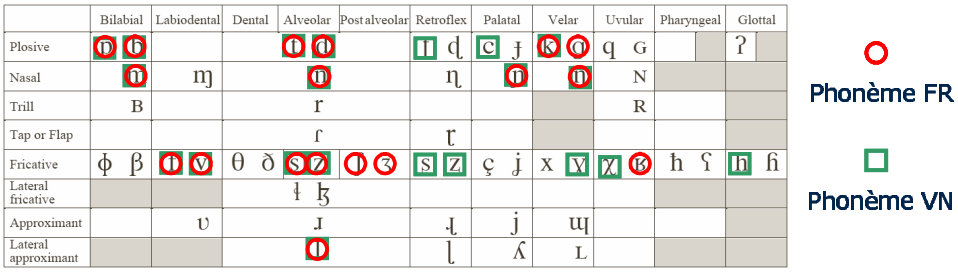


Figure 4 : couverture phonémique du français et du vietnamien pour les consonnes

3.4 Modèles acoustiques

Nous avons vu au paragraphe 3.2 qu'il existe des solutions pour collecter rapidement des ressources orales et écrites dans une nouvelle langue. Dans l'idéal, si ces ressources sont en grande quantité, et si un dictionnaire de prononciation est disponible pour la langue cible, l'adaptation du système de reconnaissance peut correspondre alors à un simple réapprentissage des modèles sur ces nouvelles données. Dans la réalité, la quantité de données collectées restent bien souvent inférieure à ce qu'elle est pour les langues bien dotées. La construction d'un système de reconnaissance automatique de la parole nécessite donc également des techniques d'adaptation rapide au niveau des modèles acoustiques comme cela est proposé dans (Choukri 2001) et (Schultz 2001) par exemple.

Une approche possible consiste à obtenir un tableau de correspondances phonémiques (*phone mapping*) entre une ou plusieurs langues sources, et la langue cible. Ensuite, les modèles acoustiques des phonèmes en langue source peuvent être dupliqués pour obtenir des modèles acoustiques en langue cible. L'avantage d'une telle approche est qu'elle ne nécessite pas ou peu de signaux d'apprentissage en langue cible puisque les modèles acoustiques du système de reconnaissance en langue cible sont en fait ceux d'une autre langue. Cependant, on retrouve dans cette approche les mêmes défauts que ceux mentionnés dans le deuxième point du paragraphe 3.3, à savoir le problème de la couverture phonémique (i.e. reconnaître du vietnamien avec des modèles acoustiques appris sur du français !). De tels systèmes peuvent cependant être améliorés en adaptant, par exemple, les modèles acoustiques avec une quantité réduite de signaux en langue cible.

Le problème est aussi d'obtenir le fameux tableau de correspondances phonémiques entre langue cible et langue source. Pour cela, on distingue les méthodes manuelles à base de connaissances (*knowledge-based*), des méthodes automatiques (*data-driven*). Les méthodes manuelles consistent à chercher les couples de phonèmes source/cible les plus proches dans le tableau d'API et nécessitent des connaissances acoustiques et phonétiques des deux langues (source et cible). Une approche automatique consiste plutôt à disposer d'un corpus vocal en quantité limitée en langue source et étiqueté (quelques minutes peuvent suffire), puis à utiliser un décodeur phonémique et calculer la matrice de confusion entre les phonèmes reconnus en langue source et les phonèmes de référence en langue cible. Une description plus détaillée des traitements réalisés et des expérimentations associées pour l'adaptation rapide de modèles acoustiques au vietnamien se trouve dans (Le, 2005).

4 Application au vietnamien

4.1 Ressources collectées

Ressources textuelles

La méthodologie décrite dans le paragraphe 3.2 a été appliquée au vietnamien. La quantité de pages Web collectées était de 2.5Go. Après filtrage, la quantité de données textuelles pouvant servir à l'apprentissage d'un modèle de langage statistique était d'environ 400Mo (5 millions de phrases). A titre de comparaison, une année complète du journal Le Monde en français correspond à 120Mo en moyenne.

Base de signaux de parole

Le corpus de parole vietnamien est toujours en cours d'enregistrement à MICA. A ce jour, il contient 35 locuteurs, 16 femmes et 19 hommes, venant des régions nord, centre et sud du vietnam. Chaque locuteur a enregistré environ 1 heure de parole ce qui fait un total de 35heures. Le corpus contient des séquences de lettres, de nombres et de mots isolés, mais aussi la lecture de phrases complètes et de paragraphes.

Des détails supplémentaires sur les ressources collectées pour le vietnamien se trouvent dans (Le, 2004).

4.2 Quelques expériences de reconnaissance automatique du vietnamien

Dictionnaire phonétique

Au début de ce travail, il n'existait à notre connaissance aucun dictionnaire phonétique sous forme électronique pour le vietnamien. Nous avons extrait tout d'abord un vocabulaire de 6,492 mots isolés à partir d'un dictionnaire franco-vietnamien issu du projet Papillon[4]. Ensuite, un analyseur phonétique (VNPhoneAnalyzer, voir Le, 2004) à base de règles a été

[4] http://www.papillon-dictionary.org/

développé pour obtenir automatiquement un dictionnaire de prononciation vietnamien. Ce dictionnaire phonétique a ensuite été vérifié par des experts de l'Institut Linguistique du Vietnam.

Système de reconnaissance automatique du vietnamien

Notre système de reconnaissance utilise la boîte à outils Janus-III de CMU. Nous avons appliqué la méthodologie décrite au paragraphe 3.4 avec le français comme langue source et le vietnamien comme langue cible. Nous sommes conscients que le choix du français comme langue source n'est pas du tout optimal et nous travaillons actuellement à l'utilisation de modèles source multilingues. Nous avons testé deux techniques d'obtention du tableau de correspondances phonémiques (*knowledge-based* et *data driven*). Les performances du système de reconnaissance automatique de parole continue du vietnamien testé sur un corpus d'une heure de dialogues sont présentées dans la *figure 5* avec respectivement l'utilisation de 0h, 1h et 2h de signal en langue vietnamienne pour adapter les modèles acoustiques empruntés au français.

Génération du tableau de correspondances phonémiques	Quantité de signaux d'adaptation en langue cible		
	0h *non-adap.*	**1h**	**2h**
Manuelle (knowledge-based)	16.13	60.4	63.6
Automatique (data-driven)	**18.52**	**61.6**	**63.8**

Figure 5 : performances (%mots corrects reconnus) de notre système de reconnaissance du vietnamien en fonction de la quantité de signaux d'adaptation utilisée et de la méthode de génération des correspondances phonémiques

Ces résultats montrent le potentiel de l'approche automatique pour la génération du tableau de correspondances phonémiques qui donne des performances équivalentes à celle obtenues avec la méthode manuelle. Nous voyons également qu'avec une quantité réduite de signaux en langue cible (2h), il est possible d'obtenir des performances acceptables (63.8% de mots correctement reconnus), même si la couverture phonémique n'est pas optimale (français vers vietnamien).

5 Travaux en cours sur le khmer

Les travaux actuels sur le khmer concernent essentiellement la collecte de ressources. Des ressources textuelles ont été recueillies avec la méthodologie présentée dans cet article et l'enregistrement d'un corpus en langue khmère est en cours à l'ITC.

Ressources textuelles

A l'aide d'étudiants cambodgiens, nous avons cherché un nombre réduit de pages Web publiées par le gouvernement cambodgien, par des organisations ou des compagnies. Nous avons d'abord remarqué que beaucoup de sites web hébergés au Cambodge sont en fait écrits

en anglais ou en français. Il y a cependant quelques sites écrits en langue khmère. Avec ceux-cis, il y a encore des difficultés de récupération automatique: sites écrits en flash[5], pages encodées par un système d'encodage spécifique ou privé[6] que nous n'avons pas réussi à convertir en un autre encodage (Unicode). Nous avons trouvé cependant un site des nouvelles en khmer[7]. La quantité de pages Web collectées à partir de ce site est actuellement de 80Mo (environ 6000 pages), ce qui reste faible par rapport aux 2.5Go de pages web collectées pour le vietnamien et par rapport aux 40Go pour le français (corpus WebFR4) (Vaufreydaz, 2002).

Concernant les modèles de langage, nous avons identifié deux pistes possibles : la première consisterait à faire un système de reconnaissance syllabique où les co-occurrences modélisées seraient des suites de syllabes ; la seconde consisterait à faire un système de reconnaissance de mots où les co-occurrences modélisées seraient des suites de mots. Dans les deux cas, se pose le problème de la segmentation (en syllabes ou en mots) qui est difficile pour les langues comme le thai et le khmer. Il existe des travaux sur ces langues pour résoudre ce problème : utilisation de grammaires de syllabes (Berment, 2004), méthodes basés sur un vocabulaire, méthodes probabilistes (Meknavin, 1997), …

Base de signaux de parole

En appliquant la boîte à outils de traitement de corpus de texte mentionnée dans la section 3.2, nous avons obtenu dans un premier temps un corpus de phrases à prononcer afin d'effectuer des enregistrements. Pour cela, nous avons utilisé un vocabulaire de 16,000 mots pour filtrer des phrases qui seront énoncés. Ce vocabulaire a été obtenu à partir du dictionnaire khmer Chuon Nat [8]. Le corpus de parole khmer est en cours d'enregistrement à l'ITC. A ce jour, 3 heures de parole ont été enregistrées par 6 locuteurs phnom-penhois.

6 Conclusion

Nous avons décrit dans cet article notre méthodologie permettant de développer et d'adapter le plus rapidement possible un système de reconnaissance automatique de la parole pour une nouvelle langue peu dotée. Des résultats expérimentaux ont été présentés pour le vietnamien et une validation sur la langue khmère est en cours. Dans le futur proche, nous travaillerons essentiellement à une meilleure couverture phonémique par l'emploi de modèles acoustiques multilingues.

---

[5] http://www.everyday.com.kh

[6] http://www.seasite.niu.edu/khmer/

[7] www.cambodiacic.org

[8] http://www.khmeros.info/

Remerciements

Ce travail a été réalisé avec l'aide de l'agence universitaire de la francophonie (AUF) dans le cadre du projet TALK.

Références

BERMENT V. (2004) Méthodes pour informatiser des langues et des groupes de langues peu dotées. Doctorat de l'Université J. Fourier – Grenoble I, Mai 2004.

CHOUKRI, R. & al. (2001) Portability of Automatic Speech Recognition Technology to New Languages: Multilinguality Issues and Speech/Text Resources. Panel Session on Automatic Speech Recognition and Understanding (ASRU-2001), Madonna di Campiglio, Italy, December 2001.

LE V.-B., BIGI B., BESACIER L., CASTELLI E. (2003), Using the Web for fast language model construction in minority languages, Eurospeech '03, Geneva, Switzerland, September 2003.

LE V.-B., TRAN D.-D., CASTELLI E., BESACIER L., SERIGNAT J.-F. (2004), Spoken and written language resources for Vietnamese, LREC 2004, Lisbon, Portugal, 26-28 May 2004.

LE V.-B., BESACIER L. (2005), First steps in fast acoustic modeling for a new target language: application to Vietnamese, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05), Philadelphia, USA, 19-23 March 2005.

MEKNAVIN, S., CHAROENPORNSAWAT, P. KIJSIRIKUL, B., (1997), Feature-based Thai Word Segmentation, In Proceedings of the Natural Language Processing Pacific Rim Symposium 1997 (NLPRS'97), Phuket, Thailand.

SCHULTZ, T., WAIBEL, A. (2001). Language independent and language adaptive acoustic modeling for speech recognition. Speech Communication, vol. 35, no. 1-2, pp. 31–51, 2001.

VAUFREYDAZ D., AKBAR M., CAELEN J., SERIGNAT J.-F., (1998) EMACOP Environnement Multimédia pour l'Acquisition et la gestion de COrpus Parole, JEP'98 (Journées d'Étude sur la Parole), Martigny (Switzerland), pp. 175-178, June 1998.

VAUFREYDAZ D., (2002) Modélisation statistique du langage à partir d'Internet pour la reconnaissance automatique de la parole continue. Doctorat de l'Université J. Fourier – Grenoble I, Janvier 2002.

Localization in the Context of a Third World Country

Dawit Bekele

Department of computer science
Addis Ababa University
P. O. Box 3479
Addis Ababa
Ethiopia
dawitb@ethiolink.com

Mots-clés : Localisation, Amharique, Ethiopie, Normalisation
Keywords: Localization, Amharic, Ethiopia, Standardization

Résumé Après avoir défini la localisation, cet article discute son importance pour les nations du tiers-monde. Ensuite, il présente les raisons essentielles pour lesquelles la localisation est de plus en plus importante pour les pays du tiers-monde à savoir: pour les permettre de travailler dans leurs langues officielles, pour permettre l'accès des NTIC aux jeunes et pour limiter la faille numérique. Cet article continue par présenter l'histoire de la localisation en Ethiopie qui consiste essentiellement d'efforts avec les trois buts suivants: 1- permettre des entrées sorties en Ethiopic 2- Localisation de la date et de l'heure 3- Localisation des interfaces de logiciels. L'article termine par présenter les questions essentielles que les pays en voie de développement doivent considérer. Ces questions sont: Coordination, Normalisation, Vision courageuse, Partenariat et Solution logiciel libre.

Abstract After defining localization, the paper discusses its importance for third world nations. Thereafter, it presents the main reasons why localization is increasingly important for third world countries: to be able to work in official languages, to provide access to the youth and to limit the digital divide. The paper then discusses the history of localization in Ethiopia that concerned mainly efforts with the following three goals: 1- enabling Ethiopic input and output, 2- localization of the date and time, 3- localization of software interfaces. Finally, it proposes the main issues that developing countries seeking localization should consider, which are: Coordination, Standardization, Bold vision but realistic approach, Partnership and Open source option.

1 Introduction

Until recently, most third world countries seemed resigned to using software products in the languages of more prosperous countries. This is especially true in African countries where local languages were rarely used in computers.

However, recently, more and more localization efforts are being made for languages that didn't get this opportunity in the past. Even Microsoft announced that it will localize its Windows operating system as well as its Microsoft Office Suite to 40 languages, many of them spoken in the third world, including Swahili and Amharic in Africa (Microsoft Press Passes, March 16 and March, 2004). The open source community is also contributing in localizing an increasing number of software products to languages in the developing world. Linux, for instance, is localized for than 80 languages, many of which are from the developing world.

Unfortunately, the increasing number of localization projects of the third world are facing unique problems that are inherent to the nature of localization and realities of these countries. Unless these problems are properly studied and adequate solutions provided, the efforts put on these projects may not provide as much fruit as expected.

This paper tries to identify the major challenges of localization by focusing on developing countries' environment. The decades long localization attempt of Ethiopia will guide the reflection. The paper will also propose some important localization strategy issues that developing nations should consider.

2 Localization

Localization can be defined as "the transfer of cultural consciousness into a computer system" (Daniel Yacob, 2004). The localization Industry Standards Association (LISA) also defines localization as follows: "Localization involves taking a product and making it linguistically and culturally appropriate to the target locale (country/region and language) where it will be used and sold". In reality, since the great majority of today's software products are produced in the United States of America and/or in the English language, most of the localization activities consist of transferring a software product developed in English/US culture to other cultures and languages.

Prior to the 1980s, software producers gave little importance for localization. In the early days of computing, even the largest computer companies ignored the importance of localization. For example, most products developed prior to the 1980s used the ASCII character set [1]. As a result, they did not even support characters such as é and è that are compulsory for French, a language used by tens of countries and hundreds of millions of people around the world. Later on, ISO-8559/1-8559/9[2] defined character sets for "major" European languages, Arabic and Hebrew which facilitated localization to these languages. However, it's only after Unicode[3]

[1] ANSI X3.4 American National Standard code for Information Interchange

[2] ISO-8559/1 - 8559/9 8-bit single Byte Code Graphic character sets

was released that most alphabets and scripts of the world have had the chance to be used by major software products (ISO, 1998).

In the late 1980s, software companies started to realize the importance of localization to capture more markets. More and more products started to be shipped with an interface in the language of the country of destination. Localization industry started to emerge with Ireland as its hub. In 1990, the Localization Industry Standards Association (LISA) was founded (Esselink Bert, 2000) to coordinate the localization efforts.

Today, localization is taken very seriously by all major software producers. Most software producers have internationalization[4] departments that insure that all the software products they produce are easily localizable[5]. Programming environments are also friendlier to localization: Java, Visual Studio as well as .Net platforms have increasing localization capabilities (Daniel Brandon, 2001).

As indicated in the definition, localization is more than translation from one language to another. It is true that translating the interface, the help system as well as the manuals is probably the most time consuming activity. However, localization also includes making sure that the software uses the particular locale's cultural conventions such as the date and time system, measurement system, number formats[6] colors and icons. Some conventions that are well established in some country may be completely unknown in some other country. For example the mail box image often used in American software products has no meaning in Ethiopian context, where mail is not delivered at home and such boxes do not exist.

3 Why is localization so important for developing countries?

Developing countries in general and African countries in particular have been using computers for at least the last four decades without necessarily benefiting from localization to their own languages. It is therefore a legitimate question to ask why not continue to use computers having western language interfaces. One could also say that localization is by no means an easy task that poor countries can readily afford, especially when these countries have hundreds of languages and different cultures. It also requires sustained effort; for example, Ge'ez Frontier Foundation translated into Amharic more than 40,000 thousands of words of the Linux operating system, which amounted to 40% of the vocabulary. However, just a year later the percentage decreased to 22% due to new words included in Linux software and the slowdown of the translation effort[7].

---

[3] Unicode/ISO 10646 32 bit character set

[4] According to LISA, Internationalization is the process of generalizing a product so that it can handle multiple languages and cultural conventions without the need for redesign

[5] Excerpt from the talk of Menassie, Zaudou, Software developer at Sun Micro Systems.

[6] French speaking countries use the comma (,) as a decimal separator while most of the remaining world uses the dot (.). This after creates confusions that can have severe consequences.

[7] http://l10n-status.gnome.org/gnome-2.10/index.html gives the level of real-time progress of all GNOME localization projects.

However, there are an increasing number of reasons why these countries have to invest in localization, in spite of the high cost that it may incur. Some of these reasons are presented below.

One of the most important reasons, at least for some of the countries, is that the local languages are the working language of the offices where the computers are used. In Ethiopia for instance, the official working language of the Federal Government is Amharic and the regional governments use their official working language. Until now, even though English is not the official language of the country, government and private organization employees who need to work with computers were required to know English. This requirement was acceptable when computers were used only in large federal organizations that have employees with, almost always, university degrees. Today, thanks to some projects such as the WeredaNet that connects each Wereda (district) administration of the country, computers are used by people who do not even speak the federal working language let alone English. This new bread of users cannot use the computers that they have available unless they are localized to their languages and cultures. Therefore, localization is no more a luxury but a necessity.

Another reason follows: on one hand computers are being increasingly used by younger people since, for example, SchoolNet programs are being launched all over the world, even in the developing world. On the other hand more and more school systems use local languages as teaching medium. In Ethiopia, the multi-billion Birr[8] SchoolNet program is installing and connecting computers in all high schools of the country. This gives to the millions of schoolchildren access to computers that they did not have before. It is expected that primary schools will have similar settings in the near future. At the same time, political leadership and education specialists are pushing for the use of local languages, at least in primary schools (African Ministers, 2002), (Kwesi Kwaa Prah, 2002). As a result starting from mid-1990s, all Ethiopian public primary schools teach in the regional languages. Consequently, since the young schoolchildren do not yet master languages other than their own, providing them with localized computers becomes a necessity.

The third reason is related to the digital divide. Most governments are now convinced that they have to avoid or at least limit the digital divide that exists between rich and poor countries but also between rich and poor within the same country, in order to avoid or at least limit some serious economic and social consequences. There are studies that show that in African countries, the overwhelming majority of the population does not speak the western languages that their countries have adopted as national languages. In Ethiopia, even though English is used in higher education and correspondences with foreign organizations, it is not an official language and few people use it on regular basis. Without localization, the use of computers is automatically restricted to the very few that speak western languages and it is therefore impossible to talk about bridging the digital gap.

The above three reasons, and probably others not given here, make localization a necessity, for an increasing number of countries. Even those that do not yet feel the need, will be confronted to it in the near future, because of the desirable and unavoidable popularization of computers.

---

[8] Birr: Ethiopian Currency equivalent to approximately 0.09 Euros in March 2005

4 History of localization in Ethiopia

Ever since the introduction of computers in Ethiopia in the 1960s, users and government highly desired localization. This is because, as already stated, the official working language of the central government is Amharic and it is only natural that users want to use the computer in their working language.

Amharic is a semetic language that derive from Ge'ez, a language now extinct except in the Ethiopian orthodox church, where it remains the liturgical language. Amharic is one of the more than eighty languages spoken in Ethiopia. It is the official language of the Ethiopian government for several centuries and the lingua franca of the country.

Amharic uses Ethiopic (Ge'ez) syllabary for its writing system. The syllabary has 461 symbols (QSAE, 2002). The number of symbols of Ethiopic has been increasing throughout the years to accommodate sounds that are not supported and that are needed for new languages that have started using the syllabary. In addition, all other Ethiopia's and the now independent Eritrea's languages used this script before the government change of 1991. In particular, Afan Oromo, the language of the largest ethnic group of the country, has been using Ethiopic until 1991.

During the last four decades, at least the following localization efforts have been vigorously pursued:

1. Enabling existing software products to accept Ethiopic inputs and outputs. This has been given highest priority by users and developers. The main problems were associated with the high number of symbols in Ethiopic compared with the Latin script overwhelmingly used in computer software products and the lack of relevant standards (Abass B. Alamnehe, 1994).

   Developers used mainly the following four strategies to solve the encoding problem: using a subset of the character set rather than the whole set; using two or more fonts to cover all Ethiopic symbols, separating the diacritic marks from the symbols; or using a two-bytes character coding system. Unfortunately, none of the above strategies were fully satisfactory. Nonetheless, today, Ethiopic is included in Unicode, an encoding system developed to accommodate all alphabets of the world and not just Latin as ASCII does. Since Unicode is supported by increasing number of software products, it can therefore be considered that the encoding problem is now solved.

2. Localization of the date and time. Ethiopia differs from rest of the world for having its own calendar and time system. The official Ethiopian calendar is 7 years, 8 months and 11 days behind the Gregorian calendar. It also differs from the Gregorian calendar for having 13 months; 12 of which have 30 days and the 13^{th} has 5 days or 6 days on leap years. In addition, the official time system is shifted by 6 hours compared to the system used in most of the world. Therefore, for an Ethiopian, it is 1 o'clock in the morning when it is 7 o'clock according to the western time system. Of course, none of the software products developed in the developed world comes with these date and time systems integrated. Some Ethiopian software developers have produced software products that display the Ethiopian date and time. However, the

most transparent localization for Ethiopian date and time systems has been on open source software such as Linux and OpenCMS that avail the source code hence make complete and transparent replacement of the Gregorian and western systems by the Ethiopian date and time systems.

3. Localization of software interfaces. Localization of software interfaces has been given low importance until recently since most users had some understanding of English and were expected to be able to work on computers using English interface. It is also because the most commonly used software products were Microsoft products and Microsoft did not allow localization by third party developers. Besides, the original software developers had no interest to invest on localization for a country with a GDP per capital of just over 100 USD (UNDP, 2003)!

4. The only software products that have localized interface were a small part of the few products developed locally, such as CMS used in courts and PIS used by the Ministry of justice (Dawit Bekele, 2003). Fortunately, in recent years, localization of software interfaces has been facilitated by open source software. Very recently, Microsoft has also shown interest to localize Windows and the Office suite to Amharic.

5 Localization strategy issues for third world countries

Very little is done in the area of localization, especially in third world countries. However, it is very probable that localization projects will increase in the coming years as they have during the past few years. These projects can learn from the decades-long localization attempts in Ethiopia and save precious time and money. This section proposes localization strategy issues that third world nations should consider in order to reach their final objectives with minimum effort and time.

The localization strategy of third world countries should consider at least the following major issues:

- Coordination
- Standardization
- Bold vision but realistic approach
- Partnership
- Open source option

It is important that all localization efforts for a specific locale are coordinated in order to avoid redundant works that only create confusion and additional cost, things that developing countries cannot afford. The coordination can be ideally done by an ICT authority or ministry. Otherwise, organizations interested in localization can form associations in the same way developed nations' localization companies established associations such as LISA.

Standardization greatly facilitates localization work and increases the quality of the localization as well as its acceptability by the general public. Adequate standards should be

developed at the appropriate time. Setting a standard before the area that is standardized is properly explored limits its acceptability since better options may be discovered once the standard is adopted. Setting a standard well after all techniques of the standardized area are explored will create the proliferation of incompatible products that will not disappear easily even after the standardization (Cargill Carl F., 1997). For example, in Ethiopia, even though an encoding standard is now available, most people continue to use the incompatible non-standard fonts they have been using for a long time. Similarly, since all sorts of incompatible keyboard layouts are being used, it is almost certain that many people will prefer to continue to use the non-standard keyboard layouts that they are used to even after a keyboard layout standard is established.

Third world countries should and can have bold visions such as localizing all major software products to all their major languages. It is important that they have this kind of vision in order not to succumb to the digital-divide. With proper strategy and with convinced leadership this is not an unrealistic vision. However, they should also realize that localization is not just translation of words and it requires a lot of sustained effort. Therefore, they should have realistic and progressive approach that enables them to see the fruit of their efforts as soon as possible. For example, localizing to all languages of the country at the same time is generally a bad approach since problems in the localization of one language can delay that of the others and also because it will drain heavily the resources of the country.

Localization will benefit if there is increasing partnership among localization project. There are a lot of intersections between localization works even between those targeting completely different cultures and languages. For example, in the Amharic localization of Open CMS performed by Addis Ababa University, many problems were encountered and most of them were solved by people involved in other localization projects around the world. Very surprisingly, the most useful assistances came from Chinese professionals who faced similar problems due to their use of a writing system different from Latin.

Partnership can also save a lot of cost for developing countries that cannot afford to spend a lot of money on localization. It also enables countries newly embarked in localization to learn from the experiences of others and reach their goals faster.

Last but not least, developing countries should seriously explore the localization of free and open source software. There are many reasons why developing nations benefit from using free and open source software products. This paper will only focus on those reasons that concern localization. Free and open source software products are most of them internationalized probably because they are developed by international teams with members from all around the world. Proprietary software products are less often internationalized. Free and open source software products are also interesting for localization since their source code is freely available. This makes any localization possible even those that were not planned by the original developers.

6 Conclusion

Localization is becoming an important issue for third world countries. It will continue to be an even bigger issue when more and more citizens of third world countries have access to computers and when their leaders are more aware of its importance.

Nonetheless, even though localization has become easier in the last few years there are still a high number of challenges and new countries should learn from others' experiences in order not to repeat the same mistakes. In this regards, the Ethiopian localization experience can give many lessons to African countries in particular and third world nations in general.

Bibliography

ABASS B. ALAMNEHE (1994), The need for a standardization of Ethiopian Script, Proceedings of EthCITA E-mail Conference, Volume III paper 3

AFRICAN MINISTERS (2002), Eighth conference of Ministers of Education of African members States (MINEDAF VIII)

CARGILL CARL F. (1997), Open Systems Standardization, A business Approach, Prentice Hall

DANIEL BRANDON (2001), Localization of Web Contents, Journal of Computing Sciences in Colleges, Volume 17 Issue 2

DANIEL YACOB (2004), "Localize or be Localized: An Assessment of Localization Frameworks", International Symposium on ICT Education and Application in Developing Countries, Addis Ababa, October 19-21, 2004.

DAWIT BEKELE (2003), Computerization Study Report, Ministry of Justice of the Federal democratic Republic of Ethiopia, Branch Office for Addis Ababa Administration

ESSELINK BERT (2000), "A practical Guide to Localization", John Benjamin's Publishing Company.

ISO (1998), ISO/IEC 10646-1:1993/Amd 10:1998(E), Amendment 10: Ethiopic Architecture and Basic Multilingual plane

KWESI KWAA PRAH (2002), NEO-COLONIALISM AND the African development challenge, TRIcontinental, Havana, Cuba, No. 150, 2002, Language

MICROSOFT PRESS PASS (March 16, 2004), Microsoft Enables Millions More to Experience Personal Computing Through Local Language Program

MICROSOFT PRESS PASS (March 2004), Microsoft Local Language Program, Quote Sheet

QSAE (2002), QSAE (Quality and Standards Authority of Ethiopia), Ethiopian Standard Ethiopic character set ICS:01.140.10, ES 781:2002

UNDP (2003), Human Development Report, www.undp.org

Aspects du traitement automatique du gallois

Johannes Heinecke
France Télécom, Division Recherche & Développement
2 avenue Pierre Marzin, F-22307 Lannion Cedex
johannes.heinecke@francetelecom.com

Mots-clefs : typologie, lexique, morphologie, syntaxe, ressources, corpora, TALN, gallois

Keywords: typology, lexicon, morphology, syntax, resources, corpora, computational linguistics, Welsh

Résumé Cet article décrit les ressources linguistiques (informatisées) du gallois, une langue celtique parlée par plus que 500.000 personnes au Pays de Galles (Royaume Uni). Après une introduction brève sur la situation actuelle du gallois, ses spécificités typologiques et des difficultés potentielles pour un traitement automatique sont présentées. Ensuite les ressources linguistiques électroniques disponibles et/ou envisageables sont discutées. Pour finir nous présentons quelques travaux dans le domaine du traitement automatique du gallois.

Abstract This article describes the (electronic) linguistic resources for Welsh, a Celtic Language spoken by more than half a million persons in Wales (UK). After a short introduction to the current situation of the Welsh language, its main typological features are presented, including potential difficulties for all areas of natural language processing. In a following section the electronically available linguistic resources are listed and discussed. Finally an eye is cast onto some work on Welsh within computational linguistics.

1 Introduction

Le gallois (Cymraeg) est, à coté de l'anglais, une des langues du Pays de Galles au Royaume Uni. C'est une langue celtique, plus précisément, une langue P-celtique insulaire.[1] Elle est la sœur du Breton (parlé en Bretagne en France) et du Cornouaillais, une langue disparue depuis la fin du XVIII^{e} siècle, qui était parlée au Cornwall (sud-ouest de la Grande Bretagne). Les autres langues celtiques (Q-celtique) sont l'irlandais, le gaélique de l'Ecosse et le manx (disparu récemment, mais toujours parlé par quelques personnes qui l'ont appris comme deuxième langue).

Dans cet article nous donnons un aperçu général sur le gallois et dans un cadre plutôt linguistique, ses caractéristiques typologiques. Dans une autre section nous présentons les ressources linguistiques électroniques existantes et leur utilité pour un traitement automatique. Nous finissons cet article avec une brève introduction aux travaux dans le domaine du TALN pour le gallois. Nous nous restreignons ici à un panorama général du gallois et son traitement automatique. Pour plus de détails voir les références bibliographiques.

2 La langue galloise

2.1 Situation actuelle

Depuis 1998[2] le gallois est reconnu comme langue officielle au Pays de Galles à coté de l'anglais. Elle est utilisée dans les administrations et enseignée ou la langue d'enseignement dans beaucoup d'écoles et dans l'université du Pays de Galles. Un l'organisme du gouvernement, le *Welsh Language Board*[3] assure que l'utilisation de la langue galloise est possible dans tous les endroits de la société et fait la promotion de la langue. Selon le dernier recensement 2001, presque 21% de la population galloise (i.e. 582.400 personnes) parle le gallois.[4] Par rapport au recensement antérieur (1991), cela signifie une augmentation d'environ 80.000 personnes (1991 : 18,7%). Grâce à l'enseignement obligatoire, même dans les régions plutôt anglophones (sud-est), le pourcentage des jeunes qui parle le gallois est en forte croissance (cf. tableau 1).

Actuellement le gallois n'est pas immédiatement menacé de disparaître, ceci grâce aux nouvelles structures mises en place pendant les dernières décennies afin de protéger et promouvoir l'utilisation du gallois. Par contre, le fait que quasiment tous les locuteurs du gallois sont bilingues (anglais) a des effets sur le vocabulaire et la syntaxe du gallois (parlé). On y trouve nombreux anglicismes et de constructions syntaxiques issues de l'anglais, comme la position finale de la proposition *o* « de » au lieu d'être avant les pronoms interrogatifs :

(1) *lle wyt ti 'n dod* ***o*** *?* syntaxe correcte : ***o*** *le wyt ti 'n dod ?*
où es tu IMPF venir de — de+SM SM-où es tu IMPF venir
« d'où est-ce que tu viens ?» (cf. « where do you come **from** ? »)

[1] Pour la distinction entre P-celtique et Q-celtique voir (Ball & Fife, 2002).
[2] Government of Wales Act 1998
[3] http://www.bwrdd-yr-iaith.org.uk/
[4] Voir l'information sur ce dernier recensement sur le site du *Welsh Language Board*.

Age	personnes	pourcentage
3-4	13.239	18,8 %
5-15	171.168	40,8 %
16-19	40.548	27,6 %
20-44	146.227	15,5 %
45-64	112.742	15,6 %
65-74	47.692	18,1 %
75+	50.752	21,1 %

TAB. 1 – résultats du recensement 2001 : personnes capables de parler le gallois

À l'exception des journaux quotidiens, il existe depuis longtemps une grande diversité d'hebdomadaires et de magazines spécialisées. La vie littéraire est en bonne santé avec plus de 500 nouvelles publications par an.[5]

2.2 Caractéristiques typologiques

Malgré le fait que les langues celtiques, et donc le gallois, sont des langues indo-européennes, elles montrent quelques traits typologiques qu'on ne trouve pas parmi les autres langues de la même famille de langues. La plupart de ces différences concerne toutes les langues celtiques insulaires. La description suivante est en partie spécifique pour le gallois. La plupart de ces traits typologiques ne pose aucun problème à un traitement automatique du langage, par contre les approches de base, développées souvent sur la typologie de l'anglais, allemand, français etc. peuvent être trop spécifiques pour traiter le gallois. Par exemple l'ordre des mots de base du gallois est VSO, i.e. le verbe prend toujours la première place dans une phrase.

Un phénomène très connu dans les langues celtiques sont les mutations des consonnes initiales des mots en fonction de leur fonction syntaxique. Le gallois connaît trois mutations, la lénition (*soft mutation*, SM), la nasalisation (*nasal mutation*, NM) et l'aspiration (*aspirate mutation*, AM).[6] Par exemple, le complément d'objet direct est muté : *Gwelodd Ioan gi* « Ioan a vu un chien ». La forme dite « radicale » du mot pour chien est *ci*, qui apparaît ici avec lénition. Certaines consonnes ne mutent pas, par contre l'aspiration des voyelles implique un *h* préfigé : *afal* « une pomme » mais *ei hafal* « sa pomme (à elle) ». Il y a des cas où seulement la mutation peut désambiguïser le syntagme : lénition *ei dŷ* « sa maison (à lui) » vs. aspiration : *ei thŷ* « sa maison (à elle) » ou encore *daeth o* « il est venu » vs *(a)*[7] *ddaeth o ?* « est-ce qu'il est venu ? ».

Toutes les langues celtiques connaissent des prépositions fléchies (nombre, personne et genre) devant des pronoms : *gennyf i* « avec moi », *ganddo fo* « avec lui », mais *gan y dyn* « avec l'homme ». Ces prépositions fléchies, l'ordre basique des mots, le comportement de l'article

[5] Welsh Books Council, http://www.wbc.org.uk/

[6] Voir Ball & Müller, 1992; Borsley, 1999.

[7] La particule interrogative *a* n'est pas obligatoire et disparaît normalement dans la langue parlée et souvent dans le gallois écritmoins formel. Ce ne laisse que la mutation comme le seul trait qui indique l'interrogation.

défini qui rend défini une construction génitive ont provoqué depuis le début du XIX^e^ siècle, les hypothèses que les langues celtiques insulaires ont été en contact avec des locuteurs des langues afro-asiatiques, notamment les langues sémitiques, qui sont les seules langues qui partagent ces phénomènes assez étonnants (Pokorny, 1927-1930; Wagner, 1959; Gensler, 1993; Vennemann, 2002).

Le gallois, notamment sa version parlée, possède un système verbal analytique complexe afin d'exprimer le temps et l'aspect (voir Heinecke, 1999). Contrairement aux autres langes indo-européennes (mais comme l'irlandais par exemple), le gallois utilise des particules (historiquement des prépositions) afin d'exprimer le temps sémantique (exemple 2) et/ou l'aspect (exemple 3) :

(2) *Mae 'm brawd* ***wedi*** *cyrraedd*
est mon frère nouveau arriver
« mon frère vient d'arriver » littéralement « mon frère est après d'arriver »

(3) *Roeddwn ni* ***'n*** *canu trwy 'r nos*
AFF-étions nous IMPF chanter pendant la nuit
« nous chantions pendant toute la nuit »

Mis à part sa structure temporelle, l'accord entre le verbe et le sujet n'est complet que si le sujet est un pronom. Pour un sujet nominal, le verbe est toujours à la 3^e^ personne du singulier.

Au niveau syntaxique, il est à observer que le gallois n'a pas de diathèse (passif) mais il y a des formes impersonnelles (une septième forme en plus des 3 personnes singulier et pluriel) qui reprennent une partie du passif : Une expression comme *argraffwyd yng Nghymru* « imprimé au Pays de Galles » littéralement « on (l') a imprimé au Pays de Galles » est utilisée, quand on ne connaît pas l'agent (où le premier actant). Syntaxiquement il ne s'agit pas d'un passif, car cette construction est similaire aux autres formes actives.[8] Il correspond plutôt au *on* français ou *man* allemand. Dans la langue parlée les constructions avec l'auxiliaire *cael* « avoir, recevoir » font émerger quelque chose qui est plus comparable avec le passif français etc. :

(4) *Mi gês i fy ngheni ym Mis Chwefror*
AFF+SM SM-avoir-PRÉT-1SG je+NM NM-mon NM-naître en mois février
« je suis né en février », littéralement « j'ai reçu mon naître en février »

Un autre phénomène syntaxique intéressant est le « function spreading ». Il s'agit d'une phrase coordonnée dont seulement le premier verbe est fléchi, les autres verbes étant des infinitifs (noms verbaux). La personne, le nombre, le temps et l'aspect du verbe initial sont implicitement sous-entendus pour les autres verbes[9] :

(5) *cododd Gwilym ei docyn a* ***rhedeg*** *am y trên*
acheter-3SG-PRÉT Gwilym son+SM SM-billet et courir-VN pour le train
« Gwilym acheta son billet et courut pour le train » (Thomas, 1996)

[8] Sauf le fait que le COD indéfini n'est pas lénié. Voir aussi Awbery, 1976.
[9] Voir plus dans le cadre de la LFG dans Sadler, 2003.

Enfin, syntaxiquement la possession et le complément d'objet direct ne sont pas différenciés : En fonction de la catégorie de sa tête, le pronom possessif[10] exprime soit le possesseur soit l'objet :

(6) *mae o 'n* ***fy*** *ngweld* ***(i)***
est il IMPF mon+NM NM-voir je
« il est en train de me voir »

mae hwn yn ***fy*** *nhŷ* ***(i)***
est ce PRED ma+NM maison (je)
« c'est ma maison »

Une difficulté à résoudre est l'existence de deux standards *de facto*. Il existe une norme plus conservatrice du gallois qui est surtout utilisée dans les textes écrits (par exemple dans les journaux ou dans la littérature). Cette version du gallois utilise plus souvent des formes verbales synthétiques au lieu des formes analytiques : Par exemple dans le gallois plutôt formel le passé antérieur est exprimé par une temps grammatical nommé « Irrealis Tense »,[11] mais de préférence par une périphrase analytique dans les occasions moins formelles : *canasem* « nous aurions eu chanté » vs *mi fasen ni wedi canu*. De plus, dans le gallois plus formel, les particules verbales ne sont pas disparues ou confondues avec les racines verbales, mais restent telles quelles :

(7) ***yr ydwyf*** *i [wedi darllen y llyfr]*
AFF être-1SG-PRÉS je après lire le livre
« j'ai lu le livre »

vs

(8) ***(ry)dwi*** *[wedi darllen y llyfr]*
être-1SG-PRÉS après lire le livre

À la lecture des grammaires descriptives de Williams, 1980 et celles de King, 1993 et Thorne, 1993 on pourrait avoir l'impression qu'il s'agit de deux langues proches mais différentes. Pour un traitement automatique il faut pourtant prévoir des grammaires (c'est à dire des ensembles de règles pour l'analyse syntaxique) qui permettent d'analyser tous ces phénomènes à la fois, car, notamment dans les textes moins formels, les deux types de syntagme peuvent être trouvés.[12]

Le gallois est une langue très bien documentée depuis son début. Il existe des nombreuses grammaires descriptives, qui couvrent, d'un coté l'aspect formel du gallois, de l'autre la langue plus spontanée.[13] De plus, le gallois est l'objet d'étude pour la validation de plusieurs théories syntaxique, comme la Generative Transformational Grammar (Awbery, 1976), la théorie X-Bar

[10]C'est pour cette raison que dans les grammaires descriptives, ces pronoms s'appellent « pronoms dépendants » au lieu de « pronom possessif ». Les pronoms personnels en fonction de sujet sont nommés « pronoms indépendants ».

[11]Heinecke, 1999.

[12]Ball *et al.*, 1988 et Jones, 1988.

[13]Par exemple : Morris-Jones, 1913; Morris-Jones, 1921; Williams, 1980; Humphreys, 1980; King, 1993; Thorne, 1993; Thomas, 1996.

(Rouveret, 1990, Roberts, 2004), la Lexical Functional Grammar (LFG, par exemple Sadler, 1998 et Sadler, 1999) ou encore la Head-Driven Phrase Structure Grammar (HPSG, Borsley, 1990). Le gallois est le sujet d'innombrables études typologiques (voir Heinecke, 1999). La recherche en linguistique comparative et des études sur les langues indo-européennes ont bien étudié l'histoire et le développement du gallois.

La description phonologique est aussi complète comme la présentation d'un standard parlé (Thomas, 1987) et l'atlas linguistique du gallois (Thomas, 1973) le montrent. Le gallois a deux groupes de dialectes principaux, qui se différencient au niveau phonologique (Thomas & Thomas, 1989) mais aussi morphologiquement et lexicalement. Pour des applications de reconnaissance vocale, deux modèles de langue sont exploités (voir ci-dessous).

3 Technologie de l'information pour le gallois

3.1 Utilisation dans l'informatique

Depuis la définition de l'encodage ISO-8859-14,[14] les caractères de l'orthographe galloise sont tous accessibles. En plus de l'alphabet latin de base, le gallois connaît des accents circonflexes sur toutes les voyelles comme marqueur de longueur ; le fait que *y* et *w* sont des graphèmes qui représentent des phonèmes vocaliques, nécessite des accents sur eux aussi : *Ŷ/ŷ* (`xDE`/`xFE`), *Ŵ/ŵ* (`xD0`/`xF0`). Un autre point à ne pas négliger sont les bigraphes *ch*, *dd*, *ff*, *ll*, *ng*, *ph*, *rh* et *th*, qui sont considérés comme étant des caractères simples. En plus, *ng* suit *g* et ne pas *n* dans l'alphabet. Donc pour un tri alphabétique il faut assurer que *didwyll* précède *didda*, *engyl* précède *ehangdeb* etc.

Depuis plusieurs années, les travaux de localisation des logiciels en gallois ont été menés. Récemment une version galloise du système d'exploitation Windows XP a été publiée. En ce qui concerne Linux, des logiciels bureautiques tels que OpenOffice, Mozilla etc. ont désormais une interface galloise.[15] Le Canolfan Bedwyr[16] de l'Université du Pays de Galles à Bangor est chargé de promouvoir et mener des projets afin d'obtenir des logiciels en gallois et pour le gallois (TALN), depuis 1999 le correcteur d'orthographe *CySill* est disponible.[17] *CySill* était une nouveauté parce que c'était le premier correcteur d'orthographe qui maîtrisait les mutations.

3.2 Ressources électroniques

Les ressources linguistiques autres que celles du Canolfan Bedwyr sont rares. Le Dictionnaire de l'Université du Pays de Galles (*Geiriadur Prifysgol Cymru*, GPC), le dictionnaire gallois

[14] L'unicode, évidemment, couvre aussi la totalité des caractères qui sont nécessaire pour utiliser le gallois dans l'informatique.

[15] http://www.meddal.org/

[16] http://www.bangor.ac.uk/ar/cb/

[17] *CySill* fait partie du package logiciels *Cysgliad* qui contient en plus des dictionnaires terminologiques *Cysgeir*. *Cysgliad* est disponible au Canolfan Bedwyr.

de référence[18]. Un CD-ROM du GPC est envisagé prochainement. Plusieurs personnes, dont l'auteur, ont créé des dictionnaires accessibles sur l'internet,[19] mais ils ne suffisent pas pour un traitement automatique du gallois, car il s'agit des listes de lemmes, sans modèles de flexions etc.

Malgré le fait que la tradition des corpus a commencé au début du XX^{e} siècle Fynes-Clinton, 1913, il n'existe que très peu de corpus électroniques. Kevin Scannell de l'Université de St. Louis (Missouri, États-Unis) a implémenté des logiciels qui permettent d'établir des corpus pour les langues peu dotées à partir des documents trouvés sur le Web. Le corpus du gallois comprend 69.502 fichiers, avec environ 95.800.000 mots (septembre 2004).[20] Un autre corpus important est le *Cronfa Electronig o Gymru* (« Archive électronique du Pays de Galles ») qui a été compilé en 1994 à partir de textes littéraires, de journaux, et de documents scientifiques et commerciaux. Ce corpus est d'environ un million de mots ; il est étiqueté et contient une statistique des mots, leurs mutations et leurs catégories grammaticales (Ellis *et al.*, 2001).

Un troisième corpus qui est en train de se constituer est le *Historical Corpus of the Welsh Language*.[21] Dans ce projet, mené à l'Université de Cambridge, on veut créer un corpus des textes de la période 1500 à 1850. Son intention est plutôt de rendre disponible une base de données pour les études de linguistique historique comme par exemple les changements syntaxiques et le développement du vocabulaire. Ce corpus sera encodé en XML en utilisant le codage développé par le *Text Encoding Initiative*[22] (TEI). Ce projet offre aussi une concordance dans le but des études philologiques. Pour l'instant (décembre 2004) ce corpus contient 30 textes (420.000 mot).

3.3 Traitement automatique du langage

À ce jour le traitement automatique du langage du gallois autre que la correction d'orthographe est quasiment inexistant. Aux Canolfan Bedwyr en collaboration avec le Irish Speech Group du Trinity College Dublin, un système de synthèse, développé à l'origine à l'Université d'Edimbourg (Williams, 1999), est réutilisé dans le projet *Welsh and Irish Speech Processing Resources* (WISPR).[23] Un corpus du gallois oral a été créé dans le cadre du projet *SpeechDat Cymru* à l'Université du Pays de Galles à Swansea (Jones *et al.*, 1998). Ce corpus contient des phrases lues au téléphone par 2000 locuteurs de toutes les régions galloises. Chaque enregistrement est ensuite annoté manuellement.

Dans l'équipe Langues Naturelles de la Division Recherche de France Télécom nous avons

[18] Thomas & Bevan, 1950-2002 : la première édition a été terminé en 2002 après autour de 80 ans de travail (qui avait commencé en 1921, bien avant la publication du premier fascicule) et contient environ 84.400 d'entrées (lemmes). Depuis, l'équipe du GPC travaille sur une réédition totale. Cf. aussi http://www.aber.ac.uk/~gpcwww/. Le dictionnaire inverse (anglais-gallois, Griffiths & Jones, 1995) est aussi basé sur une base de données dont la lettre *a* était accessible en ligne à http://www.swan.ac.uk/uwp/wa_index.htm

[19] Gallois et anglais : http://www.cs.cf.ac.uk/fun/welsh/LexiconForms.html ; gallois, anglais et catalan http://www.estelnet.com/catalunyacymru/catala/gbs_mynegai_1818.htm ; gallois et allemand (par l'auteur) : http://perso.wanadoo.fr/heinecke/dict/cymraeg/.

[20] http://borel.slu.edu/crubadan/

[21] http://www.mml.cam.ac.uk/ling/research/welshcorpuseng.html

[22] http://www.tei-c.org/

[23] http://www.bangor.ac.uk/ar/cb/wispr.php

récemment adapté notre boîte à outils pour le TALN au gallois (écrit). Un petit lexique comportant les modèles flexionnels nécessaires et un ensemble de règles grammaticales (dans le cadre d'une syntaxe de dépendance) nous permet d'analyser le gallois et générer une représentation sémantique (phrase à analyser *mi fwytodd y llygoden gaws* « la souris a mangé le fromage », figure 1).

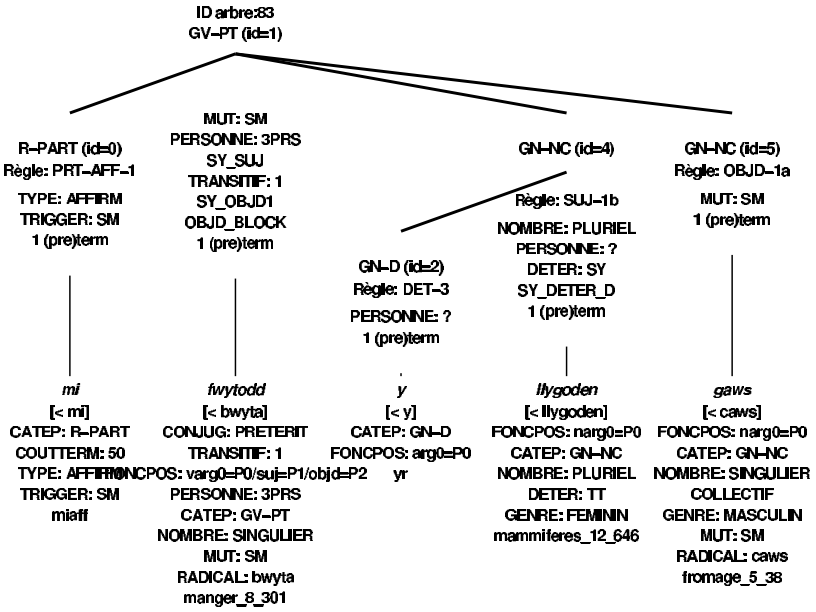


FIG. 1 – Arbre de dépendance

L'idée initiale était de vérifier si nos logiciels sont bien capables de traiter des traits typologiques du gallois absents dans les langues « bien dotées », c'est à dire la position initiale du verbe, les mutations, l'accord verbe – sujet différent selon un sujet nominal (verbe toujours en singulier) ou un sujet pronominal (accord de la personne et du nombre), ou encore la façon syntaxique de exprimer les temps sémantiques et l'aspect (*wedi*, *yn*). Le modèle sémantique pour le traitement du temps sémantique à été initialement présenté dans Heinecke, 1999, sa mise en œuvre est décrit dans Heinecke, 2005. Le lexique n'est pas (encore) lié à notre thésaurus sémantique ; ceci serait nécessaire pour aborder la traduction automatique dans le cadre de l'approche choisi (base sur des idées de la Discourse Representation Theory, Kamp & Reyle, 1993) en passant par un pivot sémantique, qui est indépendant de la langue source (Heinecke & Toumani, 2003). Après l'analyse sémantique on obtient un graphe sémantique (figure 2).

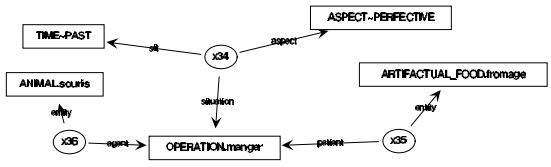


FIG. 2 – Graphe sémantique correspondant à figure 1

Malgré le structure syntaxique très différente d'une phrase qui utilise les moyens syntaxique afin d'exprimer le présent antérieur (*mae'r llygoden wedi bwyta'r caws* « la souris vient de manger le fromage ») on obtient un graphe sémantique assez similaire, à part des prédicats du temps sémantique (figure 3).

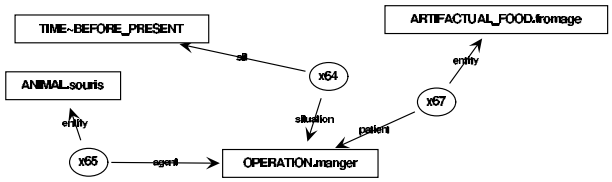


FIG. 3 – Graphe sémantique de la phrase *mae'r llygoden wedi bwyta'r caws*

Les travaux entamés seront poursuivis afin d'arriver à une traduction automatique plus complète pour au moins un domaine partiel. Le résumé automatique est l'autre objectif en vue pour notre boîte à outils ling uistiques pour le gallois. Il est cependant trop tôt pour présenter des résultats.

Les expérimentations montrent que les traits typologiques du lexique et de la syntaxe galloise ne posent aucune difficulté à nos outils du TALN. Les mutations sont déjà gérées au niveau lexical. La théorie de dépendance sur laquelle notre analyseur s'appuie est assez indépendante pour permettre un ordre de mot différent du « Standard Average European », SAE.

4 Conclusion

Le *Welsh Language Board* a récemment publié une stratégie afin d'améliorer la technologie de l'information du gallois (Welsh Language Board, 2004) ainsi qu' un rapport qui propose les travaux nécessaire pour le développement d'un système de traduction automatique gallois et anglais (Somers, 2004). à coté de la recherche scientifique, il existe un besoin d'avancement des technologies du TALN et des ressources linguistiques pour le gallois, surtout dans les domaines du traitement de l'oral (reconnaissance et synthèse vocales) et dans les technologies support pour les personnes qui utilisent le gallois dans un contexte d'informatisation (par exemple correcteurs d'orthographe, traduction automatique). Nous avons montré que le gallois n'est plus une langue « peu dotée » pour la technologie de l'information, mais le nombre de ressources reste limité. Par contre on trouve des travaux actuels dans tous les domaines autour de la technologie de l'information. En ce qui concerne le TALN, les études linguistiques du gallois ont resulté en une grande diversité des descriptions linguistiques détaillées sur tous les niveaux : phonétique/phonologie, lexique ou encore syntaxe. L'existence de lexiques bilingues gallois/anglais (sur support électronique[24]) pourra faciliter la création des données sémantiques à partir des données similaires pour l'anglais.

[24]Malgré le fait que les données ne sont pas encore accessible sous forme d'une base de données, le GPC (Thomas & Bevan, 1950-2002) et le dictionnaire anglais-gallois (Griffiths & Jones, 1995) sont sur support électronique est accessible en-ligne (voir note en base de page no. 18).

Abbréviations

AFF	affirmatif
AM	aspiration
+AM	cause l'AM sur le mot suivant
IMPF	imperfectif
NM	nasalisation
+NM	cause la NM sur le mot suivant
PRÉS	présent
PRED	particule prédicative
PRÉT	préterit
SM	lénition
+SM	cause la SM sur le mot suivant
VN	nom verbal

Références

Awbery G. M. (1976). *The Syntax of Welsh. A transformational Study of The Passive*. Cambridge : Cambridge University Press.

Ball M. J. & Fife J. (2002). *The Celtic Languages*. London : Routledge.

Ball M. J., Griffiths T. & Jones G. E. (1988). Broadcast Welsh. In M. J. Ball, Ed., *The Use of the Welsh*, p. 182–191. Clevedon : Multilingual Matters.

Ball M. J. & Müller N. (1992). *Mutations in Welsh*. London : Routledge.

Borsley R. D. (1990). Welsh Passives. In M. J. Ball, J. Fife & E. Poppe, Eds., *Celtic Linguistics. Festschrift for T. Arwyn Watkins*, Amsterdam studies in the theory and history of linguistic science : Series 4 ; 68. Amsterdam : John Benjamins.

Borsley R. D. (1999). Mutation and constituent structure in Welsh. *Lingua*, **109**, 267–300.

Ellis N. C., O'Dochartaigh C., Hicks W., Morgan M. & Laporte N. (2001). Cronfa Electroneg o Gymraeg (CEG). http://www.bangor.ac.uk/ar/cb/ceg/ceg_cym.html.

Fynes-Clinton O. (1913). *The Welsh Vocabulary of the Bangor District*. Oxford : Oxford University Press.

Gensler O. D. (1993). *A typological evaluation of Celtic/Hamito-Semitic syntactic parallels*. PhD thesis, University of California.

Griffiths B. & Jones D. G. (1995). *The Welsh Academy English-Welsh Dictionary*. Cardiff : University of Wales Press.

Heinecke J. (1999). *Temporal Deixis in Welsh and Breton*. Anglistische Forschungen 272. Heidelberg : Winter.

Heinecke J. (2005). Temps sémantique et aspect dans l'analyse semantique automatique. In *Actes de la plate-forme AFIA 2005. Atelier "Raisonnement du temps et espace"*.

Heinecke J. & Toumani F. (2003). A Natural Language Mediation System for E-Commerce applications. An ontology-based approach. In ISWC, Ed., *Proceedings of Workshop Human Language Technology for the Semantic Web and Web Services. International Semantic Web Conference, Sanibel Island, Florida, 20-23 October 2003*, p. 39–50.

Humphreys H. L. (1980). *La Langue Galloise. 2 vols. Une Présentation*. Brest : Université de Bretagne Occidental.

Jones D. G. (1988). Literary Welsh. In M. J. Ball, Ed., *The Use of the Welsh*, p. 125–171. Clevedon : Multilingual Matters.

Jones R. J., Mason J. S., Jones R. O., Helliker L. & Pawlewski M. (1998). SpeechDat Cymru. A Large-scale Welsh Telephony Database. In *Language Resources for European Minority Languages, 1998. Proceedings of the LREC Workshop, Granada, Spain 1998*.

Kamp H. & Reyle U. (1993). *From Discourse to Logic. Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Studies in Linguistics and Philosophy 42. Dordrecht : Kluwer.

King G. (1993). *Modern Welsh. A comprehensive grammar*. London, New York : Routledge.

Morris-Jones J. (1913). *A Welsh Grammar. Historical and Comparative*. Oxford : Clarendon Press.

Morris-Jones J. (1921). *An Elementary Welsh Grammar*. Oxford : Clarendon Press.

Pokorny J. (1927-30). Das nicht-indogermanische Substrat im Irischen. *Zeitschrift für Celtische Philologie*, **16 + 17.**

Roberts I. G. (2004). *Principles and Parameters in a VSO Language. A Case Study in Welsh*. Oxford Studies in Comparative Syntax. Oxford : Oxford University Press.

Rouveret A. (1990). X-Bar Theory, Minimality, and Barrierhood in Welsh. In H. Randall, Ed., *The Syntax of the Modern Celtic Languages*, Syntax and Semantics 23, p. 27–77. New York : Academic Press.

Sadler L. (1998). Welsh NPs without Head Movement. In M. Butt & T. H. King, Eds., *Proceedings of the LFG98 Conference*, Stanford : CSLI Publications.

Sadler L. (1999). Non-Distributive Features in Welsh Coordination. In M. Butt & T. H. King, Eds., *Proceedings of the LFG99 Conference*. Stanford : CSLI Publications.

Sadler L. (2003). Function Spreading in Coordinate Structures. Paper given at the 4th Celtic Linguistics Conference. http://privatewww.essex.ac.uk/~louisa/newpapers/tense-share2.pdf.

Somers H. (2004). Machine Translation and Welsh. The Way Forward. http://www.bwrdd-yr-iaith.org.uk/en/cynnwys.php?cID=6&pID=109&nID=1190.

Thomas A. R. (1973). *Linguistic Geography of Wales*. Cardiff : University of Wales Press.

Thomas A. R. (1987). A spoken standard for Welsh. Description and Pedagogy. In G. Williams, Ed., *The Sociology of Welsh*, p. 99–113. Berlin : Mouton de Gruyter.

Thomas B. & Thomas P. W. (1989). *Cymraeg, Cymrâg, Cymrêg…Cyflwyno'r Tafodieithoedd*. Caerdydd : Gwasg Tâf.

Thomas P. W. (1996). *Gramadeg y Gymraeg*. Caerdydd : Gwasg Prifysgol Cymru.

Thomas R. J. & Bevan G. A. (1950-2002). *Geiriadur Prifysgol Cymru. A Dictionary of the Welsh Language*. Caerdydd : Gwasg Prifysgol Cymru.

Thorne D. (1993). *A Comprehensive Welsh Grammar*. Oxford : Blackwell.

Vennemann T. (2002). Semitic → Celtic → English : The transitivity of language contact. In M. Filppula, J. Klemola & H. Pitkänen, Eds., *The Celtic roots of English*, Studies in Languages 37, p. 295–330. Joensuu : University of Joensuu.

Wagner H. (1959). *Das Verbum in den Sprachen der Britischen Inseln.* Tübingen : Niemeyer.

Welsh Language Board (2004). Information Technology and the Welsh Language. A strategy document.

Williams B. (1999). A Welsh speech database. Preliminary result. In *EuroSpeech 1999. Proceedings of the Sixth European Conference on Speech Communication and Technology, Budapest, Hungary, September 5-9, 1999*, Budapest.

Williams S. J. (1980). *Elements of a Welsh Grammar.* Cardiff : University of Wales Press.

TALN 2005, Dourdan, 6-10 juin 2005

Généralisation d'étiquetage morpho-syntaxique par classification supervisée

Frédérick HOUBEN, François RIOULT

GREYC – UMR 6072 – Université de Caen
campus II – BP 5186
F-14032 CAEN cedex
Frederick.Houben@info.unicaen.fr
Francois.Rioult@info.unicaen.fr

Mots-clés : Traitements multilingues, fouille de données, langues peu dotées, création de ressources, étiquetage morpho-syntaxique.

Keywords : Multilingual NLP, data mining, minority languages, means creating, tagging.

Résumé Nous présentons une méthode multilingue permettant, entre autres, de créer des ressources à moindre coût pour des langues peu dotées. Cette méthode fait appel à des propriétés très générales des langues, accessibles depuis le texte brut, ainsi qu'à des méthodes issues de la communauté de la fouille de données, notamment en apprentissage supervisé. L'objectif est essentiellement de valider la pertinence de ces propriétés à travers un ré-étiquetage réussi.

Abstract We are presenting a NLP multilingual method for low cost means for minority languages creation. This method uses very general linguistic properties, accessible from the raw text, and also engineering from data mining community, notably in supervised learning. Our first object is to validate those properties relevance through a good second tagging.

1 Introduction

Dans cet article, nous nous attacherons surtout à essayer de décrire les éléments d'une méthode de création de ressource pour des langues peu dotées à moindre coût. Cette méthode est en cours de développement et les résultats dont nous disposons ne sont qu'embryonnaires. Elle fait appel à des principes et outils issus de la communauté de la fouille de données et utilise des connaissances très générales sur les langues, connaissances qui doivent obligatoirement être accessibles directement depuis le texte brut même si, dans un premier temps, nous nous orientons vers un système d'apprentissage supervisé nécessitant l'utilisation d'un corpus étiqueté (de petite taille, 5000 mots suffisent).

Nous nous sommes fixé comme application-cadre le tagging d'un corpus d'une langue quelconque (alphabétique, non agglutinante) à partir d'un petit corpus étiqueté de cette même langue. Nous pensons qu'après une phase d'apprentissage portant sur des propriétés

facilement accessibles du corpus étiqueté, il est possible de généraliser l'étiquetage à l'ensemble du corpus. Ce qui nous intéresse réellement dans cette tâche n'est pas tant la réussite de l'étiquetage que la validation de l'hypothèse selon laquelle nous pouvons catégoriser les mots à partir de propriétés accessibles depuis le corpus brut. En cas de réussite dans cette tâche, nous pourrons alors tenter de réaliser une catégorisation lexicale sans aucune ressource, et donc sans corpus étiqueté. Dans la suite, nous appellerons aussi bien ces propriétés les attributs des mots. Il s'agit simplement de nommer les éléments constitutifs du contexte du mot qui permettront, nous l'espérons, de catégoriser correctement ce mot.

Contrairement à un étiqueteur de Brill (Brill, 1992), nous n'utilisons pas de lexique, ressource indispensable dans sa démarche, obtenue à partir d'un corpus étiqueté. Notre démarche constitue un premier pas vers une méthode sans aucune autre ressource que le seul texte brut.

Pour les mêmes raisons, nous ne souhaitons pas non plus utiliser le corpus étiqueté comme le font (Debili, 1977), (Church, 1993) ou encore (Merialdo, 1994) afin d'extraire des règles de fréquence de contiguïté de tags même si nous nous intéressons nous aussi aux rapports qui peuvent exister entre des mots consécutifs.

Enfin, nous ne souhaitons pas utiliser, comme le fait (Vergne, 1998), des règles symboliques fournies manuellement au système même si nous nous reconnaissons totalement dans son envie de minimiser les ressources nécessaires aux traitements.

Nous allons donc commencer par donné un aperçu général de la méthode puis détailler les attributs actuellement utilisés avant de nous intéresser aux outils d'apprentissage dont nous allons nous servir, introduisant notamment ce qu'est la classification supervisée, la fouille de données orientée motifs puis notre méthode actuelle. Nous présenterons enfin quelques résultats en essayant de les analyser.

2 Aperçu général de la démarche

Notre démarche se scinde en trois étapes :

1. Détermination d'un certain nombre d'attributs des mots, accessibles depuis le texte brut ;

2. Classification supervisée à partir de ces attributs. On cherche des régularités, des règles d'associations de ces attributs qui nous permettrons de déterminer la catégorie lexicale du mot ;

3. Evaluation automatique des résultats par cross-validation : on cherche à vérifier que les étiquettes attribuées à l'étape précédente correspondent aux étiquettes du corpus initial étiqueté.

A l'issue de ces trois étapes, nous pourrons alors estimer, s'il y a un bon taux de réussite, que les attributs sur lesquels nous avons travaillé permettent effectivement de discriminer les mots en fonction de rôle syntaxique. Nous pourrons alors passer à la dernière étape qui consiste en un regroupement des mots d'un corpus brut en classe d'équivalence de mots de même tag. Ce processus devient un tagging sans aucune ressource.

3 Attributs utilisés

Avant de commencer, nous souhaitons préciser que nous utilisons le terme de mot en tant que « une graphie donnée à une position particulière du corpus ». Ainsi, lorsque la même graphie se retrouve à deux endroits différents du corpus, nous parlerons de deux mots mais d'une seule graphie.

Nous nous sommes donc intéressés au contexte des mots, cherchant à trouver dans ces contextes un certain nombre de propriétés qui nous permettraient de réaliser l'apprentissage dont nous avons besoin et de discriminer correctement les différents types de mots.

Ces propriétés doivent nécessairement être accessible directement depuis le texte lui-même, sans intervention supplémentaire du locuteur, afin d'automatiser la tâche autant que possible.

Nous avons établi une première liste de cinq attributs des mots que nous allons préciser ci-dessous. Il faut cependant noter que ces attributs sont locaux, propres au mot, ce qui signifie que pour une graphie donnée, il est tout à fait possible que les propriétés qui lui sont associées ne soient pas les mêmes du fait du contexte possiblement différent de chacun des mots.

Nous allons maintenant détailler la liste de ces cinq attributs :

- Nous nous sommes d'abord penchés sur un problème déjà abordé dans (Houben, 2004) : est-ce que le mot est vide ou plein. Nous utilisons ici vide et plein dans un sens proche de (Tesnière, 1969) et considérons comme mots vides l'ensemble des mots grammaticaux ainsi que les auxiliaires. Les mots pleins sont donc tous les autres. La catégorisation en mot vide ou mot plein se fait de manière automatique en appliquant le principe de Saussure (Saussure, 1922) selon lequel « dans la langue, il n'y a que des différences » ainsi que le « principe du moindre effort » de (Zipf, 1949). Notons que cet attribut peut avoir deux autres valeurs : ponctuation et non déterminé pour les mots dont nous n'avons pas réussi à calculer s'ils étaient vides ou pleins et pour lesquels nous avons préféré ne pas faire de choix, cette incertitude étant en soi une information sur le contexte.

- Nous avons aussi souhaité conserver le type (avec les mêmes valeurs possibles que ci-dessus) du mot précédent et du mot suivant, faisant l'hypothèse que les mots vides et les mots pleins ne se succèdent pas n'importe comment. Ainsi, il y a peu de chances de trouver, en français, un déterminant avant une préposition.

- Le quatrième attribut est la position du mot dans le virgulot ((Lucas, 2003) portion de la phrase entre deux ponctuations). Cette propriété est le résultat d'une observation simple : en français, on trouve régulièrement les groupes nominaux en fin de virgulot et les groupes verbaux en début de virgulot. Nous faisons l'hypothèse que cet attribut, dans son principe, n'est pas lié qu'au français et que dans les autres langues que nous étudierons les groupes nominaux et verbaux ont une position préférentielle (hypothèse vérifiée pour un certain nombre d'autres langues).

- Enfin le dernier des attributs concerne l'influence qu'un mot vide peut avoir sur les terminaisons des mots suivants. Cette propriété n'est pas purement locale. Nous calculons, sur l'ensemble du corpus, toutes les terminaisons des mots pleins suivant

immédiatement une graphie donnée. Nous estimons que le mot peut influencer ses suivants à partir du moment où la terminaison la plus fréquente qui lui est associée représente au moins la moitié des terminaisons possibles (avec tout de même un minimum d'occurrences requis). A partir de là, nous examinons localement, pour chacun des mots, si la terminaison qui suit fait partie des terminaisons répétées (sans minimum) qui lui sont associées sur l'ensemble du corpus, auquel cas nous marquons que le mot influence son suivant. Dans tout autre cas, nous indiquons qu'il n'a pas cette influence.
Par exemple, si *les* est suivi 40 fois d'un mot terminant par *-s*, 3 fois d'un *-x* et une fois d'un -e, nous constatons qu'il peut influer sur les terminaisons des mots suivants (-s ultra majoritaire). Nous estimons alors qu'il influe effectivement sur les mots finissant par un *-s*, mais aussi sur ceux en *-x* et pas sur celui terminant avec un *-e*.

La figure suivante présente une partie du fichier une fois que nous avons extrait les cinq attributs sur l'ensemble du corpus. Cet extrait correspond au morceau de phrase «`d' ar mab henañ` », issu de notre corpus breton.

Catégorie du mot	Type du mot	Type du précédent	Type du suivant	Position dans le virgulot	Influence sur le suivant
p	v	P	v	m	o
d	v	v	n	f	n
N	n	v	P	f	n
E	P	n	s	f	n

Figure 1 : Tableau des attributs extraits du texte brut *(la première colonne de notre figure contient la catégorie du mot - p pour préposition, d pour déterminant, N pour Nom , E pour Adjectif -, résultat d'une expertise humaine sur le corpus.)*

La graphie du mot n'apparaît plus. Nous avons décidé de l'ignorer. Au mieux, nous estimons qu'elle ne sert à rien dans notre cadre et qu'elle n'est en aucun cas discriminante quant à la catégorie du mot en vue d'un tagging. Au pire, elle risque de générer un sur-apprentissage du fait d'une propriété trop « variée ». Nous n'excluons toutefois pas d'ajouter un attribut qui sera la terminaison du mot plein, celle-ci étant discriminante dans bien des langues. Par exemple, en anglais, le -ed ou le -ing sont significatifs du verbe, -ly marquant plutôt l'adverbe.

4 Quels outils pour l'apprentissage ?

Nous décrivons brièvement dans cette section la méthode utilisée pour évaluer la pertinence des attributs pour la classification supervisée.

4.1 Classification supervisée

Nous disposons d'une base de données qui décrit les propriétés des mots du corpus, parmi lesquelles figure une valeur de classe (préposition, déterminant, nom, etc.) attribuée par une expertise externe : le problème est dit *supervisé*. Dans notre cas, l'expertise fournie provient de l'analyseur de Jacques Vergne (Vergne, 1998). La *classification* est une méthode automatique qui propose une valeur de classe pour des exemples inconnus.

Pour valider la pertinence des attributs, la base de données est divisée en deux parties :

- une base d'apprentissage, qui fournit des connaissances utiles pour réaliser la classification ;
- une base de test constituée d'exemples pour lesquels nous cherchons à déterminer automatiquement une valeur de classe, à laquelle nous comparons la valeur de référence, fournie par l'expertise externe.

Dans nos expériences, nous avons évalué notre méthode en réalisant une validation croisée qui fixe la proportion de l'ensemble d'apprentissage à 90% de la base totale, les 10% restants constituant la base de test. Cette opération est répétée dix fois, afin que chaque exemple de la base serve pour l'apprentissage et pour les tests. Le score de classification est obtenu par la moyenne des dix essais, en omettant les meilleur et plus mauvais scores.

4.2 Fouille de données orientée motifs

Nous utilisons des techniques de fouille de données à base de motifs ensemblistes, c'est-à-dire que nous cherchons dans nos données des conjonctions d'attributs (les motifs ensemblistes) *a priori* intéressantes. Les méthodes les plus populaires de ce domaine sont relatives aux règles d'associations (Agrawal et al., 1996), qui expriment des régularités sous forme d'implication entre un motif prémisse et un motif conclusion.

Nous notons r une base de données sous la forme d'un triplet (A,R,O) où $A=\{a_1 \ldots a_n\}$ est l'ensemble des attributs (pour notre problème, ce sont les propriétés des mots, décrites à la section précédente), $O=\{o_1 \ldots o_m\}$ celui des objets (ce sont les mots de notre corpus), et R une relation binaire entre A et O. R indique quels attributs sont recensés dans les objets. À ce titre, un objet o pourra être considéré comme un ensemble d'attributs de A. Un *motif* X est un sous-ensemble de A, et est dit *présent* dans un objet o s'il y est inclus (on dit que o supporte X). La *fréquence* d'un motif est le nombre d'objets qui le supportent. Un motif X est *fréquent* si sa fréquence $F(X)$ dépasse un seuil γ, fixé par l'utilisateur.

Si Z est un motif, une règle d'association basée sur Z est une expression $X \rightarrow Y$ avec $X \subset Z$ et $Y=Z \setminus X$. X est la prémisse, Y est la conclusion. La confiance de $X \rightarrow Y$, notée $conf(X \rightarrow Y)$, est la proportion d'objets contenant X qui contiennent aussi Y (Agrawal, Srikant 1994), *i.e.* $conf(X \rightarrow Y) = \frac{F(X \cup Y)}{F(X)}$. Il s'agit d'une probabilité conditionnelle de présence de Y connaissant celle de X. Même si, dans la pratique, les possesseurs de bases sont plus intéressés par les règles présentes dans les données que par les motifs fréquents, celles-ci sont

dépendantes de l'obtention de ces motifs. En effet, si l'extraction des motifs fréquents est une tâche algorithmiquement difficile, la dérivation des règles d'association à partir de ces motifs ne pose pas de problème.

Plusieurs méthodes existent pour effectuer de la classification supervisée à partir des associations. Historiquement, la première et la plus simple est CBA (Liu, Hsu, Ma 1998) (Classification Based on Association). Cette méthode extrait les règles d'associations de fréquence et confiance minimales fixées par l'utilisateur, et ordonne ces règles suivant leur confiance. Lorsqu'un nouvel exemple se présente, la première règle qui peut s'appliquer lui propose une valeur de classe.

Ce procédé a été raffiné par la méthode CMAR (Li, Han, Pei 2001) (Classification based on Multiple class-Association Rules) qui ne se contente plus d'une seule règle pour prendre la décision de classification. Les règles sont cette fois-ci mesurées par un indice de corrélation fourni par un χ^2 pondéré par le maximum χ^2 possible. On évite également la redondance des règles en ne conservant que celles qui sont à prémisse minimale. Un nouvel exemple sera classé à l'issue d'un vote réalisé par toutes les règles qui s'appliquent, selon leur pondération.

4.3 Notre méthode

Pour nos expériences, nous avons implémenté une méthode proche de CMAR, capable de prendre en charge les règles *généralisées positives* (qui, en concluant sur un attribut de classe, entérinent la possibilité que l'exemple à classer appartienne à cette classe, si elle coïncide avec la prémisse) et les règles *généralisées négatives* (qui, en concluant sur la négation d'un attribut de classe, excluent la possibilité de classe correspondante) (Antonie, Zaïane 2004). Les règles *généralisées* étendent le principe des règles d'association en autorisant des négations d'attributs dans la prémisse, et sont obtenues à l'aide des motifs *k*-libres (Calders, Goethals 2003). Les règles positives sont de la forme $X\overline{Y} \rightarrow c_i$, les règles négatives de la forme $X\overline{Y} \rightarrow \overline{c}_i$, où c_i est un attribut de classe. Ces règles s'appliquent pour classer tout exemple qui contient le motif *X*, mais aucun des attributs de *Y*.

Selon le modèle de CMAR, les règles sont pondérées par le χ^2 relatif. Pour un nouvel exemple, les règles positives voient leur pondération s'ajouter au score, les règles négatives soustraient leur pondération. Au final, la classe avec le meilleur score est désignée.

5 Evaluation des résultats

5.1 Des résultats encourageants …

… mais insuffisants. Nous avons actuellement un taux moyen de 34 % de précision sur le français lors de la classification automatique. C'est déjà mieux que le hasard[1], ce qui semble

[1] Nous avons 14 classes différentes soit une probabilité de 7,1 % par classe.

montrer qu'il y a effectivement moyen de se servir de ces attributs, même si ce n'est pas encore suffisant, loin s'en faut, pour valider la méthode.

Des tests ont été effectués sur un corpus breton. Nous obtenons alors 29,40 % de réussite. Ces résultats sont à relativiser du fait de notre totale méconnaissance de cette langue et d'une projection du jeu d'étiquettes original pas forcément idéale.

À titre d'information, nous signalons que nos premiers tests ont donné 10 % de réussite alors que le deuxième et le troisième attribut (type du mot précédent / suivant) étaient regroupés en un seul attribut un peu moins informatif (nous nous contentions de vérifier que le mot était ou pas entouré de mots du même type). On s'aperçoit donc que les performances augmentent très vite grâce au simple ajout d'un attribut. Or, nous ne voyons pas de raison pour limiter le nombre d'attributs tant que chacun d'eux sera calculable sur le texte brut, sans apport de la moindre ressource extérieure, si de tels raffinements permettent l'amélioration des résultats.

5.2 Exemples de règles

Nous donnons ici des exemples avec les deux types de règles (positives et négatives).

D'abord une règle positive (rappelons qu'une règle positive nous permet d'affirmer que si les prémisses sont vérifiées alors on peut conclure sur une valeur de classe) :

type:v $\wedge$ avant:s $\wedge$ après:n $\wedge$ influence:o $\rightarrow$ classe = pronom personnel

Cette règle se lit de la manière suivante : si le type du mot étudié est vide, qu'il est précédé d'un signe de ponctuation, qu'il est suivi d'un mot de type indéterminé et qu'il influe sur les terminaisons alors on peut conclure que le mot est de la classe pronom personnel. Il faut noter ici que cette règle n'est pas la seule qui nous permette de retrouver les pronoms personnels. Globalement, nous disposons toujours de plusieurs règles qui nous amènent à conclure sur une classe donnée.

Ensuite, un exemple de règle négative (à l'inverse des règles positives, si les prémisses sont vérifiées, on peut conclure sur l'impossibilité de la valeur de classe correspondante) :

type:P $\wedge$ $\neg$après:P $\rightarrow$ $\neg$classe = déterminant

Cette règle indique que si le mot courant est de type plein et que le mot suivant n'est pas de type plein, alors le mot courant ne peut pas être un déterminant. En fait, cette règle est naturelle : un déterminant est un mot vide par définition et il est suivi d'un adjectif ou d'un nom, c'est-à-dire de mots pleins, dans la plupart des cas. Cette règle, parmi d'autres, est donc parfaitement en accord avec les connaissances que nous avons du domaine et confirme que notre démarche est cohérente.

Notons que quelque soit le type de règle dont on dispose, elles sont en concurrence les unes avec les autres. Cela signifie qu'une règle peut avoir une incidence faible par rapport à une autre règle de plus grande confiance.

Pour finir, nous voulons préciser que les 34 % de réussite ont été obtenu grâce à la combinaison de ces deux types de règles. Si nous n'avions utilisé que des règles positives,

nous aurions eu 17 % de réussite et avec les seules règles négatives, 23 %. Il s'agit donc ici d'un problème pour lequel la combinaison des deux types est impérative.

5.3 Utiliser les matrices de confusions

Ces matrices sont obtenues lors de la phase de comparaison des étiquettes calculées avec les étiquettes fournies dans le corpus étiqueté.

Elles nous donnent deux types d'information :

- la quantité de mots d'un type donné qui ont été correctement rattachés à ce type ;
- quels sont les mots qui ont été confondus avec un type donné.

Voici par exemple une des matrices obtenues lors de nos tests :

	Po	Adj	Inf	ProP	Nom	ProR	PPr	Ver	am	Det	neg	prep	Adv	Ppa
Po	33	0	0	0	0	0	0	0	0	0	0	0	0	0
Adj	0	0	0	0	15	0	0	0	0	0	0	1	2	0
Inf	0	0	0	0	9	1	0	0	0	0	1	0	0	0
ProP	0	0	1	2	0	1	0	0	0	5	0	2	0	0
Nom	0	1	0	1	56	0	0	0	0	1	1	4	0	2
ProR	3	0	1	2	1	7	0	1	0	3	0	0	0	0
PPr	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Ver	0	0	0	3	20	3	0	0	0	0	0	1	1	0
am	1	1	0	0	3	5	0	1	0	0	0	0	0	0
Det	4	2	1	9	7	7	0	0	0	6	0	0	0	0
neg	0	0	0	3	2	1	0	0	0	0	0	0	0	0
prep	3	0	0	16	8	15	0	0	1	2	0	2	0	2
Adv	1	0	0	1	11	0	0	2	0	0	0	1	0	0
Ppa	0	0	0	1	7	0	0	0	0	0	0	0	0	1

Figure 2 : Une matrice de confusion. En ligne, la classe réelle du mot, en colonne l'étiquette calculée, sur la diagonale, les mots correctement étiquetés.

On voit, sur cette matrice, qu'un grand nombre de mots d'autres types sont étiquetés comme des noms communs. Suite à cela, l'analyse des règles nous permet d'essayer de comprendre les raisons de cette confusion.

On voit aussi, et c'est essentiel pour comprendre une partie des problèmes rencontrés, que les classes ne sont pas du tout équilibrées, qu'il y a beaucoup plus de noms communs que de verbes, et plus de verbe que de pronoms…

5.4 Analyse de ces résultats

Il y a deux grandes raisons au manque de précision de ces résultats : un modèle linguistique insuffisant et un outil de fouille de donnée inadapté.

Dans le premier cas, il nous faudra trouver d'autres attributs qui permettront de mieux discriminer, nous pensons notamment utiliser la terminaison des mots, et regarder aussi la place du mot dans la phrase entière et non plus simplement dans le virgulot.

Quant à l'outil de fouille de données, il présente l'inconvénient de son avantage : il était immédiatement disponible et utilisable, mais il est fait pour des problèmes qui sont autres : ainsi, la disparité de volume des classes s'avère gênante. Il faudra donc améliorer notre outil pour mieux prendre en compte les spécificités du problème.

Notons enfin que nous regardons aussi d'autres méthodes de classification. Par exemple, un rapide test avec des arbres de décisions nous a donné un taux de réussite de 45 % avec des points réussis et des échecs très différents de notre méthode actuelle. Pour rester plus proche de notre démarche, nous avons utilisé un classifieur ensembliste, nous pourrions nous pencher sur des méthodes à base de motifs séquentiels qui ont la particularité de prendre en compte la séquentialité inhérente à l'organisation des mots dans un texte.

6 Conclusion

Nous avons posé les bases d'une méthode de validation des propriétés des mots qui présentera l'énorme qualité de ne pas nécessiter de ressources. Cette méthode semble donc particulièrement avantageuse pour des langues peu dotées. Elle permettra non seulement de travailler sans avoir besoin de ressources mais aussi, éventuellement, de créer les ressources qui manquent à moindre coût.

À partir d'un travail sur des propriétés des mots accessibles depuis le corpus brut, nous espérons pouvoir catégoriser correctement ces mots et utiliser nos résultats, soit directement pour faire un étiquetage morpho-syntaxique du corpus, soit pour la construction de ressources utilisables dans d'autres applications.

Nous n'en sommes bien sûr qu'aux prémisses de ce travail mais avons dégagé suffisamment de points encourageants et intéressants pour estimer qu'il y a là une piste en vue d'obtenir une solution originale et efficace pour les problèmes des langues sous-équipées informatiquement.

Remerciements

Merci à Tangi Ar Men, qui nous a fourni un corpus étiqueté breton.

Références

AGRAWAL R., MANNILA H., SRIKANT R., TOIVONEN H., VERKAMO A. (1996), Fast discovery of association rules. In *Advances in Knowledge Discovery and Data Mining*.

AGRAWAL R., SRIKANT R. (1994), Fast algorithms for mining association rules. In *Intl. Conference on Very Large Data Bases (VLDB'94), Santiago de Chile*.

ANTONIE M.-L., ZAÏANE O .R. (2004), An associative classifier based on positive and negative rules. In *9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD).*

ANTONIE M.-L., ZAÏANE O .R. (2004), Mining positive and negative association rules: An approach for confined rules. In *PKDD.*

BERMENT V. (2004), Méthodes pour informatiser des langues et des groupes de langues « peu dotées », Thèse de l'Université Joseph Fourier, Grenoble 1.

BERNARD G. (2003), Détection automatique de structures syntaxiques, Proceedings of the 8th *International Symposium on Social Communication*

BRILL E. (1992), A simple rule-based part of speech tagger, Proceedings of the *Third Conference on Applied Natural Language Processing*, Trento, Italy

BRILL E. (1994), Some advances in transformation-based part of speech tagging, Proceedings of the *twelfth national conference on Artificial intelligence*, Vol. 1, pp. 722-727

CALDERS T., GOETHALS B. (2003), Minimal k-free representations of frequent sets. In *Proceedings of PKDD'03.*

CHURCH K., MERCER R. (1993), Introduction to the special issue on using large corpora, *Computational Linguistics, volume 19, number 1, special issue on using large corpora: I, pp. 1-24, mars 1993.*

HOUBEN F. (2004), Mot vide, mot plein ? Comment trancher localement, Actes de RECITAL 2004, pp. 61-66

LI W., HAN J., PEI J. (2001), Cmar : Accurate and efficient classification based on multiple class-association rules. In *proceedings of the IEEE International Conference on Data Mining, ICDM 01*, San Jose, California.

LIU B., HSU W., MA Y. (1998), Integrating classification and association rules mining. In *proceedings of Fourth International Conference on Knowledge Discovery and Data Mining, KDD 98*, pages 80-86. AAAI Press.

LUCAS N., TURMEL L., CREMILLEUX B. (2003) Extraction d'associations pour la caractérisation de segments de textes en anglais avec et sans faute. *Conférence internationale sur le document électronique Cide 6.*

MARIAGE J.-J., BERNARD G. (2004), Catégorisation de patrons syntaxiques par Self Organizing Maps, Actes de *TALN 2004*,

VERGNE J., GIGUET E. (1998), Regards Théoriques sur le "Tagging". Actes de *TALN 1998*, pp. 22-31

ZIPF G. K. (1949), Human Behaviour and the Principle of Least Effort, New York, Harper.

Exemple d'écriture ignorée par Unicode : l'écriture tham du Laos

Grégory Kourilsky

Institut National des Langues et des Civilisations Orientales
2, rue de Lille 75007 Paris
gregory@kourilsky.com

Mots-clés : Unicode, écriture peu dotée, langue-π, écriture non linéaire, signes non connexes, écritures d'Asie du Sud-Est.
Keywords: Unicode, under-resourced scripts, π-language, non linear script, non-connex scripts, South-East Asian Scripts.

Résumé : L'écriture tham, employée pour noter les textes bouddhiques du Laos, est une *écriture* peu dotée informatiquement, alors que les *langues* qu'elle transcrit (lao et pali) ont fait l'objet d'une informatisation avancée. Considérant les raisons sociologiques et techniques de ce délaissement, nous proposons des pistes de réflexion pour y remédier en envisageant notamment un codage Unicode ainsi qu'une méthode de rendu.

Abstract: Tham Script, used to write Buddhist texts in Laos, is an under-resourced *script* although the *languages* that it transcribes (Lao and Pali) are well computerized. Understanding the sociological and technical reasons of this neglect, we'll propose some bases of reflection to mend it, by considering an Unicode Encoding and some rendering methods.

1 Un système d'écriture mal doté informatiquement

L'écriture tham est l'une des deux écritures officielles du Laos. Moins connue que l'écriture lao dont l'usage est courant, le tham est employé exclusivement pour noter les textes bouddhiques. Cette écriture, d'origine indienne et vraisemblablement établie par les Môn aux alentours du XV^{e} siècle (Ferlus, 1995), a la particularité de transcrire deux langues, le pali (langue du bouddhisme *theravada*) et le lao, avec des règles et certains caractères spécifiques à la notation de chacune de ces deux langues. L'écriture tham est considérée comme sacrée et son emploi est essentiellement réservé aux bonzes (moines bouddhistes) et aux lettrés[1].

[1] Le tham du Laos est très proche des écritures yuon (Nord de la Thaïlande), khün (États Shans de Birmanie) et lü (région du Xishuanbanna de Chine) si bien que l'on qualifie parfois l'ensemble de ces écritures d'« écritures tham », notamment parce que celles-ci sont surtout employées pour noter les textes bouddhiques (le terme thaï-lao *tham* dérive du mot pali *dhamma*, « doctrine bouddhique »). Précisons que les écritures khün, lü et yuon ne sont guère mieux dotées informatiquement que le tham du Laos (on note néanmoins l'existence de plusieurs polices TrueType yuon).

Écriture tham : ᨶᨾᩮᩣᨲᩔᨽᨣᩅᨲᩮᩣᩋᩁᩉᨲᩮᩣᩈᨾ᩠ᨾᩣᩈᨾ᩠ᨻᩩᨴ᩠ᨵᩔ [(2)]

Parmi les deux écritures officielles du Laos, le lao et le tham, seule la première a fait l'objet de travaux informatiques (création de nombreuses polices de caractères, prise en compte par Unicode, traitements de textes, etc.) tandis que la seconde a été presque totalement délaissée. En effet seulement trois polices existent actuellement pour saisir le tham : la *LaoDhamma* (TrueType) créée par S. Morey qui fonctionne sur Macintosh ; la *Vat Sène* (PostScript), élégante et très complète, fonctionne sur Macintosh et a été réalisée par l'École Française d'Extrême-Orient (ÉFEO). Cependant, cette dernière ne paraît stable que sur les logiciels In-design et X-Press et n'a pas été mise à la disposition du public. La troisième est la *ThamStandard* (TrueType) pour PC mise au point par nos soins et utilisée pour la rédaction du présent article[3]. Trois facteurs peuvent expliquer ce caractère anormalement peu doté de l'écriture tham :

- le tham est une écriture peu connue de la population laotienne. Elle est employée exclusivement par les moines bouddhistes à l'intérieur des pagodes,
- le tham est une écriture sacrée, possédant une fonction religieuse, voire magique. Cette dimension implique que cette écriture est plus qu'un système de transcription et son traitement informatique soulève des questions d'ordre socio-religieux,
- le système d'écriture tham est complexe : il transcrit deux langues très différentes (le pali et le lao), fait usage de signes non connexes et de nombreux signes souscrits et suscrits. Cette complexité est accentuée par le fait que cette écriture n'a jamais été l'objet d'une réelle codification, ce qui implique des graphies et des orthographes très variables selon les textes.

C'est de ce dernier facteur dont il sera question ici, les deux premiers étant traités dans (Kourilsky, 2005).

2 Trois difficultés pour une informatisation de l'écriture tham

2.1 Écriture non linéaire à signes non connexes

À l'instar des autres écritures d'origine indienne (devanagari, lao, thaï, khmère, etc.), on remarque en tham une hétérogénéité dans la disposition des signes-voyelles par rapport à la consonne. Exemples avec la consonne ᨠ *k*[4]:

Voyelle inhérente à la consonne : ᨠ /ka/,

Signe-voyelle à la suite de la consonne : ᨠᩣ /kaː/,

Signe-voyelle avant la consonne (signe *antéposé*) : ᨠᩮ /keː/,

---

[2] « Je salue le Bienheureux, le Vénérable, le Parfaitement éveillé » (Bizot, Lagirarde, 1996).

[3] Téléchargeable en ligne à l'adresse http://www.laosoftware.com.

[4] Les lettres tham isolées et les mots pali sont transcrits conformément à la transcription latine officielle du sanskrit fixée au 10e Congrès des orientalistes (1894), le pali suivant *grosso modo* les mêmes règles de prononciation que le sanskrit. Les mots lao écrits en tham sont transcrits en API et, le cas échéant, en écriture lao.

Exemple d'écriture ignorée par Unicode : l'écriture tham du Laos

Signe-voyelle au-dessus de la consonne : ◌ᩥ /kiː/,

Signe-voyelle en-dessous de la consonne : ◌ᩩ /ku/,

Signe-voyelle de part et d'autre de la consonne : ᨠᩣ /koː/, ᨠᩮᩬ /kja/, etc.

Le tham fait usage également de consonnes de forme souscrite et suscrite :

À la suite de la consonne : ᨠᩫᩣ /klaː/,

Avant la consonne (signe *antéposé*) : ᨡᩕᩣ /kʀeː/,

Au-dessus de la consonne : ◌ᩢᩩ /kuŋ/,

En-dessous de la consonne : ◌ᩣᩴ /kim/.

La position des signes-voyelles et aussi des signes consonantiques souscrits ou suscrits (dits de forme *subjointe*) incite à nuancer l'affirmation que l'écriture tham s'exécute de gauche à droite (Bizot, 2001) comme le montrent les exemples suivants :

ᨠᩕᩉᩩ *ra b h ma* = « *brahma* »

ᨶᩥᨡᩕᩣ *i n or k o* = « *nikro* »

2.2 Deux langues, deux systèmes

Le tham est un système d'écriture qui permet de transcrire deux langues, le pali et le lao. Les règles et propriétés du système d'écriture tham varient selon que celui-ci note l'une ou l'autre de ces deux langues. Aussi, il arrive qu'un même mot s'écrive différemment en tham en fonction de la langue transcrite, par exemple pour un mot pali passé dans la langue lao : Ainsi *tanhā* , « désir » sera noté (1) ᨲᨶ᩠ᩉᩣ dans un texte pali mais (2) ᨲᩢᩉᩣ [ตัมหา] dans un texte lao.

En notation du pali, la consonne qui suit une consonne dévoyellée prend sa forme subjointe ((1) : ◌᩠ᩉ *h*). En notation du lao, c'est en position finale ou médiane qu'une consonne peut prendre sa forme subjointe ((2) : ◌ᩢ *n*). Pour cette raison, nous avons choisi d'employer la terminologie suivante :

- P-tham : système d'écriture tham pour la notation de la langue pali,
- L-tham : système d'écriture tham pour la notation de la langue lao.

De plus, les graphies de certaines consonnes subjointes diffèrent selon la langue transcrite :

Valeur	Forme nominale	Forme subjointe p-tham	Forme subjointe l-tham
k	ᨠ	◌᩠ᨠ	◌᩠ᨠ ou ◌ᩮᩬ
ṅ	ᨦ	◌᩠ᨦ	◌᩠ᨦ ou ◌ᩢ
n	ᨶ	◌ᩢ	◌ᩢ ou ◌ᩮᩣ
l	ᩃ	◌ᩫ	◌ᩫ

Figure 1 : Variations graphiques entre subjointes p-tham et l-tham

2.3 Une écriture non fixée

Comme l'écriture tham n'a jamais fait l'objet de réelle codification, on trouve des mots écrits avec des orthographes différentes selon les textes. « En fait, il est pratiquement impossible de trouver deux manuscrits identiques, car chaque nouvelle copie subit des modifications ou des corrections effectuées par un lettré qui estime, à tort ou à raison, être dans son bon droit[5]. » (Peltier, 2000). En fait les copistes privilégient avant tout l'harmonie esthétique, critère subjectif par excellence. Cela est surtout flagrant en l-tham :

ໜອກ /nɔ̌:k/	ໝອກ /mɔ̌:k/	ຫຼອກ /lɔ̌:k/	ໜວກ /nwǎ:k/	ໝວກ /mwǎ:k/

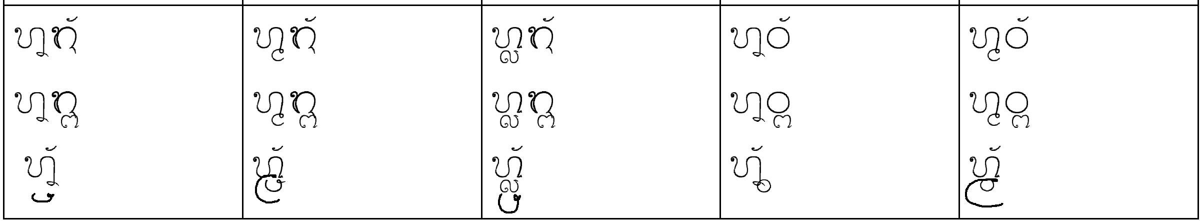


Figure 2 : Mots à plusieurs orthographes (Sena, 1957)

Ainsi, les copistes usent à l'envi d'innovations orthographiques, parfois même de simplifications dans l'écriture des mots qui ont pour objectif de réduire le nombre d'idiographèmes dans certains mots courants. On appelle ces formes simplifiées des *contractions* [ຄຳຫຍໍ້] /kʰamɲɔ̂/. Ces dernières, contrairement aux mots lao qui suivent normalement l'orthographe phonétique, ne sont pas toujours aisément déchiffrables et il est souvent nécessaire de les connaître par cœur pour déceler de quel mot il s'agit. Il faut également souligner que les contractions peuvent varier d'un texte à l'autre. Exemples :

[illegible] = [illegible] /di:lǐ:/ ; [illegible] = [illegible] /kʰanvà:/ ; [illegible] = [illegible] /anvà:/, etc.

L'aspect ultime de ces contractions se trouve dans le phénomène que nous avons appelé *glyphe syllabique*, où un signe graphique unique note une syllabe (autre que le cas consonne + /a/ bref inhérent) : [illegible] = [illegible] /lɛ/ ; [illegible] = [illegible] /rɯ:/, [illegible] = [illegible] /ao/, etc.

3 Solutions envisagées pour informatiser le tham

Nous avons en premier lieu réalisé une police de format TrueType que nous avons appelée *ThamStandard*. En tenant compte des signes multi-usages et des signes non connexes que nous avons décidé de considérer comme signes unitaires, nous avons besoin de 132 signes (i.e. *caractères*) accessibles au clavier. La police ThamStandard comporte ces 132 caractères, en plus d'une dizaine de variantes graphiques. Nous avons également conçu un clavier virtuel utilisable à partir d'un clavier AZERTY, nommé *ThamFrance* qui offre un maximum de correspondance phonétique entre les caractères latins et tham (par exemple ᨠ *k* se saisie avec la touche latine [k], ᩣ *ā* avec la touche [A], etc.). Si un tel système a l'avantage de ne nécessiter que la seule installation de la police *ThamStandard* et est donc utilisable sur n'importe quel ordinateur PC ou compatible,

[5] On peut aussi penser que le copiste souhaitait bénéficier de l'acquisition de mérites découlant de l'amélioration d'un manuscrit. Ce phénomène s'explique aussi par le fait que la copie des manuscrits est souvent confiée à des novices ou des jeunes moines qui ne maîtrisent que sommairement l'écriture tham.

l'accès à certains signes manque d'ergonomie. En effet, les touches du clavier sont insuffisantes pour saisir les 132 signes *minima* nécessaires pour représenter l'intégralité des lettres tham, et il faut alors recourir à des combinaisons de touches parfois tortueuses pour afficher certains signes. Beaucoup d'écritures indiennes et d'Asie du Sud-Est (devanagari et khmère en particulier) rencontrent le même problème[6].

Une manière de remédier au caractère peu doté de l'écriture tham tout en apportant une solution aux problèmes non résolus par la police *ThamStandard* et le système de saisie *ThamFrance*, serait d'intégrer le tham au Standard Unicode. Dans son objectif de normaliser les codages des caractères du monde entier afin de permettre une compatibilité entre les différentes polices et plates-formes, quel que soit le pays ou la langue, le Standard Unicode a établi des règles de codage pour l'ensemble des écritures figurant au Standard. Une de ces règles est de ne coder que les *caractères abstraits* (unités fondamentales de codage) et non les *glyphes*[7], ces derniers étant les représentations (parfois sous plusieurs formes) de ces caractères abstraits. Pour résumer, nous dirons qu'à un caractère abstrait correspond un ou plusieurs glyphes, et inversement plusieurs caractères abstraits peuvent être représentés par un glyphe identique. Ce système de codage permet (en théorie) de simplifier considérablement la saisie puisque c'est le système de mise en œuvre (par exemple l'Uniscribe de Windows) qui affichera les glyphes adéquats en fonction des contextes. En examinant le traitement par Unicode des écritures indiennes et dérivées (devanagari, tamoule, thaï, lao, birmane, khmère et tibétaine), qui font toutes usage de signes non connexes et de formes antéposées, souscrites et suscrites, on s'aperçoit que les problèmes ont été traités différemment selon les écritures. En fait on peut distinguer deux modèles : le modèle « indien » et le modèle « thaï-lao » qui divergent dans le codage des voyelles et dans l'ordre de stockage en mémoire[8].

Le modèle « indien » (devanagari, tamoul, khmer, etc.) suit à la lettre les principes d'Unicode en codant les voyelles en tant que phonèmes, ou reconnus comme tels dans les règles descriptives de ces langues. C'est ainsi que les voyelles composées de plusieurs idiographèmes (et quelle que soit leur position par rapport au signe consonantique) sont codées indépendamment, chacune d'elles représentant un *caractère unique*. Ainsi la voyelle tibétaine ི /i/ est-elle codée U+0F73, la voyelle khmère ើ /ɯɘ/ est codée U+0BCA, etc. Ce modèle respecte les principes d'Unicode mais présente des difficultés tant de mise en œuvre que d'utilisation (le khmer en particulier). D'autre part ce modèle code en mémoire les caractères vocaliques après la consonne, y compris les signes antéposés. Les consonnes subjointes antéposées sont stockées également après la consonne qui la précède en lecture.

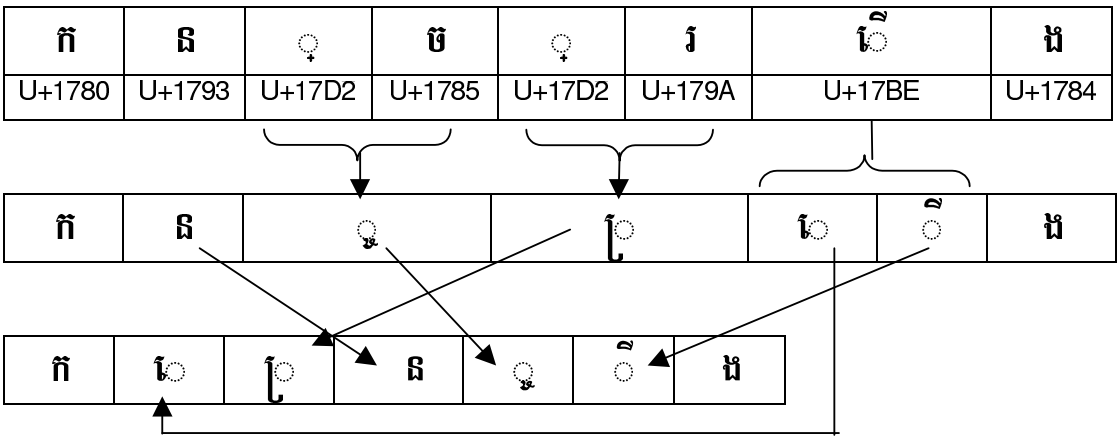
Exemple (d'après Bauhanh, 2002): saisie du mot khmer កាន់ក្រើង /kɑːncreiŋ/

---

[6] La police khmère *Kdol*, mise au point par l'ÉFEO et qui intègre un grand nombre de ligatures respectant les graphies de l'écriture khmère employées dans les manuscrits des XVII^{e}-XIX^{e} siècle, comporte 438 signes (Bizot, 1992).

[7] Le terme *glyphe*, désignant étymologiquement un signe gravé, était traditionnellement employé pour désigner les caractères maya avant d'être repris par les concepteurs d'Unicode.

[8] Nous assimilerons l'ordre logique (le stockage en mémoire) à l'ordre de saisie : « Les textes Unicode (...) sont stockés en mémoire en ordre logique. Cet ordre correspond *grosso modo* à l'ordre dans lequel le texte est saisi au clavier » (Andries, 2002).



On relève la complexité de l'ensemble des opérations que le système doit réaliser[9] :

- transformation des consonnes de forme nominale ◌ and ◌ en forme souscrite ◌[10] et ◌,
- repositionnement du glyphe ◌,
- décomposition de ◌ en ◌ et ◌,
- repositionnement de ◌ (juste avant ◌).

On pourra notamment remarquer l'opération pour le moins déroutante qui consiste à devoir décomposer ◌ en ◌ et ◌ , alors même que les concepteurs d'Unicode insistent pour coder le caractère U+17BE ◌ séparément. D'une manière plus générale, on ne peut que constater le manque de souplesse du mode de saisie des polices khmères Unicode. Pour cette raison, les concepteurs d'Unicode précisent que le système de saisie de l'écriture khmère n'est pas intuitive et doit faire l'objet d'une formation (Unicode 4.0, p. 277).

À l'inverse, le modèle « thaï-lao », mis sur pied avant l'intégration de ces écritures dans Unicode, suit un système local (la norme thaï TIS 620) qui code les idiographèmes vocaliques et non les signes-voyelles dans leur intégralité. Les caractères vocaliques sont donc des *signes visibles* et non plus obligatoirement des *phonèmes*. Les voyelles composées de plusieurs idiographèmes sont alors rendues par l'intermédiaire de plusieurs *caractères*. D'autre part, avec ce modèle, on saisit les signes dans un ordre purement graphique. Ce système déroge aux principes d'Unicode, mais offre des avantages pratiques. Exemple :

0EC0 ◌ + 0E9A ◌ + 0EB7 ◌ + 0EAD ◌ → ◌ /bɯaː/ (avec ◌ /b/ et ◌ /ɯaː/)

On comprend que le codage des caractères lao a été pensé suivant une logique graphémique, puisque le caractère U+0EAD ◌ LAO LETTER O est employé aussi bien pour saisir la consonne ◌ /ʔ/ que l'idiographème vocalique ◌ qui sert à écrire les voyelles ◌◌ /ɔː/ interconsonantique, ◌ /ɯa/ et ◌ /ɯːa/.

---

[9] Opérations auxquelles peuvent s'ajouter le rejet des saisies incorrectes, la correction automatique, etc.

[10] Voir *infra*, 3.2 pour le rendu des formes subjointes khmères. Notons que dans ce contexte, la consonne ◌ /r/ souscrite prend la forme ◌, à hampe plus allongée que le glyphe ◌ habituellement employé.

Exemple d'écriture ignorée par Unicode : l'écriture tham du Laos

Une réflexion sur une intégration du tham à Unicode permet d'envisager des solutions aux trois difficultés évoquées (*supra*, 2). Nous avons provisoirement attribué une zone de code tham (F000-F07F) dans la zone d'usage privé d'Unicode (E000-F8FF), en reprenant l'affectation proposée par V. Berment (Berment, 2004). Cette zone provisoire tham Unicode figure en fin d'article et ne doit être envisagée que comme situation temporaire en attendant l'intégration en bonne et due forme d'une zone tham définitive en accord avec le Standard Unicode[11].

3.1 Écriture non linéaire à signes non connexes

Pour le traitement des voyelles non connexes (*supra*, 2.1), nous avons donc à notre disposition un « modèle indien » qui code les voyelles en tant qu'unité phonologique, et un modèle « thaï-lao » qui code non les voyelles mais les *idiographèmes vocaliques* permettant de les composer graphiquement. Pour le principe de codage des voyelles tham, le « modèle thaï-lao » a été choisi en défaveur du « modèle indien », et ce pour deux raisons :

- le système adopté par le khmer, qui code effectivement toutes les voyelles, y compris les voyelles non connexes, nous a semblé trop complexe à mettre en œuvre et à utiliser, d'autant que l'écriture tham ne bénéficiera pas des moyens humains et financiers mis à disposition pour mettre en place le « khmer Unicode »,
- le système d'écriture l-tham est proche du système d'écriture lao (les deux systèmes notent de plus la même langue), et les graphies des voyelles tham (tant p-tham que l-tham) sont pour la plupart similaires à celles de l'écriture lao. Avec un système de codage de type « thaï-lao », les utilisateurs familiarisés avec le « lao Unicode » n'auront pas de difficulté à passer au « tham Unicode ».

Par conséquent, le mot ᨠᩢᩖᨾ /kjaʔ/ (avec ᨠ /k/ et ᨿᩖᨾ /jaʔ/) sera saisi :

F03A ᨠ + F000 ᨠ + F048 ᩢ + F041 ᩖ + F03B ᨾ → ᨠᩢᩖᨾ /kjaʔ/

et non : F000 ᨠ + F0XX ᨿᩖᨾ → ᨠᩢᩖᨾ /kjaʔ/

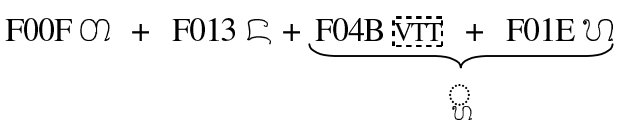
3.2 Deux caractères de rendu pour deux systèmes d'écriture

En p-tham, nous pouvons appliquer un système relativement simple pour le rendu automatique des consonnes subjointes. En effet, « la consonne au niveau de la ligne d'écriture est la finale de la consonne qui la précède (*i.e.* dévoyellée), tandis que la consonne placée au niveau inférieur est la consonne liée à la voyelle [qui suit] » (Sena, 1963). Si nous reprenons l'exemple du mot ᨾᩣᩃ᩵ᩉᩣ *tanhā*, la présence du ᩉ /h/ souscrit indique que la consonne ᩃ /n/ qui le précède est dévoyellée. On lira alors *tanhā* et non *tanahā*. Ainsi, il suffit de considérer la présence *théorique* d'un caractère équivalent au *virama* indien qui, associé à une consonne, indique que celle-ci perd sa voyelle inhérente. Le khmer Unicode use d'un tel caractère particulier pour rendre les souscrites, appelé caractère *coeng* (U+17D2), bien que celui-ci ne fasse pas partie de l'écriture khmère :

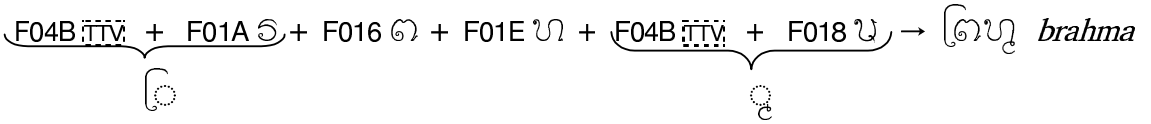
1785 ក + 17D2 ្ + 179A រ → ក្រ /craː/

[11] Une proposition de codage tham sera soumise au Comité technique d'Unicode (UTC) courant 2005.

Ce signe *coeng* fonctionne à la manière d'un *virama* en « tuant » la voyelle inhérente de la consonne précédente et indiquant que la suivante est subjointe. La règle p-tham de souscription des consonnes étant similaire à celle du khmer, nous pouvons procéder de même et considérer un *virama* « théorique », sans châsse et sans forme visible qui, associé à une consonne, affiche sa forme subjointe. Nous avons appelé ce caractère *VIRAMA THÉORIQUE THAM* (VTT). Exemple :



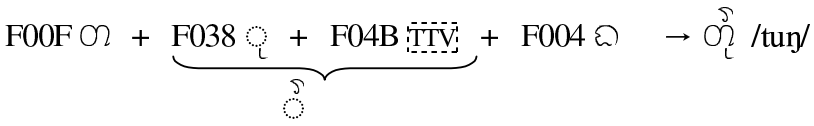
On notera le cas exceptionnel[12] du ◌ᩬ/R/ antéposé, qui sera saisi dans l'ordre graphique. Dans ce cas la logique linguistique est invalidée, le VTT n'indiquant pas que la consonne ᨻ est dévoyellée :

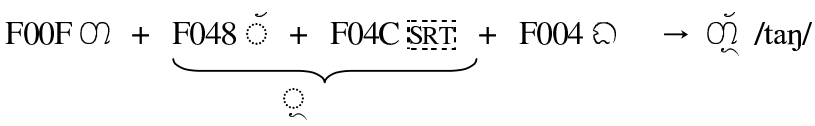


Pour la saisie du l-tham, nous pouvons conserver la méthode de rendu des glyphes p-tham par l'intermédiaire du caractère F04B VIRAMA THÉORIQUE THAM et considérer un outil supplémentaire de rendu de glyphes particuliers, sans se soucier de légitimité linguistique. Il est possible d'imaginer un deuxième caractère de rendu similaire au F04B VIRAMA THÉORIQUE THAM. Ce caractère, que nous appellerons SIGNE DE RENDU THAM, codé F04C à la suite du *virama* « théorique », permettra d'afficher les variantes graphiques de certains caractères. De cette manière, selon le glyphe qu'il souhaite voir afficher, l'utilisateur choisira librement l'un des deux caractères F04B VIRAMA THÉORIQUE THAM ou F04C SIGNE DE RENDU THAM.

+	F000 ᩋ	F004 ᩃ	F01B ᩛ	F01D ᩝ
F04B VTT	ᩋ᩠	ᩃ᩠	ᩛ᩠	ᩝ᩠
F04C SRT	ᩋᩴ	ᩃᩴ	ᩛᩴ	ᩝᩴ

Figure 3 : Rendu des variantes graphiques des formes subjointes

Exemples : 



3.3 Affichage des glyphes non contextuels

Il reste à résoudre la troisième difficulté évoquée, à savoir l'absence de codification de l'écriture tham. Un utilisateur doit être en mesure d'afficher les glyphes non contextuels (*i.e.* les formes

[12] D'autres cas particuliers sont à traiter mais ne seront pas pris en considération ici : le ◌ᩃ᩠ /ŋ/ final, les ligatures, etc. Ces signes sont également rendus par l'intermédiaire du caractère F04B VTT (Kourilsky, 2005).

graphiques non imposées par le contexte, et dépendantes du goût du copiste). La méthode d'utilisation de deux caractères de rendu décrite ci-dessus permet de résoudre en partie ce problème. Les mots l-tham pouvant s'écrire avec plusieurs orthographes possibles comme /mɔ̆ːk/ (*supra*, 2.3) seront rendus soit avec le F04B VIRAMA THEORIQUE THAM (rendu ᩈ᩠ᨠ), soit avec le F04C SIGNE DE RENDU THAM (rendu ᩈᩮᩣᨠ)[13].

D'autre part, le caractère F04C SRT peut être utilisé pour noter des variantes graphiques tham, indépendamment de la langue transcrite (lao ou pali). Avec ce système, tout caractère a, dans une même police, *potentiellement* deux variantes graphiques pour ses subjointes. Il pourra être intéressant pour un utilisateur souhaitant respecter au mieux les graphies originales d'un texte de disposer d'un éventail de choix dans une police unique. Nous donnons ci-dessous des exemples de saisie de différents glyphes rendus avec les caractères F04B VTT et F04C SRT.

+	F00C ᩌ	F018 ᩘ	F012 ᩒ	F01C ᩜ
F04B VTT	ᩈ᩠ᩌ	ᩈ᩠ᩘ	ᩈ᩠ᩒ	ᩈ᩠ᩜ
F04C SRT	ᩈᩮᩣᩌ	ᩈᩮᩣᩘ	ᩈᩮᩣᩒ	ᩈᩮᩣᩜ

Figure 3 : Rendu des variantes graphiques des consonnes subjointes

Quant aux contractions, elles ne poseront pas de problème de rendu puisque toutes les formes subjointes des consonnes sont affichables par l'un des deux signes de rendu (F04B ou F04C) (*supra*, 3.2) et que tous les idiographèmes vocaliques sont codés (*supra*, 3.1). Exemple :

F00C ᩌ + F037 ◌ᩍ + F04C SRT + F01B ᩉ → ᩌᩍᩮᩣᩉ /diːĭː/

Enfin, le problème posé par les *glyphes syllabiques* (*supra*, 2.3) peut être facilement résolu, dans la mesure où nous avons opté pour un codage « graphique » suivant le modèle « thaï-lao » (*supra*, 3.1). Il suffit en effet de leur attribuer un code. Ainsi, l'utilisation des graphies ᩋ◌ᩮ /lɛ/ et ᩍᩮᩣᩁ /ruː/ dépendent du choix du copiste (*supra*, 2.3). Afin de pouvoir disposer librement (et simplement) de ces signes, il suffit de leur attribuer un code : F060 ᩋ◌ᩮ SYLLABE THAM LEI et F05F ᩍᩮᩣᩁ SYLLABE THAM RYY. Les signes ◌ᩮᩣ /ao/ (contraction de ◌ᩮᩣ) et ◌ᩮᩍ /ɛː/ (variante graphique de ◌ᩮ) seront également codés, respectivement F045 ◌ᩮᩣ SIGNE VOCALIQUE THAM LAO AO et F04A ◌ᩮᩍ SIGNE VOCALIQUE THAM LAO MAE EI.

[13] Parmi les orthographes mentionnées en 2.3., seule la troisième, plus rare, ne pourra être rendue.

Proposition pour une zone tham Unicode

F000	Tham								F07F
		F00	F01	F02	F03	F04	F05	F06	F07	
	0	᩠	ᩡ	ᩢ	ᩣ	ᩤ	ᩥ	ᩦ		
	1	ᩧ	ᩨ	ᩩ	ᩪ	ᩫ	ᩬ			
	2	ᩭ	ᩮ	ᩯ	ᩰ	ᩱ	ᩲ			
	3	ᩳ	ᩴ	᩵	᩶	᩷	᩸			
	4	᩹	᩺	᩻	᩼	᩽	᩾			
	5	᩿	᪀	᪁	᪂	᪃	᪄			
	6	᪅	᪆	᪇	᪈	᪉	᪊			
	7	᪋	᪌	᪍	᪎	᪏	᪐			
	8	᪑	᪒	᪓	᪔	᪕	᪖			
	9	᪗	᪘	᪙	᪚	᪛	᪜			
	A	᪝	᪞	᪟	᪠	᪡	᪢			
	B	᪣	᪤	᪥	᪦	TTV	ᪧ			
	C	᪨	᪩	᪪	᪫	TSR	᪭			
	D	᪮	᪯	᪰	᪱	᪲	᪳			
	E	᪴	᪵	᪶	᪷	᪸	᪹			
	F	᪺	᪻	᪼	᪽	᪾	ᪿ			

Conclusion

Le standard Unicode et les méthodes de rendu proposées par le consortium permettent de résoudre les difficultés inhérentes au traitement informatique des écritures indiennes et dérivées (nombre très important de signes, signes antéposés et non connexes, variantes graphiques d'une même lettre, etc.). Cependant certains partis pris, assez tentant au premier abord, de « logique linguistique », impliquent une mise en œuvre complexe. Nous avons souligné la complexité du système de rendu de l'écriture khmère en raison du stockage en ordre « logique », ce qui a retardé considérablement et pénalise encore l'utilisation du khmer Unicode. Il est permis de penser qu'une intégration de l'écriture tham au standard pourra pallier le caractère informatiquement peu doté de ce système d'écriture. Encore faut-il réaliser cette intégration par un codage et une méthode de rendus simples, intuitifs et accessibles à ceux, éditeurs, étudiants, chercheurs ou encore moines bouddhistes, qui en feront usage.

Références

ANDRIES P. (2000), *Introduction à Unicode et à l'ISO 10646,* Québec, Document numérique, Volume 6, n° 3-4, 51-88.

BAUHANH M. (2002), « Rendering the World's Complex Scripts: A Case Study in Khmer », 21th International Unicode Conference, Dublin, 20 p.

BERMENT V. (2004), *Méthodes pour informatiser des langues et des groupes de langues « peu dotés »*, Grenoble, Thèse de doctorat (Université Joseph Fourier), 277 p.

BIZOT F. (1992), *Le Chemin de Laṅkā*, Paris, ÉFEO, TBC i, Paris, 352 p.

BIZOT F., BIZOT C. (2001), « Écritures bouddhiques d'Asie du Sud-est », *in* :*Histoire de l'écriture, de l'idéogramme au multimédia* (dir. A-M. Christin), Flammarion, 149-153.

BIZOT F., LAGIRARDE F. (1996), *ปฏิบัติ, La pureté par les mots*, Paris, Phnom Penh, Chiang Mai, Vientiane, ÉFEO, Textes bouddhiques du Laos, 275 p.

FERLUS M. (1995), « Les circonstances de l'introduction de l'alphabet tham lanna », *in* : *La Thaïlande des débuts de son histoire jusqu'au XV^e siècle*, Premier Symposium Franco-Thai, 18-24 juillet 1988, Université Silpakorn, 101-109.

KOURILSKY G. (2005), *Éléments pour un traitement informatique de l'écriture tham du Laos*, mémoire de DREA (Institut des Langues et des Civilisations Orientales), 330 p. (à paraître).

PELTIER, A.R (2000), *ลี้ภัยบน, La fille aux cheveux parfumés*, Vientiane, IRC, 490 p.

SENA P.L. (1957), *แบบธรรมไอธรรมอ่านหนังสือทัม ธรรมเป็นพาสาลาว [Apprendre rapidement à lire les caractères tham dans les textes lao]*, Bangkok, Kramol Tirannasur, 80 p.

SENA P.L. (1963), *แบบธรรมไอธรรมอ่านหนังสือทัม ธรรมเป็นพาสาบาลี [Apprendre rapidement à lire les caractères tham dans les textes pali]*, Bangkok, Kramol Tirannasur, 51 p.

UNICODE (2004), *The Unicode Standard 4.0*, http://www.unicode.org.

La Déclaration Universelle des Droits de l'Homme : 329 langues pour la constitution automatique de corpus et de lexiques

Hubert Naets
Laboratoire d'Ingénierie de la Connaissance Multimédia Multilingue (LIC2M)
Commissariat à l'Énergie Atomique
Bat. 38-1 ; 18, rue du Panorama ; BP 6
92265 Fontenay aux Roses Cedex ; France
naetsh@zoe.cea.fr

Mots-clefs : langues-pi, corpus, world wide web, langues peu dotées

Keywords: pi-languages, corpus, world wide web, under-resourced languages

Résumé Dans cet article, nous présentons le prototype d'un système permettant la constitution automatique de corpus depuis le web à partir des 329 langues de la Déclaration Universelle des Droits de l'Homme. Ce système exploite au maximum la proximité existant entre les langues pour augmenter la précision des requêtes envoyées au moteur de recherche et éviter ainsi la récupération de documents écrits dans des langues de même famille. Cette démarche s'inscrit dans le cadre de la production de ressources linguistiques informatisées pour les langues minoritaires dont la survie dépend entre autres de ces données.

Abstract In this article, we present the prototype of a system allowing the automatic constitution of corpora from the web, with the 329 languages of the Universal Declaration of Human Rights as bootstrapping. This system exploits the proximity existing between the languages to increase the precision of the requests sent to the search engine and to avoid the crawling of documents written in a language of the same family.

1 Introduction

En 2000, dans un article du *Courrier* de l'Unesco[1], intitulé *6000 langues: un patrimoine en danger*, Ranka Bjeljac-Babic constatait que sur les quelque 6000 langues existant dans le monde, dix d'entre elles disparaissaient chaque année. Cinquante langues seraient donc mortes depuis la parution de son texte. Pour Bjeljac-Babic et d'autres, « une langue qui n'est pas employée sur Internet "n'existe plus" dans le monde moderne. Elle est hors circuit. Elle est exclue du "commerce" », dans un cadre où « la diversité des langues [est] perçue comme une entrave aux échanges et à la diffusion du savoir ».

Le système présenté dans cet article ne pourra certes pas entraver la disparition prévue de 50 à 90% des langues au cours de ce siècle mais a entre autres buts de fournir à certaines d'entre elles des ressources nécessaires à leur prise en compte dans des analyseurs morphosyntaxiques, des parseurs syntaxiques, des moteurs de recherche ou tout autre outil issu du monde du Traitement Automatique des Langues. L'objectif que nous essayons d'atteindre est l'automatisation de la production et de l'analyse des ressources, en limitant au strict minimum toute intervention humaine. Cela correspond à un besoin croissant des industries de la langue d'étendre rapidement leurs offres à de nouvelles langues, alors que ces entreprises ne disposent pas nécessairement, dès l'introduction de nouvelles langues dans leurs systèmes, des personnes compétentes pour ces langues. Les langues peu dotées en gagneront une reconnaissance implicite.

Le programme que nous nous sommes fixé pour correspondre à ces besoins comprend la collecte de corpus pour de nouvelles langues, l'analyse morphologique des formes de ces langues, ainsi que la constitution de dictionnaires multilingues et la mise à disposition de ces ressources sur le Web.

Dans cet article, nous nous centrons sur la constitution des corpus. Il convient de noter que le traitement ici proposé n'a pas encore été testé dans sa totalité.

2 Les systèmes existants de construction de corpus à partir du web pour les langues minoritaires

Kevin P. Scannell[2] (Scannell, 2003) a réalisé *An Crúbadán*, un crawler web spécialisé dans la constitution de corpus notamment pour des langues minoritaires. À partir de textes d'amorçage d'une centaine de mots, il combine plusieurs de ces mots pour générer des requêtes qui sont transmises à l'API de Google. Le moteur de recherche renvoie une liste de documents potentiellement écrits dans la langue cible. Ces documents sont récupérés depuis le web et sont traités à l'aide d'un ensemble de techniques statistiques afin de déterminer quels documents ou parties de ceux-ci sont écrits dans la langue recherchée. Le web crawler parcourt ensuite les liens présents dans les documents identifiés comme appartenant à la langue cible. Le nouveau corpus ainsi constitué sert à amorcer la génération d'un nouvel ensemble de requêtes. *An Crúbadán* est utilisé actuellement sur 136 langues.

CorpusBuilder [(Ghani, Jones, Mladeni'c, 2001) et (Ghani, Jones, Mladeni'c, 2003)], qui a le même objectif de constitution de corpus de langues minoritaires que *An Crúbadán*, possède

[1] http://www.unesco.org/courier/2000_04/fr/doss01.htm
[2] http://borel.slu.edu/crubadan/apps.html

l'architecture suivante : une phase d'amorçage du système est assurée au moyen de deux petits ensembles de documents : des documents pertinents pour la langue recherchée et d'autres non. Une méthode de sélection de termes issus de ces deux groupes de documents est utilisée pour générer des requêtes composées de termes à inclure (provenant des documents pertinents) et de termes à exclure (issus des documents non pertinents). La requête ainsi produite est transmise à un moteur de recherche. Le document de plus haut rang renvoyé par ce moteur est téléchargé et passé à travers un filtre d'identification de langue. Selon la classification du filtre, le document est ensuite ajouté à l'ensemble des documents pertinents pour la langue ou à l'ensemble des documents non pertinents. La base de documents est alors mise à jour et le processus réitéré. R. Ghani, R. Jones et D. Mladenic insistent particulièrement sur les techniques de génération de requêtes en en testant six différentes (sélection uniforme des termes, sélection basée sur les fréquences, sélection probabiliste basée sur les fréquences, *rtfidf*, odds-ratio et odds-ratio probabiliste). *CorpusBuilder* a été utilisé pour traiter le slovène, le croate et le tchèque.

3 Vue d'ensemble du système

L'architecture de notre système (figure 1), qui est dans l'ensemble assez semblable à celles de K. P. Scannell et de R. Ghani, R. Jones et D. Mladenic, repose essentiellement sur une volonté de traiter un maximum de langues ou d'idiomes en parallèle et, partant, sur la nécessité de distinguer le plus tôt possible des idiomes très proches et ce, de la façon la plus automatisée possible.

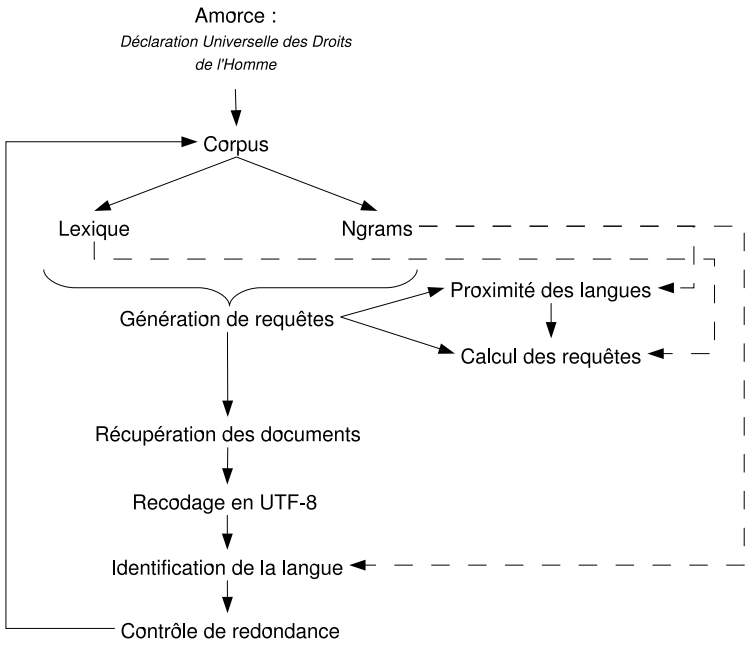


Figure 1: Schéma général du fonctionnement du système

4 La phase d'amorçage

À la suite de K. P. Scannell et de R. Ghani, R. Jones et D. Mladenic, nous partons d'un texte d'amorçage de relativement petite taille et présentant la caractéristique d'être actuellement traduit en 329 langues : la *Déclaration Universelle des Droits de l'Homme*[3] (DUDH). Parmi ces 329 versions, 275 sont au format texte, dans différents encodages de caractères, alors que les 56 documents restants sont représentés par une image (de type jpeg) ou par un fichier pdf, sans version texte. Les idiomes dans lesquels la DUDH est traduite sont très variés : des langues de grande ampleur (anglais, chinois, russe, espagnol,...) côtoient des langues régionales (breton, picard, galicien,...), des pidgins (du Nigeria par exemple) ou encore des idiomes presque éteints (l'arabela qui n'était plus utilisé que par 150 locuteurs en 1989 au Pérou, le yukaghir parlé par un nombre similaire de personnes en Russie, etc.). Nous avons remplacé un certain nombre d'images par des textes de la langue donnée, lorsque cela s'avérait possible.

Ces traductions de la DUDH, qui sont autant de corpus comparables, sont converties au format UTF-8 et servent à produire la liste des mots du texte, nécessaire pour la création des requêtes, ainsi que des ngrams de lettres permettant notamment d'identifier la langue des nouveaux documents. Les langues ne possédant pas de séparateurs sont découpées en bigrammes ou en trigrammes (en chinois par exemple) ou, lorsque cela est possible, en utilisant les changements de jeux de caractères (c'est le cas du japonais).

5 La génération des requêtes

La génération des requêtes constitue sans doute l'élément le plus important de la phase de constitution de corpus. Il s'agit en effet de produire des requêtes qui ramènent un maximum de documents dans la langue recherchée, à l'exclusion de toute autre langue. En d'autres termes, il s'agit d'atteindre les rappel et précision les plus importants possible. Il faut donc produire des requêtes composées de n mots spécifiques à une langue par rapport à des langues proches et éventuellement d'exclure m mots spécifiques aux langues proches de la langue recherchée.

> EXEMPLE : — Si l'on désire produire une requête permettant de découvrir des documents en asturien, on s'abstiendra de mettre dans la requêtes des formes telles que « que » ou « de » qui sont aussi très communes en catalan, en espagnol, en français, en galicien, en latin, en occitan auvergnat ou encore en occitan languedocien.

L'approche que nous proposons consiste donc à déterminer automatiquement, avant de générer la requête, quels sont les mots très fréquents et très spécifiques d'une langue par opposition aux langues proches. La fréquence importante de ces mots contribue à augmenter le nombre de documents découverts à l'aide d'un moteur de recherche ; quant à la spécificité, elle permet de s'assurer que le document (ou une partie de celui-ci) est bien écrit dans une langue donnée et non dans une autre.

Le problème peut être décomposé en deux sous-problèmes :

1. déterminer quelles sont les langues proches ;

[3] http://www.unhchr.ch/udhr/navigate/alpha.htm

2. déterminer les mots spécifiques à une langue par opposition à d'autres langues.

5.1 La proximité entre les langues

Le calcul de la proximité entre deux langues a pour objectif d'économiser un maximum de ressources et de temps : il est en effet peu efficace de déterminer le vocabulaire spécifique à l'asturien par rapport au japonais dans la mesure où l'intersection entre ces deux langues tend vers zéro, alors que l'asturien et l'espagnol ont une intersection beaucoup plus grande, ce qui implique une probabilité plus importante de confondre ces deux langues et donc de sélectionner erronément un document écrit dans une langue non recherchée.

Une approche naïve voudrait que l'on détermine la proximité de deux langues en comparant leurs lexiques. Plus le nombre de mots communs entre deux langues est important, plus les deux langues sont proches. C'est cependant sans tenir compte du fait que le corpus d'amorçage étant très restreint, il n'est absolument pas certain que tous les mots les plus fréquents d'une langue figurent dans le vocabulaire de ce corpus de départ. Il est donc nécessaire d'avoir un corpus d'une certaine taille avant que l'ordre des mots classés par fréquence ne se stabilise.

Une autre approche consiste à comparer les ngrams des deux langues plutôt que leur vocabulaire. Il s'agit de calculer les fréquences d'apparition de séquences de n lettres au sein du corpus. Ces séquences étant beaucoup plus fréquentes que des mots pris isolément, le modèle de langage se stabilise beaucoup plus rapidement et est donc utilisable avec de plus petits corpus.

C'est cette approche que nous avons sélectionnée pour la première version de notre collecteur de corpus. Nous sommes partis de l'algorithme de Cavnar et Trenkle (Cavnar, Trenkle, 1994), qui est utilisé également pour l'identification de la langue des documents (voir plus loin). Nous avons considéré que le corpus pour une langue donnée était le texte dont nous recherchions la langue. Une fois éliminée la meilleure langue possible pour le texte en question (cette meilleure langue est la langue du corpus), restent les n langues les plus proches, par ordre de proximité, où n peut être fixe ou variable.

5.2 Les mots spécifiques à une langue en vue de leur utilisation dans une requête

Comme nous l'avons dit précédemment, notre hypothèse est que, si l'on désire obtenir un maximum de documents pertinents pour une langue donnée, la partie positive de la requête doit être composée de termes les plus fréquents propres à la langue dont on recherche des documents ; quant à la partie négative, elle doit comporter les termes les plus fréquents dans les langues avec lesquelles la langue recherchée peut être confondue mais qui soient très peu fréquents dans la langue recherchée. Cette hypothèse correspond au score d'*odds-ratio*.

L'*odds-ratio* OR pour un mot m, étant donné le lexique d'une langue recherchée, $langue$, et le lexique confondu des n langues les proches de la langue recherchée, $languesproches$, se définit de la façon suivante :

$$OR = log_2(\frac{P(m|langue) * (1 - P(m|languesproches))}{P(m|languesproches) * (1 - P(m|langue))})$$

(Ghani, Jones, Mladeni'c, 2001) et (Ghani, Jones, Mladeni'c, 2003) comparent plusieurs méthodes de génération de requête et concluent que pour le slovène, le croate, le tchèque et le tagalog, la meilleure de ces méthodes (c'est-à-dire la méthode qui permet de récolter le plus de documents par requête) est celle des odds-ratios.

Exemple de requête produite pour l'espagnol :

+información +cinco +sentencias +ejecutado +partes -sans -au -peut -els -je -dels -pas -mais -il

6 L'identification des langues

L'étape d'identification des langues présentes dans les documents récoltés est la partie la plus sensible de ce système. Il s'agit en effet de s'assurer que chaque document est bien écrit dans la langue recherchée et d'éliminer un maximum d'éléments écrits dans une autre langue. Pour ce faire, nous avons réimplémenté l'algorithme de catégorisation de textes basé sur des ngrams de Cavnar et Trenkle (Cavnar, Trenkle, 1994), en restant compatible avec l'implémentation *TextCat* de van Noord[4]. L'identificateur de langue a été entraîné à l'aide de chaque corpus de départ constitué pour chaque langue et est dynamiquement réentraîné à chaque série d'ajouts de textes au sein du corpus. Il est utilisé pour identifier la langue générale du texte et la langue de chaque phrase de celui-ci. En complément, un identificateur de langue basé sur des tokens a également été implémenté. Même si ce type de détecteur de langue fonctionne en général moins bien (Grefenstette, 1995), il s'avère très utile lorsque le nombre de mots dont il faut identifier la langue n'est pas suffisamment important pour permettre à l'identificateur ngrams de décider.

La reconnaissance de la langue porte sur le texte dans sa totalité mais également sur chaque phrase et/ou partie du document d'origine (Prager, 1999). Les parties dont la langue est douteuse ou qui ne sont pas écrites dans la bonne langue sont rejetées.

7 Contrôle de la redondance des documents

La dernière opération consiste à vérifier qu'un même document n'a pas été récupéré plusieurs fois par le système. Ceci s'avère important pour les langues très peu présentes sur le web, particulièrement dans les premiers moments de la phase d'amorçage où le lexique de ces langues est encore très pauvre. Un simple vérification de l'URL du document ne s'avère en effet pas suffisant dans la mesure où certains textes, comme par exemple la *Déclaration Universelle des Droits de l'Homme* dans une langue donnée, sont réutilisés à plusieurs endroits du web, au sein de mises en page parfois différentes. Le contrôle de la réutilisation des documents permet d'éliminer ces textes et ainsi d'éviter de biaiser la constitution du vocabulaire du texte qui, autrement, serait sur-représenté pour certains termes et sous-représenté pour les autres.

Pour la première version du système, nous avons choisi d'utiliser la technique de chevauchement de ngrams (*ngram overlap*) (Clough et al., 2002) : pour un texte source A et un texte potentiellement dérivé B, représentés par les ensembles de ngrams $E_n(A)$ et $E_n(B)$, et la proportion de ngrams présents à la fois dans A et dans B, la similarité entre les deux textes est la suivante :

[4] http://odur.let.rug.nl/ vannoord/TextCat/

$$SIM_n(A,B) = \frac{E_n(B) \cap E_n(A)}{E_n(B)}$$

Cette méthode permet de déterminer si deux textes partagent le même vocabulaire et si les mêmes séquences de mots sont présentes dans le même ordre au sein des deux textes. Elle permet ainsi d'écarter les textes provenant de la même source.

Une fois passée cette dernière étape, les documents ou parties de documents sélectionnés sont intégrés au corpus de la langue. Le lexique et les ngrams sont recalculés pour l'ensemble du corpus et le processus complet est répété.

8 Premiers résultats

Le système n'a pas encore pu être testé dans sa totalité mais les premiers résultats concernant chaque partie s'avèrent néanmoins extrêmement encourageants.

Ainsi, en divisant chacun de nos corpus d'amorçage en deux, la première partie servant à entraîner l'identificateur de langues ngrams et la seconde à le tester, nous avons obtenu une identification correcte des langues dans 97,8 % des cas. Les erreurs concernent des langues extrêmement proches et pour lesquelles la taille du demi-corpus d'amorçage utilisé ici n'était pas suffisante. Il est possible que nous devions enrichier manuellement le corpus d'amorçage pour certaines langues. Par ailleurs, nous n'avons pas encore pu tester les effets de l'entraînement dynamique de l'identificateur ngrams à partir des nouveaux documents récoltés.

La technique des odds-ratios produit également d'excellents résultats. Nos premiers essais indiquent néanmoins que les pages web de certaines langues peu dotées sont parfois polluées par des langues de diffusion plus importante. Nous n'avons pas encore pu évaluer cette proportion ni l'efficacité des deux identificateurs de langue dans ce cas. De la même façon que nous prenons en compte les langues les plus proches dans le calcul des odds-ratios, nous envisageons de déterminer pour chaque langue la liste des langues les plus "polluantes". Nous ne savons pas encore par contre si nous utiliserons cette liste pour filtrer les pages lors de la génération des requêtes ou si nous l'emploierons pour renforcer l'identification des langues. Il est probable qu'il faudra choisir cette deuxième solution pour les langues extrêmement peu dotées, sous peine de réduire drastiquement la taille de leur corpus. Par ailleurs, d'autres idiomes de la DUDH semblent absents du web.

Le codage des documents a également posé un certain nombre de problèmes : le codage indiqué dans l'en-tête du document HTML ne correspond pas toujours au codage réel de ce document, ce qui qui a pour effet de produire des conversions inappropriées en UTF-8. Un outil maison est en cours d'évaluation pour tenter de résoudre — du moins partiellement — cette difficulté.

9 Conclusion

Nous avons présenté un système permettant de récolter des corpus de textes pour 329 langues — dont de nombreuses langues faiblement dotées —, en utilisant comme amorce la "Déclaration Universelle des Droits de l'Homme". La proximité entre certaines langues pouvant engendrer

des confusions lors de la sélection de documents pertinents pour une langue donnée, l'accent a été mis sur l'exploitation de cette proximité afin d'augmenter au maximum la précision du système. Ce système, dont la réalisation est presque terminée, n'a pu être testé jusqu'à présent que morceau par morceau mais semble d'ores et déjà être très prometteur.

Remerciements

Nous tenons à remercier particulièrement M. Gregory Grefenstette (CEA) pour ses nombreux conseils au cours de l'élaboration de l'extracteur de corpus.

Références

Cavnar W., Trenkle J. (1994), N-Gram-Based Text Categorization, *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, 161-175.

Clough P., Gaizauskas R., Piao S., Wilks Y. (2002), METER: MEasuring TExt Reuse. Proceedings of *the 40th Anniversary Meeting for the Association for Computational Linguistics (ACL-02)*, University of Pennsylvania, Philadelphia, USA, 152-159.

Ghani R., Jones R., Mladenic D. (2001), Building minority language corpora by learning to generate web search queries (Technical Report CMU-CALD-01-100).

Ghani R., Jones R., Mladenic D. (2003), Building minority language corpora by learning to generate web search queries, *Knowledge and Information Systems*.

Grefenstette G. (1995), Comparing Two language Identification Schemes, *JADT 1995: 3rd International conference on Statistical Analysis of Textual Data.*

Prager J. (1999), Linguini: Language Identification for Multilingual Documents, *Proceedings of the 32nd Hawaii International Conference on System Sciences.*

Scannell P. (2003), Automatic thesaurus generation for minority languages: an Irish example, *Proceedings of TALN 2003*, Batz-sur-Mer.

Languages of Myanmar in Cyberspace

Wunna Ko Ko[i], Dr. Mikami Yoshiki[ii]
Nagaoka University of Technology
Nagaoka, Niigata, Japan 940-2188
+81-258-46-600

1 Outlines of Selected Major Myanmar Languages

There are a total of 111 languages[1] spoken by the people living in Myanmar. Bible has been published over 50 languages spoken in Myanmar[2]. There are altogether 135 ethnic groups, also known as "Nationalities".[3] Some two or three ethnic groups speak same dialect or language and some ethnic groups speak more than one dialect. Among them, Burmese (Myanmar) is the official language and spoken by about 69 % of the population as their mother tongue. According to 1983 census,[4] the top language groups are Myanmar (Burmese) 69 %, Shan 8.5%, Kayin (Karen) 6.2%, Rakhine 4.5%, Mon 2.4%, Chin 2.2%, Kachin 1.4%.

Table 1: Selected Major Languages in Myanmar

Language	Language Group[5]	Region	Religion	Speaking Population	Scripts	Remark
Kachin (Jinphaw)	Tibeto-Burman	Kachin State	Mostly Christian	0.7 million	Latin scripts	Sample website: [6]
Kayin/ Karen	Tibeto-Burman	Kayin (Karen) State	Animism, Buddhism and Christianity	3.0 million	Karen scripts (extended Myanmar scripts)	Sample website: [7]
Chin	Tibeto-Burman	Chin State	Christianity, ethnic religion	1.0 million	Latin scripts	
Mon	Mon-Khmer	Mon State	Buddhism	1.2 million	Mon scripts (extended Myanmar scripts)	Sample website: [8]
Myanmar (Burmese)	Tibeto-Burman	Official language		34.5 million	Myanmar scripts	Sample website: [9]
Rakhine (Arakenese)	Tibeto-Burman	Rakhine (Araken) State	Buddhism	2.1 million	Myanmar scripts	Sample website: [10]

[i] Graduate School of Management and Information Systems Engineering, 045933@mis.nagaokaut.ac.jp
[ii] Professor, Management and Information System Science Department, mikami@kjs.nagaokaut.ac.jp

Shan	Tai- North West	Shan State	Buddhism	4.0 million	Shan scripts (extended Myanmar scripts)	Sample website: [11]

Figure 1: Map of Myanmar: States and Divisions

1. Thaninthayi 2. Mon 3. Yangon 4. Ayeyarwaddy 5. Kayin 6. Bago 7. Rakhine

8. Magwe 9. Mandalay 10. Kayah 11. Shan 12. Sagaing 13. Chin 14.Kachin

1.1 Myazedi Stone Inscriptions

Myazedi stone inscription, now displayed in Bagan Museum, discovered in 1886-87 by a German Pali scholar and Superintendent of the Epigraphic Office, Dr. Forchammar, is dated 1113 A.D or 474 Myanmar Era. Myazedi stone inscription was discovered near Myazedi Pogada at Myinkaba village the south of ancient Bagan (now, Bagan, Nyaung Oo Township, Mandalay Division). Another one discovered in the Ku Byauk Kyi Temple is now set up on the platform of Myazedi Pagoda. These two inscriptions are identical except the one display at the Bagan Museum is a square pillar of sand-stone and the other on the platform of Myazedi Pagoda is rectangular. The Myazedi inscribed pillar is a quadrangle, each side bearing an inscription in one of four different languages, Pali, Myanmar, Mon and Pyu. Myazedi is one of the earliest stone inscriptions so far discovered in Myanmar and the first to inscribe the Myanmar and Mon languages. Myazedi Stone Inscription, Ku Byauk Kyi Stone Inscription and Rajakumar stone Inscription (since both stone inscriptions were set up by Prince Rajakumar of Bagan Period) are the three names famous for these two stone inscriptions in Myanmar history. It is also known to many as the first Myanmar script.

Figure 2: Myazedi Stone Inscription

1.2 Kachin Language

Kachin Language, also known as Jinghpaw, is also one of the officially recognized minority languages of China. Most of the people who speak Kachin (Jinghpaw), live in Kachin State (Myanmar). Kachin Language uses Latin script. The alphabets used by Kachins were developed by Rev. Ola Hanson from American Baptist Foreign Mission Society purposely to write the Bible.

1.3 Kayin/Karen Language

There are two major types of Karen Languages, Pa'o (Pwo) Kayin (Karen) and S'gaw Kayin (Karen). Most of the people who speak Karen language live in Karen State (Myanmar), and some in Ayeyarwaddy division.

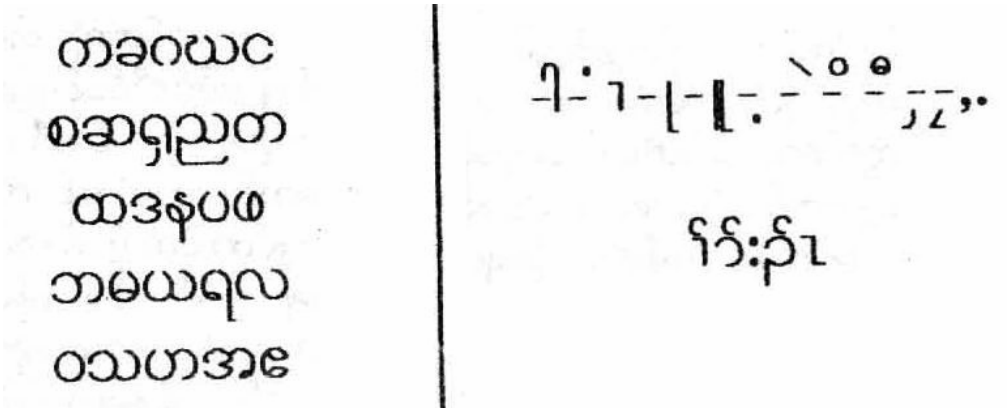


Figure 3: Kayin/Karen Alphabet: Consonants (left) and Vowels (right)

1.4 Chin Language

There are many different types of Chin Languages but mostly are quite similar. Examples are Chin-Asho, Chin-Falam, Chin-Haka, and Chin-Khumi.

1.5 Mon Language

Mon is recognized as one of the Mon-Khmer Monic group. The earliest Mon inscription, found at Lopburi in Thailand, dates from the eight century and is written in the Pallava script used at the Hinayana (Theravada) Buddhist center of Conjeeveram in the area of Madras on the east coast of India.

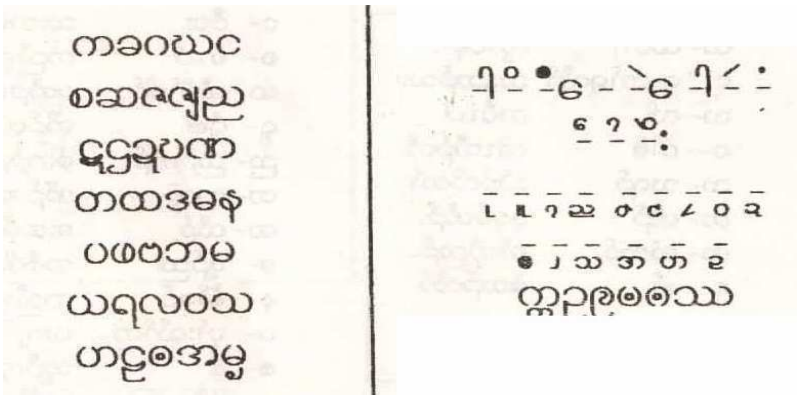


Figure 4: Mon Alphabet: Consonants (left) and Vowels (right)

1.6 Rakhine Language

Rakhine(Arakenese) language is recognized as one of the Tibeto ØBurman Lol-Burmese Southern group. Myanmar script is used.

1.7 Shan Language

Shan language is recognized as one of the Tai-Southwestern-East Central-Northwest group. The earliest reference to Shan scripts was found in a Bagan era[12] from AD1120.

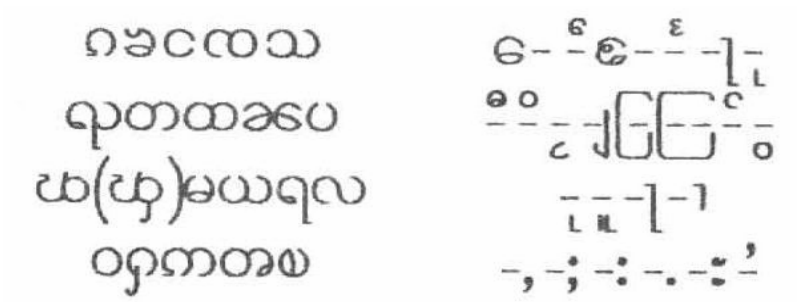


Figure 5: Shan Alphabet: Consonants (left) and Vowels (right)

1.8 Burmese (Myanmar)

Burmese (Myanmar) is the official language of Union of Myanmar (Burma). Myanmar (Burmese) is recognized as one of the Tibeto-Burman group. About 34.5 million people speak as first language and almost all the educated people in Myanmar speak as second language if their mother tongue is another ethnic language. The language of Education from primary level to secondary level is Myanmar (Burmese). From tertiary level is English. Mother tongue speakers of Myanmar/Burmese language live mostly in Myanmar Proper/Region of Myanmar, that is, seven divisions of Myanmar: Yangon, Mandalay, Sagaing, Tanintharyi, Ayeyarwaddy, Pegu and Magwe Division.

1.8.1 History of Burmese Scripts

The Burmese script derives from 11th century Mon. In A.D 1057 one of the first Burmese Kings, Aniruddha, conquered Thaton, a major Mon center, and the wise Monks, skilled artists and artisans were brought back with him to Bagan. The first inscription in Burmese dates from the following years and is written in an alphabet almost identical with Mon inscriptions. Aside from rounding of the originally square characters, this script has remained largely unchanged to the present.

1.8.2 Myanmar Language Commission

From the Rajakumar stone inscription (also known as MyaZedi stone inscription) which bear the date 475 Myanmar Era (1113 A.D.), it can be inferred that Myanmar language had developed into the written form at least more than eight centuries ago. However, although effort had been made by the monarch to establish orthography of the language and a standard orthography and a monolingual dictionary of the language were still lacking after a quarter of a century had passed since gaining independence from the British in 1948.

To correct this discrepancy and to establish the Myanmar language which is in fact the medium of the communication for the vast majority of the people either as their mother tongue or the most practical second language in this multi-ethnic society on a sound basis, the Revolutionary Council laid the corner stone for the future formation of the Myanmar Language Commission by forming the Commission for Myanmar Literature Translation in 1963. The Commission published "Burmese–Burmese dictionary" in 1968 and "The correct way of Burmese spelling" in 1978.

After a series of organizational and functional changes, the present commission assumed its present form in 1983, consisting of members appointed by the Cabinet and with a department under the Ministry of Education providing its management and operational requirements

1.9 Availability of Type-printing Technology in Myanmar

The Chronological order of History of Type Printing in Myanmar is shown in table 2.

Table 2: History of Printing and Publishing in Myanmar[13]

Year	Publication
1824 - 1826	[The First Anglo-Burmese War ended in a British victory, and Myanmar lost Assam, Manipur, Arakan and Tenasserim]
1836	The first English language newspaper published under British colonial area.
1842	The first ethnic [Karen] language newspaper under British colonial area.
1843	The Baptist mission published the newspaper in Moulmein under British colonial area.
1852	The English-Burmese dictionary published.
1852	[The Second Anglo-Burmese War ended also in British annexing Pegu province and the whole part of lower Myanmar (Burma) and renamed it Lower Burma.]
1853	The first publication by ethnic [Araken] language under British colonial area.
1854	The first publication under Burmese Monarchy: *The Life or Legend of Gaudama*.
1885	[The Third Anglo-Burmese War put the whole Myanmar (Burma) under British colonial.]
1869	The first Burmese-language newspaper published under British colonial area.
1875	-The first Burmese-language newspaper, named *Yadanabon Naypyi Daw* started publishing under Burmese Monarchy. -The first English-language newspaper, named the *Mandalay Gazette* started publishing under Burmese Monarchy.
1919	Burmese newspaper *Myanma Ahlin* started publishing.
1948	[Myanmar got independence]
1957	The *Mirror* Daily newspaper started publishing.
1963	Burmese newspaper *Loktha Pyithu Nezin* started publishing.
1964	- *Working People's Daily* started publishing - -The *Mirror* newspaper was nationalized and renamed as *Kyemon*.
1969	*Myanma Ahlin* newspaper was nationalized and later stopped publishing.
1993	-*Working People's Daily* was renamed *New Light of Myanmar*. -*Loktha Pyithu Nezin* was renamed *Myanmar Ahlin*

2 Availability of Digital Technology

Generally, there is no basic software in Burmese (Myanmar) language. Due to the lack of nationwide standardized encoding scheme in Myanmar Character set, even Myanmar Government's official website www.myanmar.gov.mm is presented in English only. But many efforts have been made on localization of basic software into Myanmar languages.

Table 3: History of Computerization in Myanmar[14]

1979	The first computer used in Myanmar at UCC (Universities' Computer Center); IBM PDB-11/70
1989	The first Computer company opened in Myanmar which provided training of usage of Software.
1992	The first Myanmar Font on Windows platform.
1993	- The technical report was submitted by Andy Daniels to Unicode Technical Committee. - Michael Everson reported to Unicode Consortium "Names of Burmese Characters: comment on Unicode Technical Report #1"
1994	Myanmar official government website www.myanmar.com launched.
1997	MCF (Myanmar Computer Federation) founded.
1998	- Limited dial-up IP connection service started by Government. - Unicode specification on Myanmar Languages, Myanmar Languages on U+1000 to U+109F[15]
1999	Amendment for Myanmar Languages in Unicode
2000	The Second Internet Service Provider, Bagan Cyber Tech established. www.bagan.net.mm.
2001	-Myanmar ICT Park opened. -The work program for Standardization of Myanmar Character Codes was presented by Pyone Maung Maung, Joint Secretary of National Task Force - Encoding of Myanmar Character Set and Implementation, Dr. Aung Maw, Myanmar Standardization Committee
2003	- Encoding Myanmar Language to Unicode, Zaw Win Aung, Project Administrator Burmese Language Support from Sourceforge.net. - Myanmar Unicode Test web page by Alan Wood - Myanmar Language Commission proposed "Myanmar Romanization"
2004	- Representing Myanmar in Unicode by Martin Hosken and Tun Tun Lwin - Analysis of Myanmar Keyboard by Myanmar NLP Research Team - Proposal of Myanmar Script Extension (Shan, Mon, Kayin (Karen) scripts) by Myanmar Unicode and NLP Research Center

2.1 Local Versions of Myanmar Font

Because of the unavailability of nationwide standardized encoding scheme in Myanmar language/script and extension of Myanmar scripts (ethnic scripts), Myanmar people try to develop local version of True Type Myanmar fonts which can easily run on Standard English version software. The first publicly used Myanmar font for Window platform, named *Shwe and Mya*, was developed around 1992. Some fonts were developed for Mac platform. But Mac is not a popular platform in Myanmar and these fonts are not widely used in Myanmar. The fonts are developed by using some graphic software and assign each character to the respective Latin key in ASCII code. These true type fonts can show the web pages correctly if they are downloaded and installed into the computer. But normally, the web sites use different fonts and just downloading one font is not enough for seeing all Myanmar web sites and also the ethnic languages. Normally, the font developers develop only the font of one language instead of developing font to use all

ethnic languages (most ethnic languages use extended Myanmar scripts). In other way, the ethnic languages also have to develop their own fonts. The sample WebPages which corresponds to three different Fonts for Burmese Language are shown in the reference. [16] The font that includes some ethnic scripts is *WinMyanmar* Font.

Most Myanmar True Type Fonts use the Myanmar Typewriter keyboard style which is the one most Myanmar people are familiar with. But it is not widely used in creating ethnic language web pages. The font *MyaZedi* is the one that has tried to be a Unicode standard font. In *MyaZedi* font, the Myanmar script graphics are not assigned to Latin ASCII codes. Instead, the Myanmar scripts are assigned to U+1000 as assigned by Unicode Standard, Version 4.0. The font designer added supplementary glyphs not included in Unicode but needed to represent ligatures and variant forms, and assign them to unused code points in the Myanmar block in Unicode. The only site which used *MyaZedi* font so far is www.etrademyanmar.com. Still in www.etrademyanmar.com, most of the pages are written in English.

Table 4: Currently Available Unicode Aware Myanmar Fonts

Font Name	Ethnic Script	Ligature
Padauk.ttf	No	Yes
Code2000	Yes	No
MyaZedi	Yes	Yes
fsex2p00_public	No	No

The unavailability of Unicode Myanmar keyboard and the inconsistent and non-standardized Myanmar fonts lead web developers to present information in English or in pdf, gif and jpg format if they try to present in Myanmar and ethnic languages.

3 Availability of Internet

Connection to internet has been available since 2000. Before that there was a Government organization that provides E-mail services, but this was mostly available for Government organizations and foreign missions. Some computer experts working at Government organizations got limited access E-mail services.

From the year 1998, the Government organization ISP provided dial-up internet Service with the maximum speed of 56 kbps. But due to the poor connection of the phone lines, the highest available speed at the user end is only about 32 kbps. Now, the total number of internet & email user through the Myanmar Post and Telecommunication (MPT), the Government organization, is about 5000 users.[17]

In 2000, another ISP, Bagan Cyber Tech, a semi-Government organization was established. It started providing dial-up service and some email service. Now, Bagan Cyber Tech provides also Broadband wireless (Yangon and Mandalay), IP Star Broadband satellite system and ADSL (Yangon and Mandalay). [18]

3.1 Development of Software in Myanmar Language

Myanmar software developers develop some software for local companies and mostly for international outsourcing companies. Even though the software is for local companies,

they use English as a language for medium of instruction. There are two main reasons: although Myanmar Language is the official language, the use of English is permitted[19] and is used in all the universities for all kinds of specialized subjects. This means the intellectuals do not hesitant to use English, and moreover, there is no standardized encoding scheme for Myanmar scripts. The use of English also makes easy for foreigners employed in Myanmar to use the locally developed software. Once, one company tried to use the Burmese/Myanmar version of locally developed software but the unavailability of interfacing with other software led to discontinuation of the use Burmese/Myanmar version.

3.2 Promoting Computer Usage for Local Languages

Local languages mostly used in computers are in the printing and publishing industry. The availability of using silk screen printing and off-setting force the publishing industry to use Myanmar Fonts and neatly printed computer print-outs. Before that, the high cost of hot printing limited the industry to government organizations and some ethnic printing houses. Using Myanmar Fonts in Computer to design for printing makes it less costly for the publishers and can produce a neatly designed product in comparison with using lead-based typefaces.

3.3 Natural Language Processing (NLP) Research Center[20]

The Natural Language Processing for Myanmar Language was formed in 1997 to develop technologies to use Myanmar Language on computer such as 1) localization of operating system and software, 2) spelling checking and grammar checking in Myanmar Language, 3) sorting in Myanmar Language, and 4) machine translation to/from Myanmar Language from/to foreign language. It formed under the guidance of Myanmar Computer Federation (MCF) and formed as NPO (Non-Profit Organization) and NGO (Non-Governmental Organization). It is formed with computer technicians and the representatives from private computer companies and government organizations. It is solely aimed to research for development of IT in Myanmar Language that the private companies cannot do by themselves alone.

4 Actions Needed for the Languages of Myanmar

4.1 Implementation of Myanmar Character Code

Although character codes for Myanmar Languages has been allocated in UCS/Unicode (U+1000 – U+109F), lack of implementation makes unavailable to local end users. Much effort had been made to develop Myanmar character codes and fonts by many international experts and local experts. They are working on their own and limited coordination between the experts, international organizations and venders slowed implementation. Bridging between international organizations, experts, venders and local end users may be the way to proper way to implementation of Myanmar Character Codes.

4.2 Extension of Myanmar Character Code

Except Kachin and Chin, other major ethnic languages have some common characters with Myanmar/Burmese language. There is not yet an official national standard for the encoding of Myanmar/Burmese Font. In setting Myanmar Font as official standard and in

Unicode, it is necessary to cover the ethnic languages as much as possible. Especially for Shan, Karen/Kayin and Mon languages, where scripts are near to Myanmar/Burmese scripts. The current encoding proposal to Unicode standard was prepared with consultation of experts from Myanmar NLP research center[21]. It may be advisable to include experts from linguistic area for these ethnic languages in developing fonts in Unicode.

4.3 Promotion of Ethnic Languages

Publications, both hard documents and soft documents (in cyberspace), in ethnic languages are in limited availability. The development of official standard font for these languages and widely access to Internet throughout Myanmar may lead not only to the development of these languages but make it easier that access opportunity of the ethnic minorities. Equal access opportunity should be secured for every language.[22]

References:

[1] Ethnologue, Languages of the World, 13th Edition, Barbara F. Grimes, Editor
[2] World Scriptures (www.worldscriptures.org)
[3] Myanmar, the country at a glance, http://www.unicef.org/myanmar/pages/a1.html, Accessed on March 8, 2005
[4] The latest census available, The figures are quoted from "Myanmar Population Changes and Fertility Survey 1991", published by Immigration and Population department, Ministry of Immigration and Population, 1995
[5] Ethnologue, Languages of the World, 13th Edition, Barbara F. Grimes, Editor
[6] http://www.mungga.tripod.com/gindai/
[7] http://www.kwekalu.net/
[8] www.kaowao.org/monversion/index.php
[9] http://www.ccdac.gov.mm/index.cfm?c=0
[10] www.rakhine.net/maharmuni.html
[11] www.mongloi.org/
[12] The History and Development of the Shan Scripts, Sai Kam Mong
[13] Chronology of the Press in Burma, by Irrawaddy, May 01, 2004, http://www.irrawaddy.org/aviewer.asp?a=3533&z=14 , Access on March 8, 2005
[14] www.myanmars.net/
[15] The Unicode Standard Version 3.0, The Unicode Consortium, www.unicode.org
[16] http://www.ccdac.gov.mm/index.cfm?c=0
http://www.etrademyanmar.com/
http://www.khitpyaing.org/
[17] MPT Data Communication department, http://www.mpt.net.mm/datacomm/index.html, Accessed on March 8, 2005
[18] http://www.bagan.net.mm
[19] 1947 Constitution "The official language of the Union shall be Burmese, provided that the use of the English language may be permitted." in Chapter 13, Article 216, General Provisions.
[20] Myanmar Unicode and NLP research center, http://myanmars.net/unicode/index.htm, Accessed on March 8, 2005
[21] Proposal of Myanmar Script Extensions: Mon, Shan, and Karen, JTC1/SC2/WG2 N 2768, May 31, 2004
[22] UNESCO Recommendation, "Concerning the promotion and use of multilingualism and universal access to cyberspace, 2003"

Approche pour un étiquetage morphosyntaxique du malais

Bali Ranaivo-Malançon

(1) Unit Terjemahan Melalui Komputer – Universiti Sains Malaysia
11800 Minden, Penang, Malaysia
ranaivo@cs.usm.my

Mots-clés : étiquetage morphosyntaxique, malais, corpus écrit
Keywords: POS tagging, Malay, written corpus

Résumé Dans cet article, nous proposons une méthode semi-automatique pour catégoriser les mots du lexique malais, langue officielle de la Malaisie, en partant d'un ensemble d'étiquettes très rudimentaires et d'une exploitation maximale de la structure morphologique des mots.

Abstract In this article, we propose a semi-automatic method to categorise words in Malay, official language of Malaysia. To achieve our aim, we start with a very simple tagset and take advantage of the morphological structure of words to predict their categories.

1 Introduction

Le malais, langue officielle de la Malaisie, n'est pas vraiment une langue sans ressource et un sujet relatif à cette langue pourrait ne pas apparaître dans cet atelier sur le « TAL et langues peut dotées ». Le malais est présent sur Internet (la recherche sur Google du mot « melayu » sur les pages éditées en Malaisie annonce 160000 pages) et il a fait depuis quelques années l'objet de recherche en traduction automatique dans l'unité de traduction automatique (UTMK[1]) de l'Université Scientifique de la Malaisie (USM) en coopération avec l'équipe GETA-CLIPS-IMAG de Grenoble, création de dictionnaire informatisé multilingue (FEM[2]), de base de données lexicales multilingue (Papillon[3]), traitement de la parole (principalement à l'Université Technologique de la Malaisie et l'UTMK). Toutefois, nous classons le malais dans le groupe des langues non pas peu dotées mais faiblement dotées pour la simple raison que le nombre de ressources linguistiques est très faible ou trop spécialisé (les dictionnaires utilisés par le système de traduction en ligne de MIMOS[4] sont des dictionnaires relatifs à l'agriculture et à la santé) et que les outils pour le traitement automatique du malais sont non réutilisables car implémentés sur des systèmes désuets et utilisables uniquement pour une seule application.

Nous avons adopté une approche empirique faisant appel à un corpus écrit et à des outils statistiques pour établir la liste de mots malais servant d'entrées pour un dictionnaire électronique formalisé unilingue du malais en cours de développement à l'UTMK et ajouter les catégories morphosyntaxiques à ces entrées. Ces informations grammaticales sont indispensables pour toute analyse automatique de textes malais. Aujourd'hui, aucun consensus n'existe sur la classification des mots malais et le *Dewan Bahasa dan Pustaka* (DBP), l'équivalent de l'Académie française en Malaisie, n'a toujours pas édité le dictionnaire *Kamus Dewan* [1], réactualisé avec les catégories lexicales.

Dans cet article, nous proposons une méthode semi-automatique pour commencer la classification des mots du malais en partant d'un ensemble d'étiquettes très rudimentaires et d'une exploitation maximale de la structure morphologique des mots.

---

1 UTMK : http://utmk.cs.usm.my/.

2 FEM – Français-English-Malay Dictionary : http://www-clips.imag.fr/geta/services/fem/.

3 Papillon project Web site : http://www.papillon-dictionary.org/.

4 MIMOS est une agence gouvernementale de recherche et de développement spécialisée dans le domaine des technologies de l'information et de la communication et de la microélectronique : http://www.mimos.my/.

2 Pourqoui l'étiqueteur grammatical du malais n'existe-t-il pas ?

2.1 Des projets trop orientés

Les recherches sur le traitement automatique du malais sont nombreuses et très variées en Malaisie. Le centre de traduction automatique (UTMK) de l'Université des Sciences de la Malaisie (USM) a eu un rôle de pionnier et reste la référence. Si la traduction automatique a été le point de départ de l'UTMK, actuellement ses axes de recherches se sont multipliés (communication homme machine, traitement de la parole, moteur de recherche, recherche d'information) avec une orientation vers la création de systèmes commercialisables (directive donnée par le Gouvernement malaisien) au dépend malheureusement de la recherche fondamentale. Un frémissement de recherche sur la reconnaissance de la parole du malais s'est fait sentir du côté de la Faculté d'Ingénierie Electrique de l'UTM depuis 2001. Le laboratoire d'ingénierie des langues de MIMOS, a créé 'FASIH' le premier synthétiseur malais (synthétiseur à diphones construit avec l'aide de MBROLA, Faculté Polytechnique de Mons, Belgique), et a mis en ligne un traducteur automatique malais-anglais (moteur fourni par l'UTMK).

Si beaucoup de projets concernant le traitement automatique du malais ont vu le jour en Malaisie, la majorité des données linguistiques et des outils créés ont été destinés à une seule application. Ainsi, les travaux sur la traduction automatique du malais auraient du avoir un étiqueteur grammatical du malais utilisable pour d'autres applications. Jusqu'à présent, chaque chercheur essaie d'adapter un étiqueteur créé pour d'autres langues sans jamais chercher à créer ou à adapter entièrement cet étiqueteur pour le malais. La raison généralement évoquée est le manque de références théoriques linguistiques. Le problème majeur que rencontre un taliste travaillant sur le malais est la rareté de ressources linguistiques. Le nombre d'années de recherche sur le traitement automatique du malais mené au sein de l'UTMK pourrait être un signe de présence de beaucoup de ressources linguistiques (dictionnaires électroniques formalisés, grammaires, bases de données lexicales, corpus) et d'outils. Malheureusement, la grande majorité de ces ressources et de ces outils est perdue ou inutilisable dû à un oubli de sauvegarde, un changement de plateforme ou tout simplement une absence totale de documentation.

2.2 Manque de linguiste informaticien

Un autre problème qui ne permet pas d'avoir un développement correct du traitement automatique du malais en Malaisie est le manque de linguiste informaticien. Il est souvent fréquent de voir dans le curriculum vitae d'un chercheur enseignant malaisien, la mention de '*natural language processing*' ou '*computational linguistics*' dans la partie 'Spécialité' bien que ces personnes n'aient eu qu'une formation partielle, parfois même inexistante, du traitement automatique des langues. Actuellement, il n'existe aucune formation complète du TAL en Malaisie. Le Département Informatique de l'USM offre deux cours de TAL en licence (cours optionnel) et en master. Le cours proposé pour la licence a été créé depuis deux ans et n'a été ouvert que cette année avec seulement six étudiants inscrits. Le cours porte sur l'introduction au TAL. Le cours en master, 'Traitement automatique de documents' (*Document Processing*), enregistre plus d'étudiants (une moyenne de vingt étudiants) mais ne crée pas plus de vocation chez les étudiants. Le Département Sciences Sociales offre un cours

de traduction automatique où les étudiants apprennent le fonctionnement des traducteurs automatiques et l'utilisation de Prolog avec des enseignants non informaticiens et non talistes. Quelques universités malaisiennes offrent un seul cours facultatif de TAL ou de linguistique computationnelle qui est généralement une brève introduction au sujet.

2.3 Rareté des références théoriques et inapplication des normes

Mis à part ces problèmes de gestion des données et de formation de spécialistes du TAL, le traitement automatique du malais souffre d'un problème politique lié à la gestion de la langue. Les normes grammaticales et orthographiques du malais sont énoncées et dictées par DBP à travers des guides pour écrire correctement le malais [2, 3, 4] ou transcrire les mots empruntés [5] et l'ouvrage de grammaire du malais [6]. Le problème est que ces règles sont rarement appliquées. L'exemple le plus frappant est l'utilisation du « di ». En tant que préposition, cette unité correspond à un mot orthographique : *di mana* 'où', *di sekolah* 'à l'école'. En tant que marqueur du passif, il est préfixé à la base : *dipakai* 'Passif-utiliser'. Les règles d'utilisation du « di » ont été énoncées depuis 1972 et expliquées clairement dans la grammaire de référence scolaire, le *Tatabahasa Dewan*. Il est encore très fréquent aujourd'hui de voir dans des lettres officielles une utilisation erronée du « di » : le préfixe est séparé d'un blanc de sa base, la préposition est jointe au mot qui le suit. Nous n'entrerons pas trop en détail dans l'utilisation des ponctuations, principalement le tiret et l'apostrophe : absence du tiret dans une forme réduplicuée (**rumahrumah* au lieu de *rumah-rumah* 'maisons') ou dans l'affixation de mots empruntés (**mengupgradekan* au lieu de *meng-upgrade-kan* 'mettre à jour'), maintien de l'apostrophe dans des mots d'origine arabe (**Juma'at* au lieu de *Jumaat* 'vendredi'), oubli de l'apostrophe dans les formes tronquées (**kan* au lieu de *'kan*, forme réduite de *akan* 'marque du futur'), etc.

3 Constitution d'un corpus écrit

Tous les problèmes évoqués précédemment nous ont amenée à orienter nos recherches sur la classification et l'étiquetage morphosyntaxique des mots malais vers une approche empirique permettant de pallier l'absence de référence linguistique théorique et de norme.

Il y a quelques années, un projet mené à l'UTMK en collaboration avec le DBP a abouti à la création d'un corpus écrit comprenant des textes littéraires et académiques. Une grande partie de ce corpus est malheureusement égarée ou inutilisable (problème de plateforme). Nous avons regroupé les textes utilisables, les avons mis pour le moment sous format de fichier texte et fait corriger manuellement. A cet ensemble de textes littéraires, nous avons ajouté des textes journalistiques provenant d'un journal national malaisien (*Bernama*). Le tableau suivant (Figure 1) montre la composition et la taille du corpus.

Nombre de fichiers	Genre	Année de publication	Nombre d'occurrences de mots
1904	Textes journalistiques	2003	551974
11	Ouvrages littéraires	1975-1992	950398
9	Publications académiques	1986-1999	586145
2	Manuels d'utilisation	1995-2000	166486
		TOTAL =	2255003

Figure 1 : Description du corpus

Nous continuons de télécharger sur Internet des articles de journaux (*Utusan Malaysia*[5] et *Berita Harian*[6]), des textes officiels et des articles publiés par des universitaires pour pouvoir tester nos hypothèses et mettre à jour les entrées du dictionnaire (le malais ne cesse d'emprunter des termes anglais et de créer de nouvelles formes affixées). Même si la fonction 'téléchargement' est automatique (nous utilisons le logiciel *Net Transport*), le choix des types de textes que nous nous sommes imposé nécessite la lecture partielle du texte et surtout l'identification de la langue. Le malais et l'indonésien sont deux langues très proches. Les identificateurs de langues en ligne ne font pas toujours la différence entre les deux. Nous avons créé un identificateur de textes écrits malais et indonésiens qui sera disponible sur le site de l'UTMK avant la fin du mois de juin de cette année.

La correction des textes était planifiée en deux étapes : correction automatique et correction manuelle. Le test effectué avec le correcteur orthographique *DewanEja* s'est avéré non satisfaisant ce qui a ramené la correction à une tâche manuelle fastidieuse et longue. Il existe actuellement deux correcteurs orthographiques commerciaux, *EjaTepat* (Accredo Multimedia Sdn. Bhd.) et *DewanEja 3000* (The Name Technology Sdn. Bhd.). L'UTMK avait déjà créé vers la fin des années 80 un correcteur orthographique du malais, tournant sur Macintosh et Microsoft Windows sur un IBM PC. Un des projets en cours de l'UTMK est le rafraîchissement de ce correcteur orthographique trop obsolète et l'ajout d'un composant pour la correction grammaticale.

---

[5] Utusan Malaysia : http://www.utusan.com.my/.

[6] Berita Harian : http://www.bharian.com.my/

4 Etiquetage du corpus

Les entrées du dictionnaire *Kamus Dewan* (format papier et électronique) ne sont pas catégorisées. La présence des bases liées (bases simples qui n'apparaissent qu'avec un affixe ou seulement dans une composition) est une des raisons sans doute pour lesquelles ces entrées ne sont pas accompagnées de catégorie grammaticale. Le DBP est en train d'ajouter ces informations dans sa prochaine édition dont la date de publication reste incertaine. Certains dictionnaires, généralement bilingues (par exemple, *Kamus Dwibahasa Oxford Fajar*, *Collins Headstart Easy Learning English-Malay bilingual dictionary*), ont commencé à ajouter les catégories grammaticales mais comme ils ne sont pas issus du DBP, ils ne sont pas reconnus. Ils restent cependant une aide pour la classification des mots malais.

4.1 Premier jeu d'étiquettes

La catégorisation des mots d'une langue repose sur un ensemble d'étiquettes défini, non ambigu et si possible ni trop grand ni trop petit. Une petite recherche sur le Web en découragera plus d'un, vu le nombre de jeux d'étiquettes disponibles. Le problème est de donc de choisir le jeu d'étiquettes qui s'adapte le plus à la langue de recherche. La question reste ouverte car comment savoir lequel est le plus adapté ? Nous avons choisi de commencer avec les étiquettes « traditionnelles » présentées dans la grammaire du DBP puis de trouver une correspondance entre ces étiquettes et le jeu d'étiquettes proposé par exemple par EAGLES[7] et MULTEXT. À la fin, nous espérons proposer un jeu d'étiquettes du malais réutilisable et comparable à un jeu d'étiquettes standard.

Nous proposons un premier jeu d'étiquettes qui va permettre d'amorcer la catégorisation morphosyntaxique du corpus. L'ensemble proposé est une réorganisation des catégories proposées par le DBP. Les classes « adverbes » et « ponctuations » y ont été ajoutées.

1. Verbe (V) : Verbe transitif (VT) ; Verbe intransitif () ; Verbe actif (VA) ; Verbe passif (VP) ; Verbe transitif actif (VTA) ; Verbe transitif passif (VTP)
2. Adjectif (A)
3. Nom (N) : Nom propre (NP) ; Nom commun (NC)
4. Pronom (P)
5. Déterminant (D)
6. Adposition (AD)
7. Conjonction (C)

---

[7] EAGLES – Recommendations for the Morphosyntactic Annotation of Corpora. http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html.

8. Interjection (I)
9. Adverbe (AV) : Négatif (AVN); Affirmatif (AVA); Interrogatif (ADI); Intensificateur (AVT)
10. Ponctuation (PO)

Nous illustrerons de manière simple les différentes étapes de notre approche à partir de trois phrases tirées du dictionnaire *Kamus Dewan*. Ces trois phrases nous servirons de fil rouge au travers de l'étiquetage des mots non ambigus, des mots affixés et à clitiques et de l'étiquetage par règles.

Exemple 1: *Buku ini baik dibaca oleh kanak-kanak.*

'Ce livre est bien pour être lu par les enfants'

Exemple 2: *Lebih baik engkau yakan sahaja akan kata-katanya.*

'Il vaut mieux pour toi que tu approuves ces remarques'

Exemple 3: *Kata-katanya itu terarah kepada lawannya.*

'Ces remarques sont dirigées contre ses opposants'

4.2 Étiquetage des mots non ambigus

Après avoir déterminé le premier jeu d'étiquettes du malais, la première étape dans l'étiquetage des mots du corpus consiste à étiqueter les mots appartenant à une seule catégorie donc non ambiguë. Cet ensemble de mots contient pour le moment les auxiliaires, les mots descriptifs, les pronoms, les déterminants, les conjonctions, les adpositions, les interjections, les adverbes et les numéraux. Au fur et à mesure que d'autres mots du corpus sont étiquetés, ils seront ajoutés à cet ensemble.

Dans les trois exemples cités précdemment, les mots non ambigus sont suivis de leur catégorie placée ici entre les deux symboles d'infériorité et superiorité.

Ex. 1: *Buku ini*<D> *baik dibaca oleh*<AD> *kanak-kanak*.<PO>

Ex. 2: *Lebih*<AVT> *baik engkau*<P> *yakan sahaja*<AV> *akan kata-katanya*.<PO>

Ex. 3: *Kata-katanya itu*<D> *terarah kepada*<AD> *lawannya*.<PO>

4.3 Étiquetage des mots affixés et des mots à clitiques

La liste des mots tirés du corpus a d'abord fait l'objet d'une correction manuelle afin d'éliminer les mots étrangers et d'isoler les abréviations et les noms propres. Le regroupement des mots par sous-chaînes partagées suivi d'une correction manuelle a permis

non seulement de mettre à jour toutes les formes dérivées d'une base, mais aussi de déterminer la propriété transitive d'un verbe actif préfixé par « me- ». Le dictionnaire *Kamus Dewan* ne cite pas comme sous-entrée la forme passive préfixée par « di- ». Si la famille morphologique d'une base contient à la fois les deux formes préfixées par « me- » et « di- », ces deux formes sont des verbes transitifs, le premier étant la forme active et le second la forme passive.

Les mots affixés malais peuvent être donc catégorisés par la reconnaissance des affixes qui les composent. Par exemple, les mots contenant le préfixe « me- » sont des verbes actifs et les mots contenant à la fois le préfixe « me- » et l'un des suffixes « -kan » ou « -i » sont des verbes actifs transitifs. La décomposition morphologique des mots est obtenue par l'analyseur de l'affixation du malais développé par Ranaivo [7]. Comme cet analyseur travaille sur des mots hors-contexte et que la liste des bases permettant la validation des découpages n'est accompagnée d'aucune information linguistique, certains mots affixés sont classés avec plusieurs étiquettes.

Dans les exemples 1 à 3, les mots affixés sont *dibaca*, *yakan* et *terarah*. Les deux premiers ont été reconnus par l'analyseur comme étant des verbes transitifs au passif (VTP). Le dernier a été analysé avec deux catégories possibles : verbe intransitif (VI) et adjectif (A).

Ex. 1 : *Buku ini*<D> *baik dibaca*<VTP> *oleh*<AD> *kanak-kanak.*<PO>

Ex. 2 : *Lebih*<AVT> *baik engkau*<P> *yakan*<VTP> *sahaja*<AV> *akan kata-katanya.*<PO>

Ex. 3 : *Kata-katanya itu*<D> *terarah*<VI;A> *kepada*<AD> *lawannya.*<PO>

La classification de certaines bases simples peut se faire par l'identification de certains clitiques. Par exemple, les mots servant de support aux clitiques « -ku » et « -mu » sont soit une de ces prépositions, *bagi*, *kepada*, *pada*, *oleh*, *untuk*, soit des noms.

- *Tidak terpikul olehku benda yang berat itu.*

  Négation – pouvoir porter – par moi – objet – Relatif – lourd – Déterminant

  Je n'ai pas pu porté cet objet lourd.
- *Adikku ialah seorang yang baik.*

  Frère/Sœur moi – est – une personne – Relatif – bien

  Mon frère / Ma sœur est une personne bien.

4.4 Étiquetage par règles

L'étiquetage du reste des mots du corpus peut s'effectuer soit en créant un étiqueteur soit en adaptant un étiqueteur existant. Cette deuxième solution n'est pas vraiment facile à réaliser. Les étiqueteurs utilisant un dictionnaire et des grammaires ne sont pas utilisables car le malais ne possède pas encore ces données linguistiques. Les étiqueteurs demandant un corpus étiqueté préalablement et manuellement ne sont pas non plus applicables car cela implique que le jeu d'étiquettes du malais soit déterminé alors que nous essayons de le construire. Un

autre obstacle à l'adaptation des outils existants est l'incompatibilité des plateformes et la difficulté d'utilisation car l'outil est écrit avec un langage de programmation inconnu du linguiste informaticien. Pour l'instant, la solution à ce problème n'est pas très claire car nous sommes encore dans la phase de test des étiqueteurs disponibles sur Internet.

Afin d'étiqueter le maximum de mots dans le corpus, nous avons établi quelques simples règles contextuelles.

- Règle 1 : Un mot précédant *ini* ou *itu* est un nom

Cette règle permet d'étiqueter les mots *buku* et *kata-katanya*.

Ex. 1 : *Buku*<N> *ini*<D> *baik dibaca*<VTP> *oleh*<AD> *kanak-kanak*.<PO>

Ex. 3 : *Kata-katanya*<N> *itu*<D> *terarah*<VI;A> *kepada*<AD> *lawannya*.<PO>

- Règle 2 : Un mot devant une adposition est un nom

Cette deuxième règle permet d'étiqueter les mots *kanak-kanak* et *lawannya*.

Ex. 1 : *Buku*<N> *ini*<D> *baik dibaca*<VTP> *oleh*<AD> *kanak-kanak*<N>.<PO>

Ex. 3 : *Kata-katanya*<N> *itu*<D> *terarah*<VI;A> *kepada*<AD> *lawannya*<N>.<PO>

- Règle 3 : Un mot devant un adverbe-intensificateur est un adjectif ou un verbe

Cette règle permet d'étiqueter le mot *baik*.

Ex. 2 : *Lebih*<AVT> *baik*<A ;V> *engkau*<P> *yakan*<VTP> *sahaja*<AV> *akan kata-katanya*.<PO>

Ces trois étapes (étiquetage des mots non ambigus, étiquetage des mots affixés et à clitique, étiquetage par règles) sont répétées jusqu'à ce que tous les mots du texte soient étiquetés ou que plus aucune règle ne soit applicable. Cette répétition du processus va permettre d'étiqueter les mot *baik* (exemple 1) et *kata-katanya* (exemple 2). Dans nos trois exemples, seul le mot *akan* (exemple 2) n'a pas pu être étiqueté et les deux mots, *baik* (exemples 1 et 2) et *terarah* (exemple 3), sont avec deux catégories.

Ex. 1 : *Buku*<N> *ini*<D> *baik*<A;V> *dibaca*<VTP> *oleh*<AD> *kanak-kanak*<N>.<PO>

Ex. 2 : *Lebih*<AVT> *baik*<A;V> *engkau*<P> *yakan*<VTP> *sahaja*<AV> *akan kata-katanya*<N>.<PO>

Ex. 3 : *Kata-katanya*<N> *itu*<D> *terarah*<VI;A> *kepada*<AD> *lawannya*<N>.<PO>

5 Conclusion

Nous avons présenté dans cette communication une méthode permettant la catégorisation des mots d'une langue, dans ce travail le malais. Cette méthode consiste tout d'abord à définir un jeu d'étiquettes morphosyntaxiques très grossier et établir une liste des mots dont la catégorie n'est pas ambiguë à partir de ces catégories prédéfinies. Puis, les mots affixés et à clitique sont étiquetés en utilisant un analyseur morphologique. Les mots restants sont étiquetés par l'application de quelques règles contextuelles. A la fin du traitement d'un texte, tous les mots étiquetés sont ajoutés à la liste des mots dont la catégorie n'est pas ambiguë. Ces différentes étapes sont répétées jusqu'à ce que tous les mots du corpus aient obtenus une étiquette ou que plus aucune règle ne soit applicable.

Les premiers résultats de nos travaux sont la création d'un corpus partiellement annoté avec des catégories morphosyntaxiques, un jeu d'étiquettes morphosyntaxiques du malais et une catégorisation morphosyntaxique des bases simples et affixées.

L'étape suivante sera de développer la méthode destinée à compléter l'étiquetage de tous les mots en utilisant un corpus partiellement étiqueté.

Remerciements

Je remercie mes deux relecteurs anonymes et Christian Boitet pour leurs commentaires constructifs.

Références

[1] *Kamus Dewan – Edisi Ketiga*, Kuala Lumpur, Dewan Bahasa dan Pustaka. 1994.

[2] *General guidelines for Malay Spelling*, Kuala Lumpur, Dewan Bahasa dan Pustaka. 1992.

[3] ISMAIL BIN DAHAMAN. *Pedoman Ejaan dan Sebutan Bahasa Melayu*, Kuala Lumpur, Dewan Bahasa dan Pustaka. 2000.

[4] ISMAIL BIN DAHAMAN. *Pedoman Ejaan Rumi Bahasa Melayu*, Kuala Lumpur, Dewan Bahasa dan Pustaka. 2000.

[5] *General guidelines for the formation of terms in Malay*, Kuala Lumpur, Dewan Bahasa dan Pustaka. 1992.

[6] NIK SAFIAH KARIM, FARID M. ONN, HASHIM HJ. MUSA, ABDUL HAMID MAHMOOD. *Tatabahasa Dewan. Edisi Baharu*, Kuala Lumpur, Dewan Bahasa dan Pustaka. 2004.

[7] RANAIVO B. *Analyse automatique de l'affixation en malais*, Thèse de doctorat, Institut National des Langues et Civilisations Orientales, Paris. 2001.

Les langues créoles de São Tomé : transcrire pour écrire

Emmanuel Schang

CORAL (UPRES-EA 3850) – Université d'Orléans
Faculté LLSH
BP 46527, 45065 ORLEANS Cedex 2 (France)
emmanuelZschang@univ-orleansZfr

Mots-clés : Créoles portugais, golfe de Guinée, São Tomé et Príncipe, corpus oraux
Keywords: Creole languages, Gulf of Guinea Portuguese-based Creoles, São Tomé and Príncipe, spoken corpora

Résumé

Je présente une initiative visant à mettre à la disposition du public (scientifiques et Organisations Non Gouvernementales) des corpus oraux transcrits sur les créoles portugais du golfe de Guinée ainsi qu'une base de données lexicales des termes utilisés dans ces corpusZ Ce projet vise à contribuer à l'essor de l'écriture en forro et en angolarZ

Abstract

This paper presents an ongoing project for the Gulf of Guinea Portuguese-based Creole languagesZ It aims at delivering transcribed spokencorpora and a lexical databaseZ

1 Introduction

Je présente dans cet article une initiative de sauvegarde et de développement des langues créoles de São Tomé (République Démocratique de São Tomé et Principe) utilisant les nouvelles technologiesZ Encore peu décrites, ces langues sont également assez peu connuesZ Ce projet a pour objectif d'augmenter la documentation sur ces langues par le biais de corpus oraux transcrits et d'une base de données lexicalesZ Il prend place dans une initiative plus vaste, CreolData, visant à créer une base de données lexicales informatisée sur les créoles portugais d'Afrique (Schang & alii, *à paraître*), rassemblant d'autres créolistes et s'appuyant sur le dictionnaire de Jean-Louis Rougé (Rougé 2004)[1]Z

[1] Pour l'instant, ces projets n'ont pas fait l'objet de demandes de financement propres et reposent sur la libre contribution de chercheurs universitairesZ

Les rapports qu'entretiennent ces langues créoles avec le portugais et les langues africaines encore parlées dans certaines communautés seront également abordés car il est nécessaire de situer ces langues dans leur environnement sociolinguistique pour appréhender les difficultés dans le passage à l'écriture et à l'*informatisation* de ces languesZ

2 Les créoles portugais du golfe de Guinée

2.1 Quelques mots d'histoire

Il est impossible de décrire précisément en quelques lignes la situation linguistique d'un pays aussi complexe que São Tomé et PrincipeZ Cependant,on peut tracer quelques grands traits qui permettront au lecteur de se faire une idée de la situationZ

La République Démocratique de São Tomé et Príncipe (capitale : São Tomé) comprend deux îles : São Tomé et PríncipeZ Indépendantes depuis 1975, ces deux îles situées dans le golfe de Guinée couvrent une superficie d'à peine 1000 km^2 au total pour une population d'environ 175000 habitantsZ

L'île de São Tomé est, selon toute vraisemblance, inhabitée lorsque les navigateurs portugais la découvrent en 1471Z Elle ne sera peuplée de façon définitive qu'à partir de 1482, par des portugais relégués (*degredados*), des esclaves venant du royaume du Bénin puis de la région Congo-Angola, un millier d'enfants juifs convertis de force en 1492 et une poignée d'aventuriers commerçants venus de toute l'Europe (vZ Caldeira 1999)Z

Au succès de l'époque de la canne à sucre (16ème siècle) succèdera une période troublée de déclin économique jusqu'au 19ème siècle, qui verra les beaux jours des plantations de cacao et de café et l'apport d'une main d'oeuvre contractuelle (parfois forcée) originaire principalement du Cap Vert, de l'Angola et du MozambiqueZ

2.2 La situation actuelle

De ce passé, l'île hérite d'une situation sociolinguistique complexeZ En effet, à côté du portugais qui est la langue officielle de l'île on trouve deux langues créoles distinctes : le *forro* (ou lungwa santomé) et l'*angolar* (ou ngola)Z Le forro est le créole majoritaire parlé à la fois dans la capitale (São Tomé) et dans l'ensemble de l'île (ainsi que sur l'île de Principe, à côté du créole local, le lung'ie)Z L'angolar est le créoleparlé actuellement par les pêcheurs de l'îleZ Mais historiquement, il s'agit de la langue des esclaves fugitifs réfugiés dans les montagnes de l'île qui seront à l'origine de révoltes importantesZ Les Angolares, chassés des terres, seront contraints au 19ème siècle à devenir pêcheurs et à vivre sur les plages de l'îleZ

A côté de ces langues créoles locales, on entend toujours parler le créole capverdien (au travers de ses différents dialectes) mais aussi ce qu'on appelle les langues des Tongas (Rougé 1992) qui sont des vestiges de langues bantoues du Mozambique et d'Angola, et un portugais local (portugais des Tongas) parlé dans les plantationsZ

Dans ces conditions, le portugais, qui est la seule langue parmi celles citées précédemment à être véritablement enseignée à l'école, constitue la langue des élites et de l'émigration (puisqu'il existe une importante diaspora santoméenne à Lisbonne)Z

Mais dans un contexte d'indépendance, le forro incarne à la fois le sentiment national (la rue reprochera volontiers lors des élections à tel ou tel homme politique de ne pas parler le forro) et la tradition culturelleZ

3 Développement linguistique

Comme le soulignait JZ-LZ Rougé lors du Premier Colloque international sur les Langues Nationales qui s'est tenu à São Tomé en octobre 2001 (JZ-LZ Rougé 2001), il existe une différence importante entre écrire en créole et transcrire le créoleZ Si quelques tentatives de transcription ont vu le jour pour le forro et l'angolar, on peut dire qu'il n'existe pas de littérature écrite dans les créoles de São Tomé, pas plus d'ailleurs que d'ouvrages techniques ou scolairesZ L'usage du créole transcrit reste le fait de quelques initiatives personnelles (généralement de la poésie comme Paga Ngunu, etcZ)ZLa Radio Nationale et quelques Organisations non Gouvernementales ont fait parfois usage de slogans écrits en forro, sans toutefois chercher l'adoption de conventions orthographiques stablesZ Des initiatives liées à la Direction de la Culture tentent cependant de lancer un cours de créole à l'Ecole Polytechnique de São Tomé, sans grand impact pour l'instantZ

On peut dire sans déformer la réalité que le portugais est la seule langue écrite de São Tomé à l'heure actuelleZ

Faute de standardisation de la graphie, des formes concurrentes voient le jour pour noter certains phonèmes (faut-il noter [k] par *k* ou *c* dans [fika] "rester/être" ?)Z Bien que ceci soit un frein pour le développement de l'écriture en langue créole, il n'y a pas là un problème insoluble ou propre aux créoles (qu'on songe à des langues comme le luxembourgeois par exemple qui font face aux mêmes problèmes)Z Des solutions techniques existent qui ont été développées pour d'autres langues créoles (CreoleConvert http://hometownZaolZcom/mit2haiti/CrConvZhtm par exemple pour le créole haïtien)Z

Il existe bien sur Internet un groupe de discussion sur São Tomé qui cherche à promouvoir la culture santoméenne et le forro, mais l'impact de cette initiative retentit plus sur la diaspora que sur la population insulaire mêmeZ Comme ailleurs, l'utilisation du créole sur la Toile, par le biais des messages, s'apparente plus à du folklore (sympathique au demeurant) qu'à une volonté concrète de développement linguistiqueZ

Les corpus oraux que nous avons réalisés[2] sur le terrain montrent bien qu'en situation de conversation spontanée, le vocabulaire utilisé par d'authentiques créolophones est largement différent de celui utilisé par les internautesZ L'érudition et l'originalité sont revendiquées par les internautesZ L'exclusion de toute forme sensée être proche du portugais (perçu par certains comme l'adversaire linguistique du créole) est la règleZ On s'interrogera plus loin sur la pertinence de cette attitudeZ Ceci se retrouve par ailleurs dans toutes les mises en écriture des langues minoritaires (les autres langues créoles n'échappent pas à ce problème)Z

[2] Rapport de mission à São Tomé, mai 2004, J-L Rougé et EZ Schang (CORAL)Z

4 Genèse et philosophie du projet

Quelques explications sur la genèse de ce projet éclaireront probablement le lecteur sur la philosophie qui nous guide iciZ

C'est au cours d'un travail à São Tomé en 1997 aux côtés des ONG dans des projets de microcrédit (Caixa de Poupança) qu'est née l'idée de favoriser l'émergence auprès des populations peu scolarisées (ici les femmes des pêcheurs angolares des plages du sud de l'île) de l'écriture en créoleZ Le projet de microcrédit n'a pas continué[3] mais l'idée est restéeZ

Le point de départ consistait à recueillir des enregistrements sur différents sujets de la vie quotidienneZ Ceux-ci devaient servir de base pour l'élaboration de documents pédagogiques et technique en créole, en réutilisant les termes employés par les informateurs lors des entretiensZ Il s'agissait de prendre le contre-pied d'une pratique courante consistant à chercher à traduire en créole des termes portugaisZ L'idée consistait donc à ne pas se servir de concepts exogènes (souvent issus de documents techniques portugais) mais d'employer les termes de la vie quotidienne autant que possibleZ Si un organisme voulait à relancer des projets de ce type, la base d'entretiens annotés que nous proposons lui serait disponible, évitant le coût d'une étude préliminaireZ

Mais à l'heure actuelle[4], les travaux visent essentiellement la communauté des linguistes en cherchant à sauvegarder des traces (orales) de ces langues créoles (le lung'Ie de Príncipe étant quasiment éteint, les autres créoles risquent de suivre cette voie)Z

Le projet est pionnier pour ce groupe de langues, mais modeste tant par les moyens mis en œuvre que par l'ambition avouéeZ Il s'agit pour l'essentiel de constituer des enregistrements transcrits et annotésZ Ceux-ci seront accessibles rapidement à la communauté scientifique et à qui le souhaitera pour une utilisation compatible avec la philosophie de ce projetZ

Ce projet concernant les créoles de São Tomé s'inscrit dans une initiative plus large visant à constituer une base de données lexicales sur les créoles portugais d'Afrique, CreolData (vZ Schang & alii *à paraître* et note 1)Z

Bien entendu, une telle approche va de paire avec le parti pris d'adopter autant que possible les logiciels libres et gratuitsZ

5 Démarche suivie

La démarche consiste à recueillir des enregistrements des créoles de São Tomé lors de missions de recherche sur le terrainZ Les entretiens enregistrés portent aussi bien sur les contes et la tradition orale que sur la vie quotidienneZ Les informateurs enregistrés sont des locuteurs de langue maternelle créole et utilisent le créole quotidiennementZ Les enregistrements récents sont faits directement au format numérique (fichiers Zwav, 16-bit, 44100hz)Z

---

[3] Pour des raisons qui n'ont pas de rapport avec la linguistiqueZ

[4] En raison d'un contexte politique peu favorable, pour de nombreuses raisons, tant nationales qu'internationales, qui ne permet pas de placer l'éducation en créole parmi les prioritésZ

Les entretiens sont transcrits ensuite par le ou les linguistes sous le contrôle d'un informateur à l'aide du logiciel Transcriber (Barras & alii 2001)Z

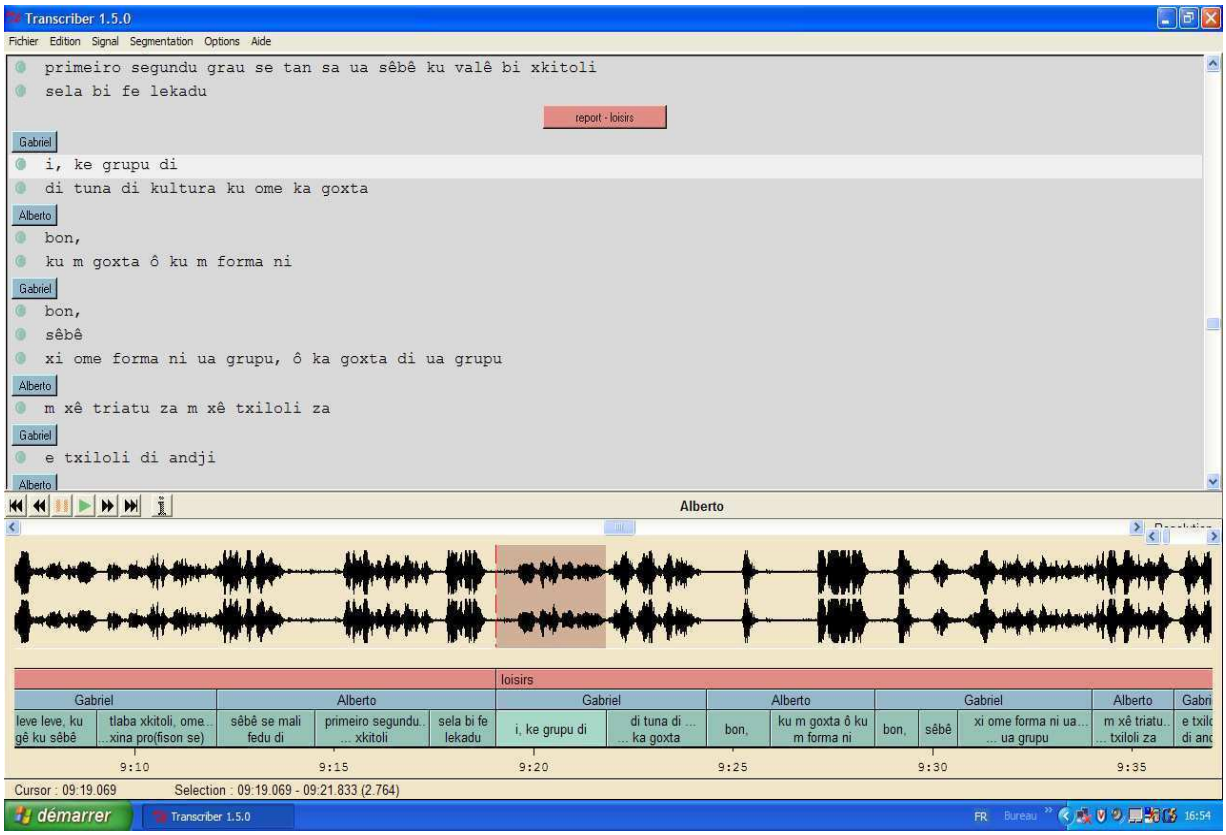


Figure 1 : transcription des entretiens avec Transcriber.

La graphie choisie pour transcrire les entretiens est celle qui est utilisée par JZ-LZ Rougé dans son dictionnaire (Rougé 2004)Z Aucune graphie officielle n'étant adoptée à l'heure actuelle à São Tomé (même si des propositions dans ce sens ont été faites au Colloque International sur les Langues Nationales de São Tomé et Príncipe), il paraît intéressant d'opter pour une graphie qui soit compatible avec les différents créoles portugais d'Afrique car certains termes sont communs aux différents créoles (par exemple *kume* "manger")Z Nous reviendrons plus tard sur ce pointZ

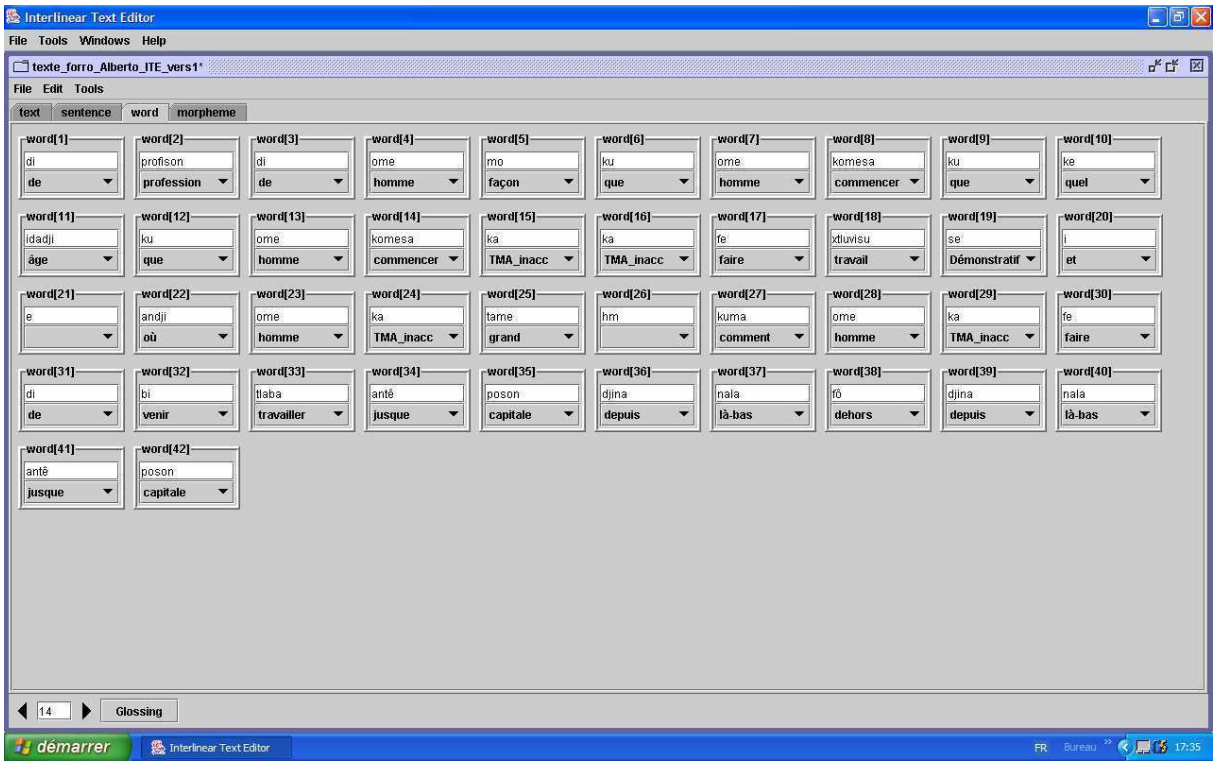
Transcriber permet par la suite de travailler sur des fichiers texte ou des fichiers XML contenant la transcriptionZ

```
<Section type="report" topic="to2" startTime="559.069" endTime="1040.730">
<Turn speaker="spk1" startTime="559.069" endTime="564.326">
<Sync time="559.069"/>
i, ke grupu di
<Sync time="561.833"/>
di tuna di kultura ku ome ka goxta
</Turn>
<Turn speaker="spk2" startTime="564.326" endTime="568.552">
<Sync time="564.326"/>
bon,
<Sync time="565.938"/>
ku m goxta ô ku m forma ni
</Turn>
<Turn speaker="spk1" startTime="568.552" endTime="574.0">
<Sync time="568.552"/>
bon,
<Sync time="569.644"/>
```

```
sêbê
<Sync time="570.635"/>
xi ome forma ni ua grupu, ô ka goxta di ua grupu
</Turn>
<Turn speaker="spk2" startTime="574.0" endTime="576.053">
<Sync time="574.0"/>
m xê triatu za m xê txiloli za
</Turn>
<Turn speaker="spk1" startTime="576.053" endTime="577.425">
<Sync time="576.053"/>
e txiloli di andji
</Turn>
<Turn speaker="spk2" startTime="577.425" endTime="578.897">
<Sync time="577.425"/>
txiloli Santana
</Turn>
<Turn speaker="spk1" startTime="578.897" endTime="581.28">
<Sync time="578.897"/>
sa ome sa blabu ni kwa se
</Turn>
```

Figure 2 : segment du fichier XML contenant la transcription de l'entretien

Ensuite, la transcription sous format texte sert d'entrée à l'élaboration d'une glose à l'aide de Interlinear Texte Editor (désormais ITE; Lowe & alii 2004 et http://michelZjacobsonZfreeZfr/ Les mots sont alors glosés[5] en français (figZ 3) et une proposition de traduction du texte est élaboréeZ ITE permet la réalisation d'une nomenclature (figZ 4) des termes utilisés dans les entretiensZ Il comprend également un concordancier très utile au linguisteZ Son principal avantage dans le cadre de ce projet est qu'il utilise XML et nous permet alors, grâce aux feuilles de style XLST, de pouvoir mettre en forme les données selon nos impératifs du momentZ



[5] La glose étant entendue comme une traduction littérale du mot créole en françaisZ

Figure 3 : la glose mot par mot des entretiens avec ITE.

Cet outil apparaît aussi utile pour les travaux de recherche en linguistique que pour la mise en forme des données recueilliesZ S'il est conçu pourpermettre une transcription juxtalinéaire aisée (vZ le programme Archivage du LACITO : http://lacitoZvjfZcnrsZfr/archivage/), l'utilisation qui en est faite dans ce projet est autreZ En effet, le lexique récupéré dans les entretiens (figZ 4) sert d'entrée à la base de données lexicales que nous élaboronsZ Les items lexicaux sont repris à partir du fichier XML contenant le lexique et constituent le point de départ d'une entrée lexicale de CreolDataZ

```
<?xml version="1.0" encoding="UTF-8"?>
<lexique>
  <item nb="2">
    <transcription>mendu</transcription>
    <glose>peur</glose>
  </item>
  <item nb="1">
    <transcription>matu</transcription>
    <glose>forêt</glose>
  </item>
  <item nb="2">
    <transcription>mindjan</transcription>
    <glose>médicament</glose>
  </item>
```

Figure 4 : extrait du lexique obtenu avec ITE.

La glose est alors abandonnée au profit d'une définition (s'inspirant ou provenant pour l'essentiel de Rougé 2004)Z

Les données sont alors présentées sous une forme s'inspirant largement des propositions de norme en cours (Lexique pour le TAL et Lexical Markup Framework[6])Z

---

[6] Documents et discussions disponibles sur wwwZnormalangueZorg et wwwZtc37sc4Zorg

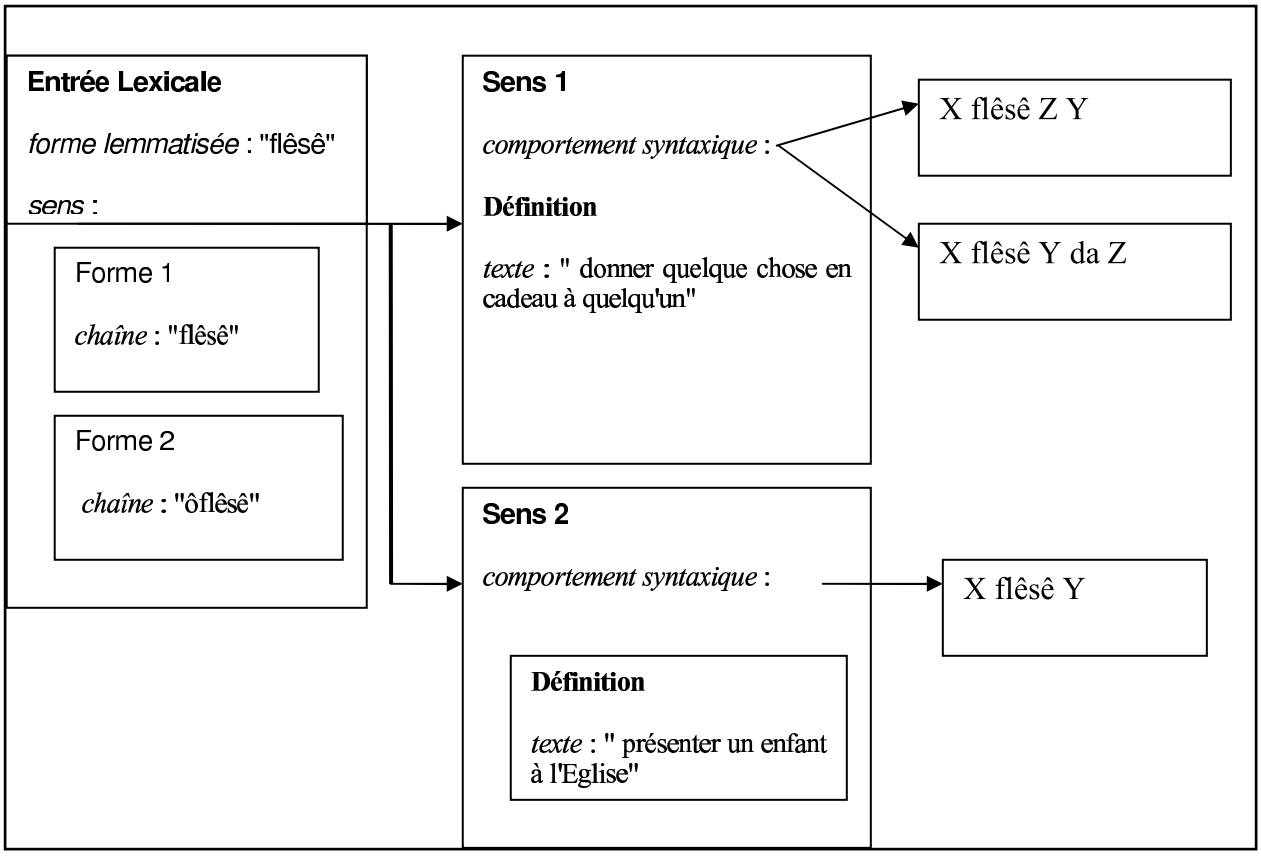


Figure 5 : schéma représentant l'entrée lexicale de flêsê "offrir".

La figure 5 illustre de façon schématique le traitement du mot *flêsê* "offrir"Z Il existe en forro deux variantes (au moins) *flêsê* et *oflêsê* qui ont deux sens distincts : 'présenter un enfant à l'Eglise' et 'offrir'Z Ce dernier sens autorise deux comportements syntaxiques (<X flêsê Z Y> et < X flêsê Y da Z>)Z

Pour l'instant, il n'y a pas de véritable traitement de l'interface syntaxe-sémantique car celle-ci pose un grand nombre de problèmes tant pratiques que théoriquesZ Faut-il utiliser la théorie classique des rôles thématiques (comme proposé dans Schang & alii *à paraître*) ou bien opter pour une autre théorie ? A l'heure actuelle, nous ne disposons pas d'éléments de réponse satisfaisantsZ D'un côté, la théorie des rôles thématiques pose problème : il n'est pas plus simple en forro qu'en français de savoir quels rôles attribuer pour les verbes *lire* ou *persuader*Z De l'autre, d'autres modèles plus complets existent mais ils sont lourds à mettre en œuvre (que ce soit la Théorie Sens-Texte ou le Lexique Génératif, vZ Bouillon & Busa 2001 pour une discussion)Z La morphologie (peu importante dans ces créoles en l'absence de flexion nominale et verbale ainsi que de procédés dérivationnels propres à la morphologie, vZ Schang (2000)) est également absente et fera l'objet de travaux à venirZ Le lexique étant issu des enregistrements recueillis, il se compose des atomes reconnus par la syntaxeZ Ce n'est probablement pas satisfaisant dans l'absolu, mais dans une première approche, il s'agit d'un choix de raisonZ

Quoi qu'il en soit, ces problèmes n'ont rien de spécifique aux créoles portugais et se posent à l'identique pour toutes les languesZ La différencetient ici essentiellement dans les moyens à disposition pour le traitement de ces problèmesZ Eneffet, le choix de coller au plus près à la proposition de norme Lexical Markup Framework relève ici plus de l'exercice de style que de

la nécessité économique : qui développera des logiciels de traduction pour une langue de moins de deux cent mille locuteurs ? C'est la volonté de partager les données qui prévaut ici et non pas les intérêts économiquesZ

6 Problèmes et développements futurs

Nous avons déjà vu dans les paragraphes précédents quelques problèmes liés à l'élaboration du projetZ J'insisterai ici sur la question de la variation (et de sa productivité), qui se pose dès lors qu'on cherche à décrire le créole ou à proposer des transcriptions du créoleZ

Certes, celle-ci n'est pas un phénomène propre aux langues créoles mais les créoles posent depuis toujours la question de l'identification des langues (vZ notamment Schang 2004 pour une discussion)Z Je ne parle pas ici de la coexistence de variantes d'un même mot (vZ figure 5 pour une solution à ce problème) mais de la difficulté de cerner les limites de ce qu'on appelle le créoleZ

Le problème que l'on rencontre tient à l'imbrication des systèmesZ Les locuteurs passent d'un système (devrais-je dire d'une langue ?) à l'autre, tant au plan lexical qu'au plan grammaticalZ C'est ce qu'illustre (1) où figure en italique un mot portugais fléchi au milieu d'une phrase en forroZ

(1) non na xka te *condições* (…)

/nous/négZ/Temps-Mode-Aspect/avoir/conditions/

nous n'avons pas la vie facile (...)

De façon générale, il est absolument arbitraire d'exclure les mots portugais du forro et de chercher à créer un créole "pur" sachant que les mots d'origine portugaise constituent près de 95 % du lexique du forroZ

Par ailleurs, s'il est aisé de repérer ce qui est typiquement angolar (défini en creux comme ce qui n'est ni forro ni portugais), on peut remarquer que l'angolar et le forro sont très proches, tant par leur vocabulaire que par leurs structures grammaticalesZ La vision organique des langues en tant qu'entités distinctes qui vivent et meurent est dans ce contexte mise à malZ Dans des enregistrements spontanés, il est très souvent difficile d'estimer s'il s'agit de portugais (avec un fort accent local) contenant des mots créoles ou bien d'une variété acrolectale (proche du portugais) du créoleZ Il enva de même entre les deux créoles forro et angolaresZ

La conception que je défends consiste à dire que les mots n'appartiennent pas à une langue mais que celle-ci les utiliseZ En effet, dire, parexemple, que le mot *taxi* est un mot français, portugais ou forro n'a guère de sensZ On trouvera dans les trois langues quelque chose qui, à la variante de prononciation près, est le mot *taxi*Z Doit-on noter qu'il s'agit d'un mot d'origine grecque ? Pourquoi pas, tant qu'il s'agit de dire qu'il s'agit d'une origine grecque et non pas d'un mot grecZ Ainsi, je prends le parti de ne pasexclure les mots d'origine portugaise des transcriptions et du lexique que je mets en placeZMais s'agit-il alors d'un lexique forro ou d'un lexique correspondant à un/des locuteurs(s) dans une situation de communication donnée ? C'est assurément la seconde réponse qui est la bonneZ Dans ce cas, je ne peux que proposer la description de situations de communication plutôt que la description d'une langueZ De ce point

de vue, on ne peut qu'être d'accord avec la devise de linguasphère (http://wwwZlinguasphereZorgZ): "dans la galaxie des langues, la voix de chaque personne est une étoile"Z

Remerciements

Je remercie mon collègue Jean-Louis Rougé pour sa collaboration à ce travail, ainsi que Laurent Romary et Gil Francopoulo pour leurs conseils ponctuelsZ

Références

BARRAS, CZ,GEOFFROIS, ÉZ,WU, ZZ,LIBERMAN, MZ(2001), Transcriber: development and use of a tool for assisting speech corpora productionZ*Speech Communication*, *33*(1-2): 5–22Z

BOUILLON, PZ,BUSA, FZ (2001)Z*The language of word meaning*Z Studies in NLP, Cambridge University PressZ

CALDEIRA, AZ (1999),*Mulheres, sexualidade e casamento em São Tomé e Príncipe.* Lisboa: CosmosZ

LOWE, JZ,JACOBSON, MZ,MICHAILOVSKY, BZ (2004), Interlinear Text Editor Demonstration and Projet Archivage Progress ReportZ Actes de :*4th E-MELD (Electronic Metastructure for Endangered Languages Data) workshop on language engineering: Linguistic Databases and Best Practice. Detroit. 15-18 juillet 2004*Z

ROUGE, JZ-LZ(1992), Les langues des TongaZ Actes de :*Colóquio sobre 'Crioulos de base lexical portuguesa', Universidade de Lisboa, junho de 1991, E. d'Andrade e A. Kihm (eds)*Z

ROUGÉ, JZ-LZ(2001), Escrever i/o transcrever as línguas de São Tomé e PríncipeZ Actes de : *Collóquio Internacional sobre as línguas nacionais. São Tomé, oct. 2001.*

ROUGE, JZ-LZ (2004),*Dictionnaire étymologique des créoles portugais d'Afrique*Z Paris : KarthalaZ

SCHANG, EZ (2000), L'émergence des créoles portugais du golfe de GuinéeZ Presses Universitaires du SeptentrionZ Thèse de Doctorat del'Université Nancy 2Z

SCHANG, EZ (2004), Identification automatique des langues :que faire des créoles ? Actes de : *Workshop MIDL 2004. Paris, nov. 2004. Presses de l'ENST.*

SCHANG, EZ,ROUGE, JLZ,ESHKOL, IZ,PETIT, MZ (à paraître en 2005), CreolData : une base de données lexicales informatisée sur les créoles portugaisZ*Les créoles, Revue Française de Linguistique Appliquée*, DZ Fattier (edZ)Z

Methods, Models and Standardization Issues for the Creation of Linguistic Resources: the Case of Under-Represented Languages

Claudia Soria, Monica Monachini

ILC-CNR
Via Moruzzi 1, Pisa, Italy
{claudia.soria, monica.monachini}@ilc.cnr.it

Keywords: less widely available languages, multilingual terminological resources, resource production methodology, and standards.

Abstract Availability of Linguistic Resources for the development of Human Language Technology applications is nowadays recognized as a critical issue with both political and economic impact and implications on the sphere of cultural identity. Many languages have little or no information technology available. This paper reports about the experience gained during the *INTERA* European project for the production of multilingual terminological lexicons for those languages that suffer from poor representation over the net and from scarce computational resources, but yet are requested by the market. It discusses the procedure followed within the project, focuses on the problems faced which had an impact on the initial goals, presents the necessary modifications that resulted from these problems, evaluates the market needs as attested by various surveys, and describes the methodology that is proposed for the efficient production of Multilingual Terminological Lexicons.

1 Introduction

Language Resources are central components for the development of Human Language Technology applications. The availability of adequate Linguistic Resources for as many languages as possible plays a critical role in view of the development of a truly multilingual information society. It is no mystery that the task of producing language resources is an extremely long and expensive process. For most western languages, and English in particular, however, this is partly softened by large and often free data availability, good representativeness, and significant size, together with availability of language processing tools. There is plenty of languages, however, for which this picture is far from being adequate. Many languages actually suffer from poor representation and scarcity of raw material, not to mention the availability of robust processing tools. It is imperative to try to reach a balance of language coverage in order to avoid a *two-speed Europe* (Maegaard *et al.* 2003). This awareness gave rise to coordinated efforts, both at national and European level, in the direction of reducing this gap (Calzolari *et al.* 2004) and enabling less-favoured languages with respect to language technology. The concepts of BLARK, i.e. the definition and adoption of a standard Basic Language Resource Kit for all languages, a minimal set of language

resources necessary to the development of language and speech technology (Krauwer 1998), goes exactly in this direction.

In *INTERA*, the expression "less widely available languages" has to be interpreted in the sense of (Gavrilidou *et al.* 2003, Gavrilidou *et al.* 2004), that is, of "less widely available in the digital world". This concept has been developed in response to a survey, also conducted in the framework of the *INTERA* Project, aimed at the identification of users' needs and expectations concerning language resources. Although western European languages have been confirmed as having the highest amount of request, the survey clearly demonstrates that there is an increase in demand for Balkan and eastern European languages. Under this respect, it has to be noted that the notion of "less widely available languages" is by no means to be interpreted as a synonym to "less widely spoken languages"; in fact, many of the languages to which the former concept seems to apply, such as eastern European languages and Balkan ones, actually appear among the forty most widely spoken languages all over the world (http://www.globallanguages.com). For instance, Russian, Polish, Ukrainian, Romanian, and Serbo-Croatian appear in the 8th, 22nd, 23rd, 32nd, 37th rank respectively. Nevertheless, these languages still seem to be rather under-represented as regards digital content, although there is an increase in Balkan LRs production - tendency encouraged by national and EU activities as well - as attested by surveys conducted in the framework of other European projects (i.e. ENABLER, www.enabler-network.org). Despite their limited availability, Balkan and eastern European languages are highly requested by the digital content market, thus making the issue of resource production even more crucial.

The *INTERA* European Project had a twofold task: on the one hand it was aimed at producing multilingual parallel corpora and terminological lexica for some languages that were identified as belonging to the class of "less available ones", i.e. Greek, Serbian, Bulgarian and Slovene. On the other hand, another aim of the project was establishing a reference production model for multilingual resources, which is up-to-date, compliant with existing standards and yet viable and attractive for digital content producers. In this paper we report about the experience gained during the *INTERA* Project for the production of a model for multilingual terminological lexicons.

The particular point of view of trying to design a production model for multilingual terminological lexicons inevitably brought us to try to conciliate two different, opposite forces. On the one hand, there are the needs and requirements expressed by users of the digital content market, as emerged through a thorough examination of eContent professionals' practices and policies as for LRs. From this point of view, any prospective resource should aim at completely satisfying user needs and requirements, as well as complying with existing standards. On the other hand, there is evidence of unavoidable production gaps and the users' documented desiderata are in conflict with the actual viability and feasibility of language resources, as is documented by various surveys of language resources (Gavrilidou and Desipri 2003; Gavrilidou *et al.* 2003).

This means that a realistic production model should also take into consideration very basic problems of data availability, representativeness, and size, together with availability of language processing tools. A realistic model for the production of multilingual terminological lexicon, thus, rather than describing an ideal situation, which would be far from reality, should consider a variety of possible situations thus anticipating possible shortcomings in all stages of production.

The paper is organized as follows: Section 2 describes user needs; Section 3 illustrates possible scenarios to be faced as for data availability, format, and annotation, whereas Section 4 presents the actual scenario we were confronted with. Section 5 presents the methodologies and techniques used for the construction of the *INTERA* multilingual terminological resource. Finally, Section 6 presents the conclusions.

2 User Needs

According to the requirements emerged from *INTERA*'s user needs survey and from the outcomes of relevant past initiatives, a multilingual terminological lexicon should fulfil both resource-related and content-related requirements. Resource-related requirements are those referring to overall characteristics such as use of common, widely accepted and widely used standards, availability of readily accessible information, and validation. Content-related requirements, on the other hand, refer to the availability of particular types of information in the terminological entries, which users assess as essential, desirable, or irrelevant. Equivalents in other languages and grammatical information about a given terminological entry are considered as essential information, while definitions, conceptual relations, a domain code indication, and reliability codes are seen as desirable.

3 Foreseen Scenarios

The quality of the final resources produced is strictly intertwined with the existing source material, its (re)usability, the presence of linguistic annotation or the availability of tools to perform linguistic analysis, the compatibility/reusability of different linguistic analyses, etc. Data availability, representativeness, and size, together with availability of language processing tools are crucial factors to be taken into account in a very realistic production model, since they are variables with strong repercussions on the task and the process of corpus production and terminology extraction. Different possible scenarios can thus be foreseen for the production of multilingual resources, all of them having an impact over the task of term extraction. The quality of the multilingual resources dramatically differs depending on the scenario, i.e. according to whether a corpus is *parallel* or *comparable*, whether there is a unique pivot language available or not, etc. The possible scenarios envisaged can be summarized as follows:

In the "ideal" scenario, the extraction task can be performed working on parallel specialized texts with a pivot language (hopefully English) for which NLP tools and resources are available. It can be expected that sufficiently "clean" lists of candidate terms can be extracted that are to be confirmed by the terminologists. In this situation, we can aim at a *truly* multilingual database, where the lexicon will be the same across languages and terms will be all interconnected and corresponding to each other.

In the "worst" solution, where there are no truly parallel texts but sets of parallel texts without any pivot language, we only can resort on statistical procedures of term recognition. The

greatest risk is to produce a list of candidate terms with possible *noise* or *silence*[1] and where human involvement is massive.

In a "mid-way" solution where sets of pairs of corpora are available, the resulting terminological lexicon is not a truly multilingual homogeneous one, but a set of terminological lexicons in different languages where terms are likely not to be the same throughout the lexicons.

4 Deviation from initial goals: The Actual Scenario

The actual configuration of parallel corpora found in *INTERA* was actually different from expected. The scenario we were confronted with was rather similar to the above mid-way solution: the final collection is represented by a *comparable multilingual corpus* as opposed to a parallel multilingual corpus. Instead of having the same texts in all languages, there were pairs of different texts loosely belonging to the same domains (namely, *law, tourism, health, and education*). For each pair, English always represents one member (the *pivot* language), while the other is Greek, Bulgarian, Slovene, or Serbian.

The Table below illustrates this situation in detail.

Domain	**Languages**			
	Greek	**Bulgarian**	**Serbian**	**Slovene**
Law	x	x	x	x
Health	x		x	
Education	x	x	x	
Tourism	x			
Environment	x			
Finance			x	
Politics		x		

Figure 1 : Distribution of languages and domains

The first consideration to be made concerns the *degree of multilingualism* of the terminology: since the texts are not truly parallel, the final terminology is not a truly multilingual one. In other words, the lexicon is not the same across languages and terms are not all interconnected and corresponding to each other. Instead, for each English–*X* pair we derived the

[1] Noise and silence are commonly used in terminology as complementary of precision and recall respectively.

corresponding terminology, thus arriving at a bilingual (English-language–*X*) terminology for each domain. Since the domains are at least partially overlapping, some terms occurring in one terminology also occur in another one, thus enabling to build truly multilingual links at least for a subset of terms. Given the situation illustrated in the Table above, the only domain for which a quadri-lingual partial terminology is feasible is the Law domain. The Education domain yields a tri-lingual terminology, and the health domain a bi-lingual one. The Tourism, Environment, Finance, and Politics domains are monolingual terminologies.

The second consideration is related to the range of technical solutions adopted for automatic term extraction. The availability of the same pivot-language for all target languages proved useful, especially because the target languages are under-represented ones, for which few reference corpora and NLP tools are available. On the contrary, there is a huge amount of resources (corpora, lexica and tools) available for the English language, and this allowed us to opt for a combination of statistical and NLP procedures, as illustrated in more detail in the next Section.

5 Terminology Extraction

Terminology can be considered the surface realization of relevant domain concepts (Cabré, 1992; Sager, 1990). Candidate terminological expressions are identified either by hand, or in a semi-automatic manner. Semi-automatic procedures for terminology extraction usually consist in shallow techniques that range from stochastic methods to more sophisticated syntactic approaches (Jacquemin, 2001; Bourigault, Jacquemin, L'Homme, 2001).

All of them, however, converge in identifying terms mostly on statistical grounds, on the basis of its relative frequency in a corpus, possibly augmenting this measures with filters capturing the domain specificity of a term. Although not theoretically correct (as the status of "termhood" is in principle independent on the number of occurrences, and a *hapax* might well be a term), this practice is rooted in computerized terminology, where computer-aided text analysis and the possibility of processing large amount of information have changed the bases of terminology compilation, as well as how the appropriateness of terms is conceived and the degree of human intervention in the whole process. In this particular context, we adopted a hybrid approach to terminology extraction from multilingual parallel texts, combining statistical and symbolic techniques.

5.1 The data

As introduced above, the data available for the task of automatic term extraction come under the form of four parallel corpora: English-Greek, English-Serbian, English-Slovene and English-Bulgarian. Each parallel corpus is further organized according to the particular domain to which the texts of the corpora belong: while the English-Slovene corpus covers the law domain only, the English-Greek corpus, for instance, covers as many as five domains, i.e. law, education, health, tourism, and environment.

The size of available data is important for determining the coverage of the terminological resource, since more data mean more terms. However, it is important also for the quality of the terminological resource, as the automatic procedure needs an amount of data reaching a

level of statistical relevance to yield high-quality data. Unfortunately, the available data dramatically differed in size both across the different domains and across the different languages. The biggest data were available for Greek and Serbian (59 and 69 Mb respectively), while Bulgarian and Slovene were represented globally only with 24 and 33 Mb.

The richest domain is represented by law (129 Mb), followed at a distance by education (20 Mb) and health (14 Mb). This difference among domains has an obvious consequence over the overall amount of terms that can be made available as a result of the extraction process. In other words, there will be domain-specific terminologies that will be very different in size and hence term coverage. This situation is clearly depicted by the case of the terminology for the health domain. The corpus data amount to 13 Mb for Greek and 1Mb for Serbian. The highest quantity for Greek allows to extract more candidate English terms, as easily foreseen, but, most importantly, to produce less candidate translators and of better quality: while for Greek the candidate translators/terms ratio is of 1,5, for Serbian it is of 4,1.

	Candidate terms	**Terms**	**Candidate translators**	**Translators**
Serbian	734	488	2012	201
Greek	1710	1052	1580	826

Figure 2: Candidate translators/terms ratio for Greek vs. Serbian

5.2 Extraction procedure

The task of automatic term extraction was organized around three main phases:

1. Automatic extraction of terms from the English components of the parallel corpora. The English language is henceforth defined as the "pivot language";

2. Automatic identification of candidate translators in the target languages;

3. Manual verification of the candidate translators found with the automatic procedure.

5.2.1 Extraction of English candidate terms

The objective of the first step is the identification of terms for a given sub-language; it is assumed that these terms should represent those that most probably are peculiar for a specific sub-domain. Under this assumption, the terms that will be identified will represent the candidate terms for a specialised (domain-specific) lexicon.

Candidate single terms are extracted by comparing the relative frequency of lemmas inside each domain and language specific subcorpus against a lemma-based frequency lexicon of the *British National Corpus*, which was used as a reference corpus.

In more detail, the comparison between the frequency distributions of terms in the general lexicon and that of the different domain-specific lexicons was performed adopting a mathematical function evaluating "the distance of the frequency of domain-specific terms from the frequency which was expected on the basis of the general lexicon".

We compared the lists generated adopting several different mathematical formulae, among which are the following:

$d1 = f_{r\,(specialized\ lexicon)} - f_{r\,(general\ lexicon)}$

$d2 = f_{r\,(specialized\ lexicon)} / f_{r\,(general\ lexicon)}$

$d3 = \log(f_{r\,(specialized\ lexicon)} / f_{r\,(general\ lexicon)})$

where $f_r$represents the relative frequency of a term inside the lexicon.

Terms, however, are not simply represented by single terms. Examples include compounds (*credit card*), adjective-noun (*administrative procedure*) or complex noun phrases (*principle of equal treatment*). We thus specified a bunch of basic syntactic rules expressing constraints over syntactic patterns in order to select candidate multi-word terms.

In order to avoid over-generation problems, some corrective measures have been applied, most notably by specifying either lists of words to be discarded *a priori* (stop-word lists) or different values of the threshold under which a candidate is automatically rejected. The threshold is each time adjusted depending on the overall size of the parallel corpora under analysis and empirical measures.

5.2.2 Extraction of candidate translators

Once candidate terms are identified for English, we turn to the task of automatic identification of candidate translators in the target languages. To this end we exploited the structuring information available in the parallel corpora from which the terminology was to be extracted. Since the sentences in the target language texts are aligned to those of the pivot language, it is easy to select a suitable search space for any candidate term. The algorithm for the extraction of candidate translators consists of the following steps:

1. Selection of the *source region set* from the pivot language corpus;
2. Extraction of *target region set* from the target language corpus;
3. Search Extraction of lemmas from target region set;
4. Ordering of the lemmas contained in the search target region set according to a *ranking function*;
5. Selection of candidates.

Given a candidate term *t* (in English), the target region set inside the target language corpus is easily identified thanks to the parallel structure of files to be processed: each region of the

English corpus containing the term *t* is uniquely associated with a region of the target language corpus.

Then, the lemmas from the target region set are extracted, filtering out lemmas belonging to "non significant grammatical categories" (e.g. conjunctions, prepositions).

It was observed that the target language lemmas could be classified on the basis of their "probability" of being a translation of a given term by means of simple frequency analyses.

This classification is obtained through the synthesis of a ranking function. Several hypotheses were considered, all of them aiming at highlighting the statistical "idiosyncrasies" of the translating lemma.

The best performing measure is the following:

$f(l) = r(l) - q(l) * |I|$

Where $r(l)$ is the number of regions of the target region set containing at least one occurrence of lemma l, $q(l)$ is the ratio between the number of regions containing lemma l and the total number of regions in the corpus; $|I|$ is the total number of regions of the target region set.

5.2.3 Validation and production of terminological entries

The lists of candidate English terms and their corresponding candidate translations in the other languages (lists of single word terms and multiword terms as produced by the tool) were presented to human validators, who were all native speakers of the selected languages. The validators' task consisted in examining the lists and marking bilingual pairs of terms (English – their mother tongue) as correct, based on the following criteria:

(a) the pair is indeed a term of the specific domain (and not a general vocabulary word) AND

(b) the translational equivalence is also correct.

The terms identified as correct pairs were further lemmatized (single word terms and multiword terms), and the final lists produced by the validators were suited for the production of the multilingual terminological entries.

Since compliance to a reliable standard framework is a pre-requisite for ensuring sharing, reusability and exchangeability of data, the TMF family of formats was taken as the reference model to encode the *INTERA* terminological entries. TMF stands for Terminological Mark-up Framework (ISO 16642 2001), an international standard designed in the framework of the ISO initiatives to support the creation and use of computer applications for terminological data and exchange of such data between different applications. Being a meta-model for terminology mark-up, TMF allows for the specification of user-defined mark-up languages (called TMLs). A TML makes it possible to design the encoding format of a terminological collection according to specific needs. In designing the *INTERA* TML we tried to harmonize users' needs with the realistic considerations when dealing with under-represented languages.

6 Conclusions

The approach to multilingual terminology lexicons adopted and described in this paper cannot be seen as a standard practice nor is it to be considered as a recommended practice in terminology building. In fact, there is no such practice for all purposes. There only are better solutions under certain conditions. We claim, however, that the procedure adopted is a viable and fruitful one given the following conditions:

- Data are sparse
- No NLP tools are available for the target languages
- No reference corpora are available for the target languages but many NLP tools and reference corpora are available for one language (the *pivot*)

Moreover, this experience taught us some lessons about the more general task of building terminological resources for languages suffering from scarcity of widely available data and processing tools. Thus, besides the actual production of the resources, a parallel result has been the identification of gaps and shortcomings in the process usually employed by LRs producers (or users who might wish to create their own LRs) and to suggest ways of remedying them.

At a general level, the production methodology is heavily influenced by the following factors:

- Lack of integration among computer tools working at different levels of analysis.
- Lack of compatibility among the resources themselves. This means, for instance, not only enforcing compatibility in data encoding and representation, but also ensuring that the resources are compatible from the point of view of the additional, linguistic and non-linguistic information, which is added to the raw data. Once again, compliance with agreed-upon standards is recommended, as well as harmonization among the different tag sets used in the various resources. Ideally, all resources should use the same convention of linguistic annotation; when this is not possible, it is recommended that a harmonized tag set is used, or that conversion procedures from the proprietary tag set to a common, standardized one is provided.
- The limited number of existing corpora, especially in languages other than English.
- The particular configuration of resources available. The particular methodology to be adopted for the production of multilingual terminological resources must be carefully adjusted to the idiosyncratic situation to be handled, where by situation we mean the type of languages, the quantity and quality of resources, and the purposes for which the resource is being built.

In conclusion, the experience gained during the *INTERA* project calls for more resources, of good quality, and compliant with sound standards. Lesser-favoured languages can benefit from building parallel resources where English represents one language. Another general recommendation is that a criterion of *practical feasibility* be followed, thus balancing the constraints imposed by corpus size, languages, and users' and standards

requirements. This is deemed the only viable and reasonable solution, especially from the point of view of prospective users that will have to apply the model that is the outcome of the *INTERA* project.

References

BOURIGAULT D., JACQUEMIN C., L'HOMME M.-C. (EDS) (2001), *Recent Advances in Computational Terminology*, Amsterdam & Philadelphia, John Benjamins.

CABRÉ, M.T. (1992), *Terminology. Theory, methods and applications*, Amsterdam & Philadelphia, John Benjamins.

CALZOLARI N., CHOUKRI K., GAVRILIDOU M., MAEGAARD B., BARONI P., FERSØE H., LENCI A, MAPELLI V., MONACHINI M., PIPERIDIS S., (2004), *ENABLER Thematic Network of National Projects: Technical, Strategic and Political Issues of LRs,* in LREC-2004 Proceedings, Lisbon.

GAVRILIDOU M., DESIPRI E. (2003), *Final Version of the Survey*, ENABLER Deliverable 2.1.

GAVRILIDOU M., DESIPRI E., LABROPOLOU P., PIPERIDIS S., MONACHINI M., SORIA C. (2003), *Technical specifications for the selection and encoding of multilingual resources,* INTERA Deliverable D5.1.

GAVRILIDOU M., GIOULI V., DESIPRI E., LABROPOLOU P., MONACHINI M., SORIA C., PICCHI E., RUFFOLO P., SASSOLINI E. (2004), *Report on the multilingual resources production,* INTERA Deliverable D5.2.

KRAUWER S. (1998), *ELSNET and ELRA: A common past and a common future,* in *ELRA Newsletter,* Vol.3, N. 2.

MAEGAARD B., CHOUKRI K., MAPELLI V., NIKKOU M., POVLSEN C. (2003), *Language Resource – Industrial Needs,* ENABLER Deliverable D4.2, Copenhagen.

JACQUEMIN, C. (2001), *Spotting and Discovering Terms through Natural Language Processing*, Cambridge, MA and London, The MIT Press.

SAGER, J.C. (1990), *A Practical Course in Terminology Processing*, Amsterdam & Philadelphia, John Benjamins.

TALN 2005
Developments Towards an Electronic Amharic Corpus

Daniel Yacob
Ge'ez Frontier Foundation
7802 Solomon Seal Dr, Springfield, VA 22152, USA
yacob@geez.org

Abstract

The state of Amharic natural language processing was aptly assessed at TALN 2003 by Atelach, Asker and Mesfin. A public Amharic corpus and a comprehensive lexicon were two of the most needed items in absence for Amharic language researchers. Since the 2003 assessment some progress has been made in these two areas and researchers have begun informal collaboration to address the common goal of developing these public resources. In this same period Ethiopia's legal system has changed to cloud the issue over what the legal status of an Amharic corpus would be. While a promising start is underway, corpus developers and researchers alike will have to familiarize themselves with the new legislature in Ethiopia and reexamine the status of their holding to avoid potential unintended violations.

1 Introduction

Amharic is the most studied and best understood language of Ethiopia, it also serves as the country's lingua franca. Researchers today, both inside and outside of Ethiopia, are increasingly interested in computational investigations of the Amharic language. The lack of a freely available electronic corpora, lexicon, and transcription standard, coupled with the complexities of Amharic orthography are a significant barrier to would be researchers.

Amharic, along with its ten sibling Ethio-Semitic languages found in Ethiopia and neighboring Eritrea, is written in the Ethiopic syllabary. Amharic has been a written language for roughly 600 years and has as rich legacy of both typeset and calligraphic literature. Significant amounts of electronic corpora in Amharic, however, did not exist prior to the 1990s. The rise of desktop and internet publishing over the last decade has helped accumulate a body of data but of limited value. This paper will review the available materials, their usefulness, complications in working with Amharic orthography, and the new legal ramifications that face corpus development in the research community.

2 Electronic Corpus

Prior to the boom in desktop publishing in the late 1980s electronic text in Amharic existed as the product of experimental explorations into computer environs and was unappreciable in quantity. As personal computers came down in price in the late

1980s and early 1990s and word processing software became more practical and extensible the stage was set for Amharic publishing to get underway in a significant manner. However, due to the political status at that time the vast majority of Amharic desktop publishing would occur outside of Ethiopia amongst the Ethiopian Diaspora.

The largest bodies of work developed would then be in periodical literature of the Diaspora which were news publications and largely of political content. Very little of this content likely survives in the present day and may no longer be accessible under modern computer systems. Ethiopic script lacked a standard for representing the letters of the syllabary electronically until Amendment 10 to the ISO-10646 standard in 1997 which later becoming a part of a formal Unicode standard in 2000 when major software vendors would begin to support Ethiopic script. In this time more than seventy computer encoding systems were devised for Ethiopic script, none compatible with the other, and supported presently by only a handful of surviving software companies. Unlike the well established western languages, having an electronic document in Amharic does not equate readily to being able to read the document.

Following the change of governments in Ethiopia in 1991, the liberalization of press laws and downward cost of personal computer becoming affordable to Ethiopian businesses, the stage was then set by the mid 1990s for the sustained generation of electronic materials. Periodical publications such as weekly newspapers and magazines would lead this production of electronic content. Unfortunately hard disk capacities were still limited and there was little to no appreciation of the electronic record. A newspaper team producing a publication one week would delete the files on the following week without giving it a second thought.

Book authors and many magazine producers did not publish their own materials. Rather, they would take handwritten manuscripts to a publishing house that would typeset the text electronically and produce the publication for a fee. The electronic version was generally not preserved for the future except in the case of books. Authors would not automatically be given an electronic copy however since the typing and typesetting was considered a service with the cost absorbed by the publishing house. An electronic copy could, reluctantly, be made available to the author for a significant larger fee.

Perception would begin to change with the arrival of the internet and exposure to online newspapers. Electronic information exchange was nearly non-existent in Ethiopia prior to email service and given the encoding difficulties there was little expectation that a document written on one computer should be readable on another.

Eight months following the arrival of public internet service to Ethiopia the Ethiopian News Headlines (ENH) service was launched that featured a selection of articles from the capital city's newspapers. These articles were retyped from the newspapers and published in the Unicode character set amongst others. To date the news service has generated over 10,000 articles from more than 110 newspapers that are available in its archives and in zipped archived bundles for each month and year that may be downloaded freely and used under "fair use" rules governing literature. This corpus served as the basis of the collocation extraction research of Sisay and Haller in 2003 as well as the speech recognition research of Solomon Abate among others.

While the ENH offers a large collection of newspaper articles, the typical article is under 500 words in length, is largely of political content, and as retyped content is subject to having more typographic errors than the original document.

Other online publications in Amharic with large numbers of documents include the government news organ the Ethiopian News Agency (ENA) and the private, yet government affiliated, Walta Information Service. The ENA is noted for offering good quality Amharic language but has not yet begun publishing in Unicode leaving researchers to find a way to convert the ENA materials into a usable form. Walta on the other hand has begun partial publishing in Unicode since the middle of 2003 and has nearly made a full transition to Unicode (archives remain in a legacy encoding). The archives of each are available on a per-article basis. Like the ENH articles they are relatively short but the content characteristic of these services is broader, offering more than political subjects.

Research into Amharic machine translation is underway in 2004 at Stockholm University under the direction of Dr. Lars Asker where work on a parallel lexicon is being developed. Parallel corpus is even more limited for Amharic and again relies heavily on news services, in this case the bilingual Walta. Translation of Open Source software made in recent years offers a significant amount of translated phrases (over 40,000 phrases). However, the phrases are usually of 3 words or less, perhaps 10% representing one more sentences, having 20% redundancy, are of technical content, and reflecting the mixed quality of untrained volunteer translators.

A small collection of books have been typed up and made available to the research community. Under arrangement with the authors, the researcher must first sign a contract assuring that he or she will use the content for research purposes only and not commercially. This library made available by the Ge'ez Frontier Foundation offers much lengthier content under broader subject matter. The typographic correctness of the materials remains uncertain at this time.

3 Lexicon & Spelling

Atelach, et.al. (2003) identified the need for an Amharic lexicon and spelling checker as essential to the research community. The Ge'ez Frontier Foundation has been working actively in 2004 and continuing into 2005 to provide a basic word list. The lexicons given in the dictionaries of Amsalu Aklilu (1979 EC), Desta Tekle Wold (1962 EC) and Tesema Habte (1951 EC) are being extracted to form the basis for a comprehensive public lexicon. A rough version of the Amsalu lexicon comes with the Aspell version 0.60.2 open source spelling checker and has some initial tagging for affix rules.

A more refined version of the Amsalu lexicon will be available in the coming months and affix tagging will be an ongoing effort for some years to come. Amharic is a highly inflected language where the affix rules are largely governed by the presence of a midfix and many 10s of thousands of derived forms of some nouns and verbs become possible. The tagging of the lexicon as per their derivational classes will be essential to detecting proper word formation.

The comprehensive lexicon effort should produce its first unified lexicon in 2005. The next stage in refining the lexicon will be to resolve differences in spelling

that may emerge. The three dictionaries that form the basis of the lexicon are widely considered to be of the highest quality so discrepancies are expected to be minimal.

A complexity that enters into Amharic spelling are the presence of Ge'ez loan words and words derived from a Ge'ez root. Ge'ez is the ancient language of Ethiopia that is analogous in the role that Latin played for the Romance language of Europe. Ge'ez had a richer phonemic inventory and required additional letters for its orthography. In Amharic orthography these additional letters from Ge'ez would take on the phonemic value of its nearest neighbor. The result being two syllabic series for 's' ('ሰ' and 'ሠ'), two series for 'ts' ('ጸ' and 'ፀ'), two for 'a' ('አ' and 'ዐ') and 4 for 'h' ('ሀ', 'ሐ', 'ኀ' and 'ኸ'). This redundancy in Amharic becomes a source of confusion and the letters are treated as interchangeable by the lay person. Common Amharic spelling then becomes highly flexible and "correctness" is not a matter of precision but one of acceptable proximity. For example the fourth month of the year, canonically "ታኅሣሥ" ("Tahsas") may have any of the logical, phonetically equivalent, forms:

ታኅሣሥ	ታሕሣሥ	ታህሣሥ	ታኽሣሥ
ታኅሣስ	ታሕሣስ	ታህሣስ	ታኽሣስ
ታኅሳሥ	ታሕሳሥ	ታህሳሥ	ታኽሳሥ
ታኅሳስ	ታሕሳስ	ታህሳስ	ታኽሳስ

Table 1: Logical Amharic Renderings of the month "Tahsas".

While logical under the rules of the syllabary, the final column however is not also probable, leaving us with only 12 renderings likely to be found. Equating of the various spellings in text processing has been accomplished through the device of equivalence classes for the Amharic syllabary whereby:

[=ህ=] ≡ {ህ,ሕ,ኅ,ኽ} *(all 'h' syllables)*
[=ሳ=] ≡ {ሣ,ሳ} *(all "sa" syllables)*
[=ስ=] ≡ {ሥ,ስ} *(all 's' syllables)*

The equivalence classes may then be applied to form the regular expression for the possible renderings in the expression string "ታ[=ህ=][=ሳ=][=ስ=]". A metaphone matching approach has also been devised for Amharic where all renderings simplify into the single string "ትህስስ" ("thss"). The Perl language package, Text::Metaphone::Amharic in fact applies the Amharic character classes in its metaphone implementation also via the Perl package Regexp::Ethiopic::Amharic.

The Regexp::Ethiopic package contains classes for a few other Ethiopian languages using Ethiopic script. The same month in Tigrinya would render canonically as "ታሕሣሥ" (taħsas) but could not be matched by the Regular expression derived for Amharic as 'ሕ' ('ħ') would no longer be a member of the [=ህ=] equivalence set (likewise 'ኽ' which becomes 'x' in Tigrinya). The correct regular expression for valid Tiginya renderings is then "ታሕ[=ሳ=][=ስ=]" matching the 2nd column in the above table and the metaphone key becomes "ትሕስስ".

Without exception the Ethio-Semitic languages of Ethiopia and Eritrea use the Ethiopic script, many but not all of the members of the other language families (Cushitic, Omotic and Nilo-Saharan) will also use Ethiopic script. Unlike the Ethio-Semitic languages that have literal histories of some length, the primary issue that has prevented recent corpora development for the less populous languages has been the lack of character encoding support for their written elements. Only since Unicode 4.1.0 (March 31. 2005) has there been a standard supporting the written syllables of Bench, Blin, Me'en, Mursi, Sebatbeit, Suri and Xamtanga.

Prior to the last change of governments the early 1990s, it was legally permissible by the central government for publications to be in only one of Amharic, Tigrinya or Afan Oromo. Needless to say, significant corpora in the remaining 75 or so languages are not to be found. Since then some language communities have elected to adapt Latin script as the basis for their orthography. Spelling problems encountered by such communities are primarily related to the representation of long and short vowels and geminated consonants, none of which could have been represented in Ethiopic. Hence "Adooleessa", the seventh month of the year in Afan Oromo, is likely to be rendered with any number of doubled letters, e.g. "Adolleesaa", "Addoolessaa", etc. Prevalence of a multitude of renderings here is due in large part to user confusion over where lengthening is needed and the inherited acceptance of lax spelling practices from Amharic. By migrating to Latin script however these languages may enjoy the benefit of the vast computational resources and methodologies developed for western languages.

The resources discussed for Ethiopic script are effective for coping with the complexities of modern orthography for tasks such as pattern matching applied in stages of text retrieval and spelling correction. Spelling correction very much depends upon a good quality lexicon. Amharic does not have an authoritative reference for spelling nor a recognized authority responsible for defining Amharic rules and vocabulary. The closest such authority would either be the Ethiopian Orthodox Church or The Amharic Language School at Addis Ababa University.

The combined lexicons of the three authors mentioned will provide a strong basis for a spelling checker as a database for canonical spellings. In a number of cases it will likely be necessary to allow for two and three acceptable renderings of a word. The faculty of the Amharic Language School at Addis Ababa University will be enlisted to refine and add to the unified lexicon as well as decide which amongst alternative spellings can be deemed acceptable.

4 Ethiopian Intellectual Property Laws

Until very recently Ethiopia was without copyright or intellectual property laws. Foreign and domestic works could be republished with impunity. Researchers could enjoy the luxury of using textual materials without any concern for lawsuits but will now have to reexamine their holdings.

Wishing to join the World Trade Organization (WTO) Ethiopia has begun bosltering its legal system to protect intellectual properties. In 2004 Ethiopia passed a highly progressive copyright proclamation which covers a wide range of media from books, pamphlets and speeches to music, photographs, software and databases. Copyright is protected for the life of the author plus fifty years. The copyright

directorate is under the Ethiopian Intellectual Property Office established in 2003 (Wondwossen 2004). Ethiopia has become a member of the World Intellectual Property Organization (WIPO) in 1998.

The new laws are not yet well understood by the public nor the judicial system and may be difficult to apply as both go through a learning period. Lawsuits have already been brought against suspected violators by copyright holders and in some cases dismissed when the suite was initiated out of context. When the copyright law applies will, in a practical matter, be learnt by the society through judicial trial and error.

5 Status & Conclusion

The informal collaboration of researchers toward the development of an Amharic corpus has advanced passed the recognition of the problem and participants are presently reviewing applicable standards, tools, and related initiatives to launch a formalized effort. Oxford University has generously come forward and offered to serve as the corpus repository under its Open Archives Initiative which requires that the Text Encoding Initiative guidelines be followed. The home for the Amharic corpus will most likely be at Oxford but the requirements are still being studied. The formal phase of the Amharic corpus initiative is expected to get underway in early June following the resolution of these matters.

Internet newspaper archives, the few available books to researchers, and software translations provide a sizable corpus of Amharic text in electronic form but do not yet represent a balanced corpus as prescribed in Atelach, et.al. (2003). The unified lexicon in development while a promising start, represents no more than a collection of raw materials and may not be of a widely usable quality for several years. The critical mass of resources required for natural language processing of Amharic to "take off" is likewise some years away though we can now say that important steps are underway. While the resource collection is building, corpus providers and Amharic researchers are advised to familiarize themselves on Ethiopia's new copyright and intellectual property laws avoid the unintended missteps.

References

Amsalu Aklilu (1979 EC), **አማርኛ-እንግሊዝኛ መዝገበ ቃላት** *Amharic-English Dictionary*, Kuraz Publishing Agency.

Atelach Alemu, Lars Asker, and Mesfin Getachew (2003). *Natural Language Processing for Amharic: Overview and Suggestions for a Way Forward*, In Proceedings of TALN 2003 Workshop on Natural Language Processing of Minority Languages and Small Languages, Batz-sur-Mer, France, June, 2003.

Atelach Alemu, Lars Asker, and Gunnar Eriksson (2004). *Building an Amharic Lexicon from Parallel Texts*, In Proceedings of First Steps for Language Documentation of Minority Languages: Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilation, a Workshop at LREC2004.

Atkisonson, Kevin (2004), *GNU Aspell,* http://aspell.net/, GNU, v0.60.2, December 27, 2004.

Daniel Yacob (2004a), Regexp::Ethiopic, http://search.cpan.org/~dyacob/Regexp-Ethiopic/, CPAN, Version 0.14.

Daniel Yacob (2004b), Text::Metaphone::Amharic, http://search.cpan.org/~dyacob/Text-Metaphone-Amharic/, CPAN, v0.11.

Desta Tekle Wold (1962 EC), **አዲስ ያማርኛ መዝገበ ቃላት** *Addis Yamarña Mäzgba Qalat*, Artistic Printers, Addis Ababa.

Ethiopian News Headlines (1989-1997 EC), Newspaper Archives, ftp://archives.news.com.et/pub/ENH/, Addis Ababa.

Sisay Fissaha and Johann Haller (2003). *Application of Corpus-based Techniques to Amharic Texts*, In *Proceedings of the MT Summit IX Workshop on Machine Translation for Semitic Languages: Issues and Approaches*, New Orleans, Louisana, September 2003.

Tesema Habte Mikael Gisew (1951 EC), **የአማርኛ መዝገበ ቃላት** *(yä'ämärñä mäzgäbä qalat),* Addis Ababa.

Wondwossen Belete (2004), *The Intellectual Property System in Ethiopia,* Ethiopian Intellectual Property Office, Addis Ababa, December 2004.

TALN 2005 - RECITAL 2005

12ème conférence annuelle sur le Traitement Automatique des Langues Naturelles

9ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues Naturelles

---

ATELIER

LANGUE DES SIGNES

Atelier Traitement Automatique des Langues des Signes TALS 2005

A. Braffort, C. Cuxac, P. Dalle,
B. Garcia, António C. da Rocha Costa, R. Sabria

Motivation

Le Traitement Automatique des langues des Signes est un domaine de recherche très récent et encore peu développé, particulièrement en France. L'objectif de l'atelier Traitement Automatique de la Langue des Signes (TALS) est de réunir les chercheurs s'intéressant à la modélisation de la Langue des Signes, qu'ils soient linguistes ou informaticiens, afin de donner une impulsion aux collaborations pluridisciplinaires dans le cadre du TALS.

Nous avons à ce titre décidé de réserver une demi-journée à des présentations de travaux de recherche et à une table ronde consacrés à un corpus d'étude commun, réalisé spécialement pour cet atelier. Un domaine encore jamais abordé en informatique et quasiment pas en linguistique concerne le dialogue en LS. Il s'agit pourtant d'un sujet important et d'actualité, qui constitue d'ailleurs un des axes de travail du RTP « Dialogue et Interaction » du département STIC du CNRS. À cette fin nous avons choisi de réaliser un corpus de dialogue en LSF, qui est le premier du genre, afin d'amorcer une réflexion sur cet aspect[1].

Du fait d'un travail direct des linguistes français sur le terrain, dès l'origine, leurs descriptions se caractérisent par une appréhension de la LSF dans toutes ses dimensions, et, en particulier, par une prise en compte centrale de l'iconicité et des modes pertinents d'utilisation sémantico-syntaxique de l'espace. Ceci requiert le développement de modèles bien spécifiques, tant en linguistique qu'en informatique. On peut espérer que ces recherches originales pourront ouvrir des pistes fécondes, aussi bien pour l'étude de la gestualité coverbale que, plus largement, pour celle de l'ensemble des langues naturelles.

Thèmes abordés par l'atelier

Le nombre important de soumissions proposées pour cet atelier montre un besoin de la communauté de disposer d'un lieu de rencontre et d'échange. Pour cette année, les communications retenues concernent spécifiquement les thèmes liés au Traitement Automatique des Langues des Signes. Il serait certainement utile d'envisager pour la suite un

---

[1] Merci à l'institut IRIS de Toulouse d'avoir conçu et réalisé ce corpus de dialogues dans des délais très courts pour nous permettre de le proposer à la communauté scientifique (IRIS : http://www.lesiris.free.fr/).

atelier d'une durée plus importante, incluant les autres aspects des sciences humaines non couverts par TALS'2005.

Les thèmes couverts par les articles ci-après portent sur l'analyse et la compréhension d'énoncés signés (lexique, morphosyntaxe, sémantique, structuration du bas niveau), la modélisation de l'espace de signation, les approches cognitives, les formes graphiques, l'annotation de corpus, l'analyse et la génération automatique.

Programme de l'atelier

Session portant sur le corpus de dialogue « TALS 2005 »

« Passif et inverse en langue des signes française » *P. Guitteny*

« Travail contrastif sur les moyens d'annotation de corpus LSF (partition et SignWriting) visant l'analyse linguistique du domaine référentiel » *I. Fusellier et L. Boutora*

« Modélisation des relations spatiales en langue des signes française » *A. Braffort, B. Bossard, J. Segouat, L. Bolot et F. Lejeune*

« Modélisation de l'espace discursif pour l'analyse de la langue des signes » *B. Lenseigne et P. Dalle*

Autres Sessions

« Pour une iconicité corporelle » *D. Boutet*

« Construction/déconstruction de l'espace de signation » *A. Risler*

« Verbes et actants en Langue des Signes Française » *L. Kervajan, E. Guimier De Neef et J. Véronis*

« Problèmes et méthodes pour l'analyse d'énoncés en LSF » *A. Balvet et M.-A. Sallandre*

« Système d'annotation et de segmentation de gestes de communication capturés » *A. Héloir, S. Gibet, N. Courty et M. Raynaud*

« Using SignWriting as a Phonetic Notation System » *V. Bonow Boeira, L. R. Volz de Oliveira, D. Souza Madeira, A. C. da Rocha*

« Semantic searching for SignWriting » *S. Aerts, B. Braem, K. Van Mulders et K. de Weerdt*

« Variations dans la représentation écrite d'un signe en SignWriting » *G. Aznar et P. Dalle*

Passif et inverse en langue des signes française

Pierre Guitteny

Signes – Université Michel de Montaigne, Bordeaux III
33 chemin Pomerol 33000 Bordeaux
pierreguitteny@wanadoo.fr

Mots-clés : Langue des signes, Syntaxe, Passif, Inverse

Keywords: Sign Language, Syntax, Passive, Inverse

Résumé Les verbes en langue des signes sont souvent dotés d'un mouvement réversible. Il est alors difficile de parler d'actif et de passif : un décalage de point de vue permet de passer de l'une à l'autre interprétation. Cependant, la langue des signes n'est pas exempte de passif prototypique. C'est pourquoi nous proposons de distinguer, en langue des signes, passif et inverse – comme cela est le cas pour d'autres langues.

Abstract In sign language, verbs can often have a reversible movement. It is then difficult to speak about active or passive voice: a shift from point of view passes from the one to other interpretation. However, sign language is not without prototypic passive. This is why we propose to distinguish, in sign language, passive and inverse - as that is the case for other languages.

1 Introduction

C. Cuxac (Cuxac, 2000 : 209) remarque que : « *pour l'orientation active ou passive du verbe, c'est la direction du mouvement du verbe qui change son orientation sémantique, au sens propre comme au sens figuré. L'opposition actif/passif, en quelque sorte neutralisée en raison de la spatialisation des relations actancielles, n'a donc pas de raison d'être. [...] un verbe comme [INFORMER] est tout autant 'informer' qu' 'être informé', un verbe comme [INVITER], tout autant 'inviter' qu' 'être invité'*. » Il est vrai que beaucoup de verbes, de diathèses peuvent être inversés en langue des signes : il suffit de changer l'orientation d'un verbe pour intervertir les rôles d'agent et de patient. Nous souhaiterions toutefois apporter une nuance à ces affirmations : à notre avis, en langue des signes, comme dans d'autres langues, il convient de distinguer passif et inverse...

2 Caractéristiques du passif

Les manuels scolaires utilisent souvent des définitions adaptées à une langue particulière, et des définitions générales qui ne couvrent pas l'ensemble des cas de la langue. Ainsi, concernant le passif, il est souvent écrit qu'*un verbe est à la voix passive quand le sujet désigne l'être ou la chose qui subit l'action indiquée par le verbe*. Or de nombreux verbes à la voix active s'emploient avec un sujet qui subit l'action, comme 'je subis' ou 'j'endure' – comme en langue des signes le signe [se retourner contre soi] (le dos de la main plate venant toucher le nez). D'autres définitions insistent sur les liens entre phrases actives et phrases passives, mais de nombreuses phrases actives ne peuvent pas être passivées, et certaines formes passives n'ont pas de correspondance à l'actif.

Les recherches linguistiques concernant le passif ont été nombreuses, et ont produit de nombreuses définitions. Parmi celles-ci, nous retiendrons les plus répandues. Ainsi, D. Creissels écrit : « *Le passif canonique est un mécanisme qui, opérant sur un verbe transitif, produit une forme intransitive dérivée dont le sujet reçoit exactement le même rôle que l'objet de la construction transitive.* » (Creissels, 2004, ch. 12) Pour C. Muller, le passif consiste à : « *reléguer au rang de relation facultative de dernier rang la relation prédicative du premier argument au verbe, avec ou sans modification (temporelle, aspectuelle) dans la sémantique du verbe.* » (Muller, 2002 : 223) Avec d'autres termes et d'autres outils, la grammaire générative sous sa forme classique analyse le passif en notant que la morphologie passive absorbe le rôle thématique du sujet (de l'argument externe) ainsi que l'assignation d'un cas accusatif, ce qui provoque le mouvement de l'argument interne venant occuper la place restée vide. Autrement dit, le verbe transitif devient intransitif, inaccusatif, et l'objet se déplace pour occuper la place du sujet. (Chomsky, 1981, ch. 2) Le point commun entre ces différentes définitions est – pour employer des termes plus simples – que le passif consiste essentiellement à supprimer, ou tout au moins à rendre secondaire, la place de l'agent. Il s'agit donc d'un point de vue particulier sur la scène présentée.

3 Le passif en langue des signes

Est-il possible d'utiliser, pour l'analyse de la langue des signes, des catégories créées pour les langues vocales ? En ce qui concerne le passif, les définitions des manuels scolaires insistant sur le participe passé, l'auxiliaire ou le C.O.D. sont évidemment inadaptées. Par contre, les définitions linguistiques mettant l'accent sur les places respectives de l'agent et du patient peuvent tout à fait concerner la langue des signes.

Est-il souhaitable d'utiliser ces catégories ? Il s'agit là d'une question largement débattue... Un des intérêts de cette reprise peut consister justement à comparer les structures des langues vocales à celles des langues des signes, ne serait-ce que pour montrer leurs différences...

Si l'on reprend donc la définition linguistique du passif : structure consistant à reléguer au second plan – voire à supprimer totalement – l'agent, la langue des signes permet une telle expression. La manière la plus courante, comme le note C. Cuxac, consiste à inverser l'orientation des verbes : en transfert personnel par exemple, les verbes transitifs 'actifs' ont souvent une direction et/ou un mouvement de l'agent vers le patient. Ils prennent un sens passif avec le mouvement ou la direction inverse.

4 Un exemple du corpus TALS (séquence GB) :

La locutrice B parle d'une séance d'arts plastiques au Musée avec des enfants. Elle dit qu'elle était sollicitée par beaucoup de questions de la part des enfants. Au début de cette intervention, B signe le verbe [appeler] à plusieurs reprises, orienté vers elle-même : elle était appelée par les enfants. Ce qui est particulier ici est que ce verbe est signé de différentes façons : tantôt avec la main droite, tantôt avec la main gauche, avec différents mouvements d'épaule, différentes orientations, montrant la multiplicité des appels de la part des enfants. L'origine de ces appels, l'agent de ces verbes est connu 'en général' : il s'agit des enfants présents lors de cette animation. Cependant, chacun de ces appels provient d'un enfant situé spatialement – à droite, à gauche, en face, etc., mais non identifié personnellement. Cette origine non précise des appels, non clairement identifiée, ces agents non nommés, produisent un énoncé qui met clairement l'accent sur le rôle du patient, sur le point de vue de B – peu importe que ces appels proviennent de Pierre, Paul ou Jacques – d'ailleurs, elle ne s'en souvient peut-être pas elle-même. Dans cet énoncé, le point de vue adopté est celui du patient ; l'agent est mis au second plan. Une bonne traduction devra prendre une forme passive pour bien refléter le signifié de cet énoncé – comme beaucoup de phrases passives, une traduction par la contrepartie active est possible, mais elle reflèterait moins bien le point de vue exprimé.

Pour prendre une image cinématographique, la caméra se dirige sur le patient et, par un mouvement de focale et de profondeur de champ, les enfants sont mis au second plan ou dans le flou – par exemple, on ne voit plus des enfants que les bras qui se lèvent.

5 Passif et inverse

Pour certaines langues comme l'Algonquin, les recherches linguistiques distinguent passif et inverse (Bresnan J., Dingare S. et al., 2001).

Dans ces langues, l'inverse est une construction transitive, où le patient occupe la place de l'objet, l'agent est mentionné directement – sans ajout d'un oblique (comme une préposition en français), et où l'agent ne peut pas être omis.

Le passif, lui, est une construction intransitive, où le patient occupe la place du sujet, où l'agent peut être omis ; et s'il n'est pas omis, il est marqué par l'ajout d'un oblique.

6 Passif et inverse en langue des signes

Dans la séquence AB du corpus TALS 2005, une occurrence typique de construction inverse se trouve dans le dialogue entre les deux locuteurs : la locutrice B, pose une question sur ce qui est présenté dans le site de Websourd, avec le signe [présenter] tourné vers l'extérieur, et le locuteur A reprend (très brièvement) le même signe [présenter], mais tourné vers lui-même. On peut traduire ces deux occurrences du même signe par une alternance actif/passif : « qu'est-ce que présente ce site ? », « Ce qui est présenté consiste... ». Et cette alternance rend bien compte de ce qui est signifié par l'orientation de ce verbe : B intervient à plusieurs reprises avec des verbes actifs, du point de vue de ceux qui construisent et alimentent le site internet, tandis que A prend souvent le point de vue de l'internaute qui navigue sur le site internet de Websourd et qui voit défiler devant son écran de multiples informations.

On pourrait voir là un exemple de passif : la première phrase exprime le point de vue de l'agent, la deuxième celui du patient. Nous préférons y voir un exemple d'inverse. Une des différences entre ces deux structures est que dans la construction inverse, l'agent est clairement identifié – et, en langue des signes, une place précise lui est attribuée dans l'espace de signation. Il est impossible de changer cet emplacement – sauf raison particulière qui se doit d'être explicitée. Il s'agit d'un simple changement d'emplacement du point de vue ; l'information transmise est identique. Au contraire, dans le cas prototypique du passif, l'agent est soit inconnu, soit connu mais mis au second plan. La place qui lui est attribuée dans l'espace de signation est donc moins stricte : celle-ci peut varier sans que cela perturbe la compréhension du message. L'information transmise est modifiée.

Le fait que ces exemples concernent tantôt un récit, tantôt un dialogue n'a pas d'influence sur la question du passif et de l'inverse : dans toutes formes de discours – y compris l'expression poétique, il est possible d'utiliser ces différentes structures.

Pour continuer la comparaison avec le langage cinématographique, l'inverse est un simple mouvement de champ/contrechamp : la focale est identique, la profondeur de champ également, les personnages sont toujours présents et identifiés, seul l'emplacement de la caméra change.

7 Compléments

D'autre part, au passif, la mention de l'agent est possible ; mais l'absence de cette mention est également possible. Dans une phrase inverse, l'agent ne peut pas être effacé : même s'il n'est pas explicitement mentionné, sa place fixe suffit à signifier sa présence. Au passif, on peut ajouter, à la suite du verbe, un complément – soit sous forme prépositionnelle [par] ou [à cause de], soit sous forme de phrase indépendante [qui ? ...].

D. Creissels (Creissels, 1995 : 278s.) remarque d'ailleurs que dans certaines langues comme l'arabe ou le nahuatl, le passif n'admet pas de complément d'agent. Dans d'autres langues, le complément d'agent résulte d'un processus récent par lequel a été intégré à la phrase passive ce qui à l'origine constituait une phrase distincte. En tswana, une phrase comme : « l'enfant a été mordu par le chien » peut s'analyser comme : « l'enfant a été mordu, c'est le chien ».

On peut se poser la question d'un processus similaire en langue des signes : un complément d'agent prend souvent la forme d'une phrase séparée, introduite par une question [Je suis appelé. (Par) qui ?...]. Peut-être les compléments prépositionnels, comme ceux introduits par [par] ou [à cause de] proviennent-ils, à l'origine, de phrases indépendantes ?...

8 Exemples du corpus TALS 2005 (séquence AB) :

Au début de la séquence AB, le locuteur A prend le rôle de l'internaute qui navigue sur le site de Websourd, et qui peut choisir entre une présentation écrite ou une traduction signée – le verbe [signer] est orienté vers le locuteur, d'après la situation de l'internaute qui regarde les explications signées, diffusées par le site internet.

De même, à la fin de la séquence AB, il est question de contes présentés par deux personnes sourdes, et le locuteur A prend encore le rôle de l'internaute qui surfe sur le site et regarde les séquences vidéo qui lui sont signées. Le verbe [raconter] est donc orienté vers le locuteur.

Toutefois, si, dans ces exemples, nous voyons des exemples de verbes orientés vers le locuteur, donc avec un sens passif et une traduction possible par une phrase passive, l'orientation de ces verbes est précise quant à son origine : à chaque fois, l'origine du mouvement du verbe est clairement située à la place du site internet de Websourd. Le locuteur aurait très bien pu prendre le rôle des concepteurs de Websourd – et non celui de l'internaute qui surfe – comme le fait B. Les situations, et l'orientation des verbes, sont tout à fait inversables sans changement de signification.

9 Conclusion

Pour rendre compte des différentes formes que peuvent prendre les phrases de sens passif en langue des signes, nous proposons de distinguer deux formes : le passif et l'inverse. L'inverse est un simple changement de point de vue sans aucun changement de signification, tandis que le passif modifie l'information transmise, du fait du changement de focalisation portée sur les actants : l'agent est soit complètement occulté, soit mis au second plan, laissé dans le flou ; ce qui se traduit notamment pour l'agent par un emplacement moins précis. D'autre part, le passif peut être suivi d'une mention de l'agent, souvent sous forme de phrase indépendante, tandis que l'agent est toujours explicitement présent dans la structure inverse.

Distinguer ces structures peut, par ailleurs, permettre d'affiner les analyses du Traitement automatique des LS. Ainsi, le logiciel de modélisation de l'espace de signation, développé à l'IRIT de Toulouse, permet de réaliser aisément le mouvement inverse. Il pourrait être agrémenté d'une fonction de focale et de profondeur de champ, permettant de mieux refléter les nuances possibles de l'expression du passif en langue des signes.

Références

BRESNAN J., DINGARE S. & MANNING C.D., (2001), Soft Constraints Mirror Hard Constraints : Voice and Person in English and Lummi, *Proceedings of the LFG 01 Conference*.

CHOMSKY N. (1981, trad. 1991), *Lectures on Government and Binding*, Dordrecht, Foris Publications.

CREISSELS D. (1995), *Eléments de syntaxe générale*, Paris, P.U.F.

CREISSELS D. (2004), *Cours de syntaxe générale*, disponible sur : http://lesla.univ-lyon2.fr/

CUXAC C. (2000), *La Langue des Signes Française, Les voies de l'iconicité*, Paris, Ophrys.

MULLER C. (2002), *Les Bases de la syntaxe*, Pessac, Presses universitaires de Bordeaux.

Travail contrastif sur les moyens d'annotation de corpus de LSF (partition et Sign Writing) visant l'analyse linguistique du domaine référentiel

Ivani Fusellier-Souza (1) et Leïla Boutora (2)

Laboratoire SFL, UMR 7023 – Université Paris 8
2, rue de la Liberté – 93526 Saint-Denis
(1) ivani.fusellier@wanadoo.fr
(2) leila.boutora@neuf.fr

Mots-clés : Langue des signes, annotation de corpus, référentialisation de l'espace, forme graphique
Keywords : Sign language, corpus annotation, construction of spatial references, graphic form.

Résumé
L'objectif de ce travail est d'effectuer une confrontation entre une transcription en partition et une transcription au moyen de Sign Writing (SW) en termes de notation d'éléments linguistiques afin de faire émerger des solutions intermédiaires pour la transcription d'énoncés en langue des signes, notamment pour la notation du domaine référentiel[1] (personne, espace, temps et modalité).

Abstract
The aim of this work is to perform a confrontation between two different notation systems for signed languages, for the same video data (dialogue) : a "partition" transcription and a Sign Writing (SW) transcription. We want to note linguistic elements to find intermediate solutions to transcribe signed utterances, particularly for the notation of referential field (person, space, time and modality).

1 Les systèmes de transcription

On commencera par rappeler brièvement les grands systèmes de transcription[2] dont on dispose aujourd'hui pour transcrire et/ou annoter des corpus vidéo en langue des signes (monolinéaires, plurilinéaires, en partition, multimédia), ainsi que leurs points forts (compacts, gloses, simultanéité, image...) et leurs limites (paramètres non notés, séquentialité,

[1] Notamment dans les approches énonciatives (Kerbrat-Orecchioni, 1990 et Culioli, 1999) qui considèrent la langue dans son usage.
[2] Une grande partie de ces systèmes est présentée dans (Wilbur, 2001).

lourdeur...), et ce dans le but de bien faire ressortir les spécificités des systèmes étudiés ici.

On oppose donc : les systèmes monolinéaires qui peuvent utiliser soit des symboles phonétiques ou phonologiques (respectivement HamNoSys et Stokoe : problèmes de lisibilité) notant les éléments paramétriques du signe, soit le lexique des LV pour noter directement le sens du signe (BTS qui compense la séquentialité par des procédés de type factorisation et indices) ; les systèmes plurilinéaires (Johnston : lignes dédiées à la transcription séquentielle des éléments au moyen d'HamNoSys, à la glose et à la traduction en LV) ; les partitions (Cuxac, Bouvet, Sallandre, Fusellier-Souza) où l'on affecte une ligne à chaque élément manuel ou non manuel ; les systèmes multimédia qui intégrent la vidéo, en dynamique ou en statique (en partition : Sign Stream, Elan, Anvil, Ancolin ; plurilinéaire : Sync Writer (reprise de la trancription de Johnston avec la mimique faciale en plus) ; lexique seul : Kheiros (manuel) et Sign PS (manuel et non manuel). Une place à part doit être faite à Sign Writing, brièvement décrit dans la section suivante, qui s'apparente aux systèmes monolinéaires (symboles phonétiques), bien qu'il tente une exploitation particulière de la surface graphique.

2 Présentation de Sign Writing et du système en partition

Dans la perspective de ses créateurs, Sign Writing s'apparente plus à un système phonographique qu'idéographique. Il permet de noter la plupart des éléments de la LS dans un mode articulatoire et du point de vue du locuteur (en émission) : configuration, orientation, emplacement, contact et mouvement des mains, expressions du visage, direction du regard, mouvements de la tête et des épaules, éléments de prosodie. Son apport essentiel par rapport aux autres systèmes phonétiques réside dans la disposition relative des éléments articulatoires au sein d'une vignette, permettant ainsi une appréhension globale des éléments d'un signe, lorsque les autres systèmes placent ces éléments de manière successive. Les symboles qui le constituent donnent un accès à la face signifiante des différents paramètres du signe. L'iconicité qui caractérise certains de ses symboles ne concerne d'ailleurs que le niveau signifiant et n'est donc pas liée au référent. Une description détaillée du système et des réflexions plus théoriques à son sujet sont présentées dans (Boutora, 2003). On notera que ce système a connu des modifications profondes initiées par ses utilisateurs et qu'il évolue toujours.

La partition se définit comme un système de notation (description) visant à établir les relations de multilinéarité et de simultanéité qui existent entre les éléments paramétriques, tout en essayant de restituer l'information véhiculée par la face signifiée (sens). Ce système s'inspire de la partition musicale où chaque paramètre est traité indépendamment sur l'axe horizontal tandis qu'en lecture verticale les paramètres sont analysés comme un ensemble s'articulant lors de la construction du sens. Un des avantages de ce système est qu'il peut être paramétrable en fonction des objectifs du chercheur. Actuellement, certains linguistes segmentent la partition selon le principe de ruptures verticales par des fragments isolés (barres séparatrices) des unités de temps matérialisées (Monteillard, projet LS-Colin). Ce procédé de segmentation d'unités de sens est pertinent si l'on veut faire une analyse quantitative ou bien si l'on veut faire référence à un fragment spécifique de la séquence transcrite.

3 Méthodologie

Dans un premier temps, nous avons procédé à l'analyse de l'extrait en nous appuyant sur la transcription en partition décrite précédemment et finalisée au cours d'un précédent travail (Fusellier-Souza, 2004). Nous avons ensuite juxtaposé les vignettes SW aux images extraites

de la vidéo pour en faciliter la comparaison. Les séquences discursives analysées se présentent sous forme d'interactions. Sur la vidéo nous avons deux interlocuteurs. Par conséquent, un dédoublement de la grille était nécessaire pour montrer la dynamique interactive du discours.

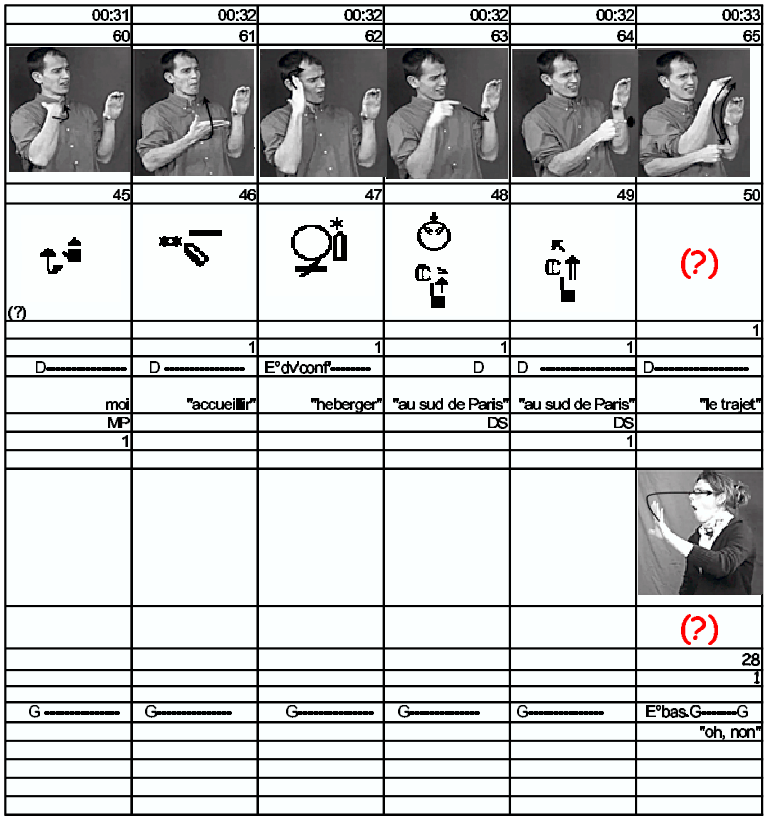


Figure 1 : Extrait de la double transcription partition/Sign Writing

On notera cependant que le corpus transcrit en SW que nous exploitons ici a été recueilli selon des modalités qui nous poussent à garder une certaine réserve quant aux conclusions que l'on pourra tirer de la confrontation des deux systèmes. D'une part, sans même être dans le cadre d'une démarche de production écrite spontanée puisqu'il s'agissait de retranscrire la vidéo d'un dialogue en LSF, l'utilisateur n'avait pas reçu la consigne d'effectuer une transcription linguistique de ce dialogue. D'autre part, l'utilisateur n'est pas locuteur de la LSF mais d'une LS étrangère.

4 Observations et pistes de réflexion

4.1 Remarques sur les données quantitatives et qualitatives

L'étude et la confrontation des deux transcriptions ont fait ressortir des données intéressantes à la fois quantitatives et qualitatives.

4.1.1 Données quantitatives

Sur une durée de 00'48", la séquence étudiée (CD1) comprend 96 unités temporelles et 114 unités de sens produites (77 pour G et 37 pour D, dont certaines se superposent, ce qui

explique que le nombre d'unités de sens est supérieur au nombre d'unités temporelles). Comme le montre le tableau ci-dessous, sur ces 114 unités de sens produites, certaines sont notées dans la transcription en SW, d'autres non.

Données de notation de SW	total		Locuteurs		Types d'unités
			G	D	
Unités de sens notées	86	dont	63	23	SS (signe standard), pointages
Unités de sens non notées	25	dont	2	2	SS (signe standard)
			5	7	RS (référence spécifique) : pointage anaphorique, personnel et spatial
			5	4	à valeurs modales

Figure 2 : Unités de sens notées et non notées en Sign Writing

4.1.2 Données qualitatives

1. L'utilisation de SW a permis de retranscrire la plupart des unités de sens construites par les paramètres manuels en notant le signifiant de manière très fidèle ;

2. On observe cependant l'absence de notation en SW de certains paramètres non manuels, pertinents pour l'analyse de la LSF (Cuxac, 2000) :

- La direction du regard dans la construction de la référence spécifique ou dans l'expression de la modalité et de la détermination (ex. fermeture des yeux : pertinente ou pas) ;

- La mimique faciale dans l'expression de valeurs modales (marqueurs négatifs, interrogatifs, évaluatifs et argumentatifs) et de valeurs qualitatives ou quantitatives (adjectivales et adverbiales) ;

- Les mouvements de tête et du tronc dans l'expression de valeurs modales de type argumentatif et dans l'expression des liens de coordination et/ou de subordination entre les énoncés.

3. De la même manière, on constate l'absence de certaines informations fondamentales pour l'interprétation du sens dans la notation de constructions de références spécifiques :

- Relation spatio-géographique exprimée par le proforme (reprise de frontière de lieu - Paris) et référentialisation des endroits au nord et au sud de la référence (Paris) ;

- Activation de l'espace par le regard : information linguistique sur la cible impliquée dans le procès de type dynamique (déplacer, aller/venir).

- Reprise anaphorique des entités discursives par des pointages anaphoriques.

4.2 Ouverture

Nous pensons à terme pouvoir tirer bénéfice de l'utilisation des fonctionnalités respectives des deux systèmes pour la transcription linguistique. Par exemple, il nous semblerait pertinent d'étudier une façon d'utiliser les symboles existants dans le système SW pour noter (dans la partition) certaines valeurs linguistiques de types modal et aspectuel (mimique faciale) ainsi que les valeurs référentielles de la direction du regard. Cette étude montre en outre qu'un travail de réflexion sur la notation de l'espace en langue des signes doit encore être mené.

Il serait aussi très intéressant de pouvoir comparer dans un avenir proche la production écrite d'un locuteur 1) de la LSF, 2) conscient du rôle linguistique assuré par ces éléments qui semblent indispensables à la compréhension d'un énoncé, avec l'extrait que nous venons d'analyser. Ceci nous permettrait de savoir si, dans des conditions « idéales », SW permet de noter les éléments qui nous renseignent sur l'utilisation linguistique de l'espace en LSF ? Une question subsidiaire pourrait être : ce même locuteur serait-il en mesure d'interpréter sans ambiguïtés le texte SW que nous venons de présenter.

Remerciements

Nous tenons ici à remercier Marianne Stumpf pour le travail de retranscription qu'elle a effectué, ainsi que pour sa collaboration qui nous aura permis d'enrichir notre réflexion. Nous remercions aussi le groupe de recherche LS-Script pour nous avoir permis de mener à bien cette étude.

Références

BOUTORA L. (2003), *Etude des systèmes d'écritures des langues vocales et des langues signées. Description et analyse comparatives de deux systèmes « idéographiques » et de Sign Writing.* Mémoire de DEA, Université Paris 8, Saint-Denis.

CULIOLI A. (1999), *Pour une linguistique de l'énonciation*, tomes 1, 2 et 3, Paris, Ophrys.

CUXAC C. (2000), La Langue des Signes Française : les Voies de l'Iconicité, *Faits de Langues* Vol. 15-16, Paris, Ophrys.

FUSELLIER-SOUZA I. (2004), *Sémiogenèse des langues des signes. Primitives conceptuelles et linguistiques des Langues des Signes Primaires (LSP). Étude descriptive et comparative de trois LSP pratiquées par des personnes sourdes vivant exclusivement en entourage entendant.* Thèse de doctorat. Université Paris 8, Saint Denis.

KERBRAT-ORECCHIONI C. (1990), *Les interactions verbales,* tomes 1 et 2, Paris, Armand Colin.

LS-COLIN : http://www.irit.fr/ACTIVITES/EQ_TCI/EQUIPE/dalle/cognitique/

LS-SCRIPT : http://lsscript.limsi.fr/

SIGN WRITING : http://www.SignWriting.org/

WILBUR R. (ed) (2001), Sign Transcription and Database Storage of Sign Information, *Sign Language and Linguistics*, Vol. 4 – 1/2 .

Modélisation des relations spatiales en langue des signes française

A. Braffort, B. Bossard, J. Segouat, L. Bolot et F. Lejeune

LIMSI/CNRS
Bat 508, Campus d'Orsay, 91 403 Orsay cedex
{annelies.braffort, bruno.bossard, jeremie.segouat, laurence.bolot}@limsi.fr

Mots-clés : Langue des Signes Française, Relation spatiales, Modélisation Sémantico-Cognitive, Avatar Signant

Keywords: French Sign Language, Spatial relations, Semantico-Cognitive Modelling, Signing Avatar

Résumé Nous présentons dans cet article les bases d'un modèle générique permettant de représenter les relations spatiales entre entités en langue des signes française (LSF), ainsi que la manière d'utiliser ce modèle pour la génération automatique d'énoncés.

Abstract We present in this paper the fundaments of a generic modelling allowing us to represent spatial relations between entities in French Sign Language, and the way to use this model for automatic generation.

1 Introduction

Comme toutes les langues des signes, la LSF utilise l'espace de signation placé devant le signeur pour décrire des entités, conjuguer des verbes directionnels, exprimer des relations spatiales... (Cuxac, 2000). Ces dernières (« dans », « au nord de », « à côté de »...) sont généralement exprimées par l'intermédiaire de pointages et de configurations spécifiques réalisées à des emplacements donnés dans l'espace de signation. Une formalisation de ce type de relations est possible par le biais des grammaires cognitives (Lejeune, 2004). Cet article présente les principes de la modélisation informatique basée sur l'utilisation de cette formalisation et qui se veut complètement indépendante de son implémentation au sein d'une application informatique (Braffort, Lejeune, 2005). Une implémentation dans le cas de la génération d'énoncés réalisés par un avatar signant est proposée à des fins d'évaluation du modèle.

2 Traitement automatique des langues des signes

La plupart des études portant sur la modélisation informatique des structures linguistiques des langues des signes (LS) sont menées dans le contexte d'une application donnée, typiquement la génération automatique d'énoncés en LS réalisés par un avatar signant. Dans plusieurs études, la modélisation comporte des représentations syntaxiques, voire sémantiques. Parfois, ces modèles sont très éloignés de la réalité du fonctionnement des LS (génération de dactylologie, d'américain signé…). Certains projets intègrent des modules permettant de représenter des phénomènes linguistiques propres aux LS, tels que les verbes directionnels (projet européen eSign), mais ne permettent pas de représenter d'autres types de spatialisations. Une étude propose une approche vraiment spécifique aux LS, basée sur la modélisation de l'espace de signation (Huenerfauth, 2004). Le système, pas encore implémenté devrait permettre de représenter des énoncés de type « transfert situationnel ». En France, les premières études ont été menées dans un contexte spécifique. Dans (Braffort, 1996), la modélisation de l'espace de narration est utilisée pour l'interprétation d'énoncés comportant des verbes directionnels. Plus récemment, une modélisation de cet espace a été proposée dans le contexte de l'analyse de la LSF par traitement d'image (Lenseigne, 2004). Enfin, la formalisation proposée dans (Lejeune, 2004) a pour objectif la génération automatique. Ces deux approches nous semble cependant présenter des propriétés de généricité prometteuses. Nous nous basons ici sur le formalisme proposé dans (lejeune, 2004).

3 Modélisation des relations spatiales statiques

3.1 Formalisme

Dans sa thèse, F. Lejeune propose un ensemble de représentations spécifiques à la LSF. Une de ces représentations concerne les relations spatiales statiques, telles que « L'université est au nord de Paris ». Ce type d'énoncé est composé de **signes stabilisés** désignant des entités (humain, animal, objet, notion abstraite, action…), notés $[SIGNE]_{LSF}$, de **proformes** permettant de spatialiser les entités en un lieu donné, notés PF(signe, lieu), du **regard**, noté REG(lieu), qui est utilisé pour « instancier » ou « réactiver » un emplacement dans l'espace de signation, en particulier juste avant d'y placer un proforme (Cuxac, 2000) et aussi du **pointage** sur un emplacement dans l'espace de signation, noté POINT(lieu). Cet exemple est illustré Figure 1.

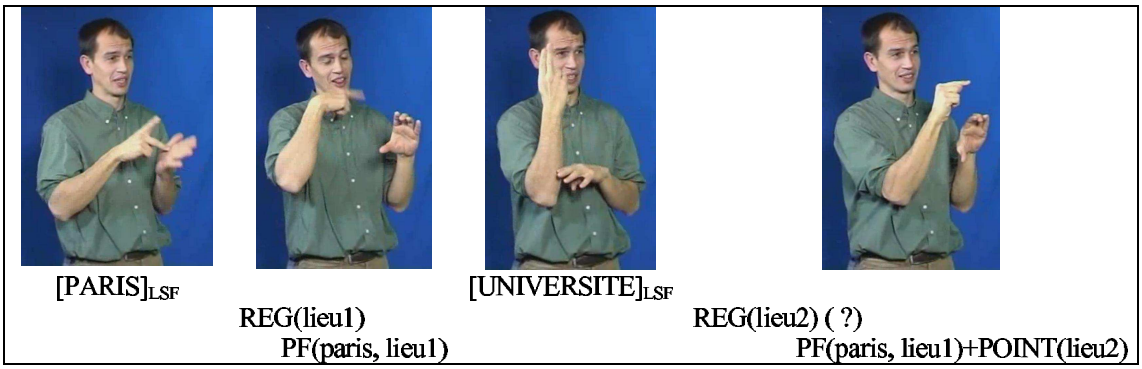


Figure 1 : Séquence « $[PARIS]_{LSF}$, là, $[UNIVERSITE]_{LSF}$, au nord »

Dans le formalisme proposé par Lejeune, la notion de repérage d'un entité par rapport à une autre se représente par la description formelle **<x REP y>** qui indique qu'une entité x est repérée par rapport à une entité y. Ce schème générique est instancié dans notre exemple de la manière suivante : **<L=OR$_{Nord}$(DET(LOC(Paris))) & Université REP IN(L)>**, qui précise la manière dont le repérage est réalisée. Cette représentation utilise des primitives sémantico-cognitives, telles que des propriétés des entités, des opérateurs et des relateurs entre entités. Les opérateurs et relateurs prennent comme arguments des schèmes et renvoient des schèmes. Ils peuvent ainsi s'utiliser de manière imbriquée. Dans notre exemple, on utilise :

- LOC(x), opérateur qui spécifie qu'une entité x, dans le contexte de l'énoncé, est un lieu ; elle peut donc servir à localiser une autre entité.

- DET(x), opérateur qui détermine un point de vue sur une entité x, par l'intermédiaire d'un proforme précis, précédé d'un regard en un lieu donné.

- OR(x) est un opérateur qui oriente un entité, ici il s'agit d'un lieu orienté selon le repère absolu des points cardinaux. OR donne une orientation au proforme.

- IN(x), opérateur topologique faisant référence à l'intérieur d'une entité x de type lieu. Ici, on fait référence à l'intérieur de l'espace induit par OR.

- x REP y, relateur indiquant que par rapport à une entité lieu y, l'entité x est repérée. Cette application du relateur correspond à un ordre privilégié repère-repéré.

Ainsi, cette description formelle permet d'exprimer que l'entité université est repérée dans un espace orienté au nord de l'entité Paris. Ce type de relation s'exprime à l'aide d'un proforme et d'un pointage. Le proforme fait référence à Paris et le pointage à l'université.

Ces différents relateurs et opérateurs sont définis de manière formelle au sein d'un système à base de règle. Ce système permet de construire des schèmes ou élaborer des séquences d'instructions permettant de passer d'un schème à l'énoncé correspondant.

3.2 Utilisation dans un contexte de génération

Ce formalisme est actuellement utilisé au sein d'une maquette informatique pour générer automatiquement des énoncés exprimant une relation spatiale entre deux entités. Le système prend en entrée le triplet {entité repère, relation, entité repérée} et affiche en sortie un avatar 3d qui signe l'énoncé correspondant en LSF. Le système procède en trois étapes.

- La première étape consiste à élaborer le schème correspondant à la relation. Pour cela, on dispose de bases de connaissance, comportant des informations sur certaines relations spatiales et certaines entités. Concernant les relations, les informations stockées spécifient par exemple le sens des mots « nord », « sud » en terme de relation spatiale dans le cas d'une relation de nature géographique. Par ailleurs, pour chaque entité disponible à ce jour dans la maquette, on spécifie la liste des proformes qu'il est possible de lui assigner et dans quelles conditions (selon que l'entité est considérée comme un individu, un lieu…).

- La deuxième étape consiste à élaborer à partir du schème la séquence d'unités gestuelles à faire réaliser par l'avatar. Pour cela, une séquence d'instructions correspondant au schème permet de construire les syntagmes correspondant à la relation. Dans notre exemple, l'énoncé est composé d'un syntagme {Paris, regard, proforme C}, suivi d'un syntagme {université, regard, pointage}. On assigne à chaque syntagme une cible, qui va permettre de contrôler la direction du regard et la position du proforme et du pointage.

- La dernière étape consiste à appliquer la séquence d'unités gestuelles à l'avatar. Ce dernier est créé à partir de logiciels grand public (Poser, 3ds) et est importé dans un module d'animation qui permet de contrôler son animation. La représentation bas niveau fournie au logiciel d'animation correspond typiquement à une partition ou sont notées les valeurs des différents articulateurs et leurs synchronisations éventuelles. Ainsi, l'application peut gérer le contrôle de l'emplacement des proformes et la direction du regard, correspondant aux cibles définies précédemment. Un des points durs concerne le contrôle du regard, qui doit « précéder » ou « être synchroniser » avec le déplacement des mains. Une autre difficulté est de gérer les expressions du visage. Ce dernier point n'est pas traité pour l'instant.

Figure 2 : Exemple de synchronisation du regard avec l'emplacement des proformes

4 Conclusion et perspectives

Nous avons montré dans cet article comment modéliser de manière générique un certain type de relations spatialisées entre entités en LSF, ainsi que la manière d'utiliser ce modèle pour la génération d'énoncés. Nous n'avons pour l'instant intégré qu'une parcelle du formalisme, qui comporte aussi des opérateurs et relateurs permettant de décrire des situations cinématiques et dynamiques.

Dans le cas d'une application de reconnaissance automatique, certains des mécanismes mis-en œuvre dans le cas de la génération peuvent être utilisés. Ils vont typiquement permettre d'aider le système à choisir le bon signe lors plusieurs choix sont possibles (par exemple entre un proforme et un signe stabilisé). Ils vont permettent également, pour un système tel que celui proposé dans (Bossard, 2003), de détecter la présence d'informations sémantiques (en l'occurence une relation spatiale) qui nécessitent un traitement spécifique.

Références

BOSSARD B. (2003), Some issues in Sign Language Processing, Actes de *Gesture Workshop* 2003, LNIA 2915, Springer.

BRAFFORT A. (1996), Reconnaissance et compréhension de gestes, application à la langue des signes. Thèse de l'université d'Orsay, France.

BRAFFORT A. ET LEJEUNE F. (2005), Spatialised semantic relation in French Sign language : Toward a computational modelling. Actes de *Gesture Workshop*, France (à paraître).

CUXAC C. (2000), La langue des signes française ; les voies de l'iconicité. *Faits de Langues* 15/16, Paris, Ophrys.

HUENERFAUTH M. (2004), Spatial Representation of Classifier Predicates for Machine Translation into American Sign Language. Actes de *Workshop on the Representation and Processing of Signed Languages, LREC 2004*, Portugal.

LEJEUNE F. (2004), Analyse sémantico-cognitive d'énoncés en Langue des Signes Française pour une génération automatique de séquences gestuelles. Thèse de l'université d'Orsay, France.

LENSEIGNE B. (2004), Intégration de connaissances linguistiques dans un système de vision, application à l'étude de la langue des signes. Thèse de l'université Toulouse 3, France.

eSIGN, projet européen : http://www.sign-lang.uni-hamburg.de/eSIGN/

Modélisation de l'espace discursif pour l'analyse de la langue des signes

Boris Lenseigne et Patrice Dalle
IRIT-TCI - Université Toulouse 3
118 route de Narbonne 31062 Toulouse cedex 4
lenseign@irit.fr
dalle@irit.fr

Mots-clefs : Espace de signation, monologue, dialogue, analyse de corpus vidéo

Keywords: Signing space, monolog, dialog, video corpora analysis

Résumé Cet article présente un modèle de la structure des énoncés en langue de signes (LS) qui s'articule autour de la notion d'espace de signation. Nous présentons dans un premier temps ce modèle tel qu'il a été conçu pour l'analyse de séquences d'images dans un contexte réduit. Puis nous faisons le bilan des éléments manquants pour étendre ce contexte. Dans une seconde partie, nous présentons une extension de ce modèle aux situations de dialogue et continuons la discussion sur les éléments qui doivent y apparaître dans ce second cas. Cette discussion ouvre la voie à l'introduction de modèles additionnels pour pouvoir analyser la LS dans son ensemble.

Abstract We propose a model for the structure of sign language utterances (SL) that is based on the signing space notion. We first present this model as it was designed for the analysis of reduced-context utterances. After that, we discuss about the elements that lack to generalize that model. In the second part, we present the extension of the model to the representations of dialogs and we continue the discussion about the elements that are missing in the second case. This discussion opens a wide file for the study of additionnal models that will help the complete analysis of the SL.

1 Introduction

Les recherches linguistiques sur la LS ont connu ces dernières années un essor important qui a abouti à des descriptions de cette langue suffisament précises et complètes pour pouvoir envisager leur intégration dans des systèmes informatiques. Notre recherche s'appuie fortement sur ces résultats et vise à proposer des modèles informatiques permettant de rendre compte de la structure des énoncés en LS et de la relier à la façon dont ceux-ci sont réalisés. Ces travaux se situent dans le cadre de la conception de systèmes d'analyse d'images dédiés et constituent une approche originale de cette langue qui trouve des applications dans de nombreux autres domaines (animation d'avatars, études linguistiques, langages gestuels,...).

En effet, dans la grande majorité des systèmes existants, qu'il s'agisse d'interpréter la LS ou de faire signer un personnage de synthèse, l'énoncé est considéré comme une succession de signes isolés, éventuellement coarticulés. En interprétation, les connaissances sur la structure de l'énoncé concernent alors l'ordre dans lequel les signes sont produits. Il peut s'agir de connaissances statistiques (Hienz *et al.*, 1999) (Liang & Ouhyoung, 1996) ou de contraintes sur la structure des énoncés (Starner & Pentland, 1995). La structure spatiale des énoncés reste, en revanche, rarement prise en compte : les systèmes de reconnaissance se limitent le plus souvent à l'interprétation des verbes directionnels (Sagawa *et al.*, 1997). L'interprétation d'énoncés spatialisés ne reste possible qu'à l'aide d'hypothèses fortes sur leur structure (Braffort, 1996). En synthèse, la structure spatiale des énoncés apparaît de façon implicite dans (Lebourque, 1998) par l'introduction de la notion de cible pour un mouvement. Elle est en revanche explicitement représentée dans le système proposé dans (Huenerfauth, 2004). Pour l'heure toutefois, aucun modèle formel de l'espace de signation n'a été proposé et la question de son exploitation pour le traitement automatique de la LS a été peu abordée.

Nous présenterons, ici, un tel modèle élaboré initialement dans un cas mono-locuteur. Nous en commenterons les limites actuelles et présenterons son extension aux situations de dialogue.

2 Modélisation de l'espace discursif dans le cas mono-locuteur

2.1 Modèle symbolique de l'espace de signation

Ce modèle est basé sur la théorie de l'iconicité présentée dans (Cuxac, 2000). Il représente l'agencement des différentes entités de l'énoncé et les relations sémantiques qui les relient. Le type de chaque entité est défini à partir du type des relations susceptibles de la relier aux autres.

Une première version du modèle a été proposée dans (Lenseigne, 2004) pour l'interprétation d'énoncés de requêtes. Il s'agissait donc de ne représenter qu'un sous-ensemble de la LS. Nous abordons l'analyse de l'énoncé par le biais de la construction de l'espace de signation. On accède alors au sens de l'énoncé en considérant les entités évoquées et leur fonction dans l'énoncé à partir de leur agencement dans cet espace. Dans le contexte de requêtes, on distingue quatre types de fonctions : les *localisations temporelles absolues* ou *relatives* au temps courant de l'énoncé, les *localisations spatiales* des différentes entités dans l'espace de signation et la fonction *d'actance* qui peut, le cas échéant, impliquer plusieurs entités. A ce niveau, ne sont représentés que les éléments du discours susceptibles d'être référencés par le signeur et d'être impliqués dans les relations décrites ci-dessus. Ces éléments sont instanciés via une modélisation du comportement qui permet de déduire le type de l'entité à partir des attitudes, mouvements, gestes et expressions du locuteur. D'autres niveaux syntaxiques et lexicaux seront nécessaires pour construire le sens de l'énoncé.

2.2 Représentation interne et externe du modèle

Du point de vue informatique, l'espace de signation (classe `EspaceSignation`[1]) est représenté comme un volume cubique divisé régulièrement en `Emplacements`. Chaque `Emplacement` peut contenir zéro, une ou plusieurs `Entités`. Réciproquement, une `Entité` appartient à

[1] Les termes écrits dans une police "`machine à écrire`" sont des élements du modèle

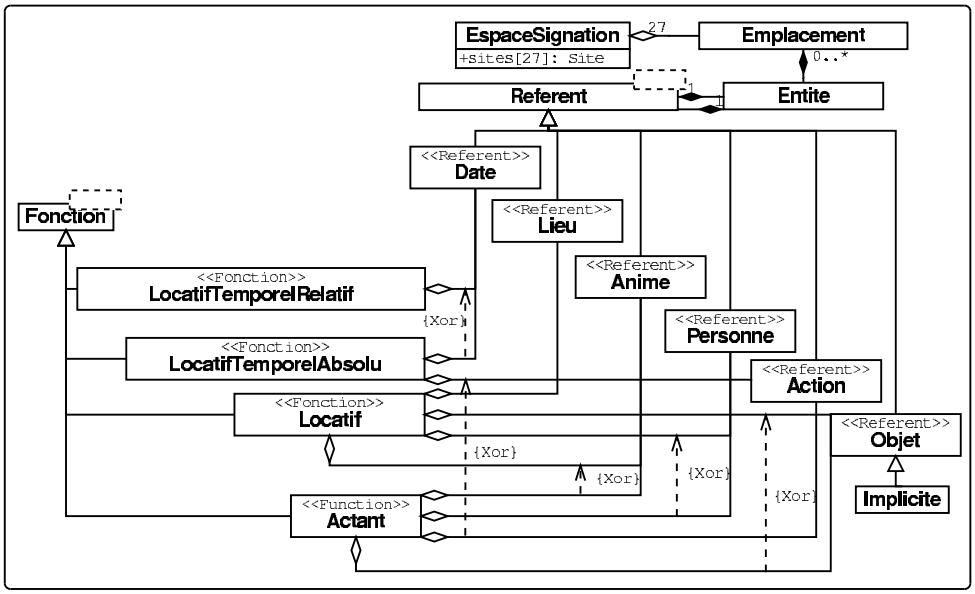


Figure 1: Diagramme de classes UML de la représentation symbolique de l'espace de signation.

un ou plusieurs `Emplacements` et possède un `type` qui est un `Référent`. Les différents `types` d'`Entité` permettent de contraindre les relations susceptibles de les relier, c'est-à-dire les différentes `Fonctions` que l'entité pourra prendre dans l'énoncé (locatif, actant, ...). L'architecture générale du modèle est donnée sous forme de diagramme de classes UML (fig. 1). Cette structure de données est en outre dotée de mécanismes permettant d'assurer la cohérence de la représentation lors de la création d'une `Entité` de type donné. Un outil interactif a été réalisé pour construire et visualiser l'utilisation de l'espace de signation (fig. 2a). Notre objectif est d'automatiser cette construction par analyse d'image.

2.3 Discussion à propos du modèle

Ce modèle ne permet, actuellement, de représenter que les relations temporelles, spatiales et d'actance. Ceci est insuffisant pour représenter l'ensemble des références potentielles

Ainsi les notions qui viennent qualifier une entité n'apparaissent pas initialement dans l'espace de signation. Pourtant, elles peuvent être réifiées et devenir alors référençables. D'autre part ce modèle n'inclut ni les relations de composition, c'est-à-dire le fait de pouvoir référencer l'élément d'un ensemble ni celles de de généralisation/spécialisation. Enfin il ne permet pas les références aux localisations spatiales dans la mesure où ces dernières sont représentées implicitement dans l'agencement des `Entités` dans l'espace de signation.

3 Extension aux situations de dialogue

La LS est presque toujours utilisée en situation de dialogue; les locuteurs partagent alors le même espace discursif et les entités évoquées par un locuteur peuvent être référencées indifféremment par chacun des locuteurs. Cette version du modèle permet de prendre en compte ces interactions.

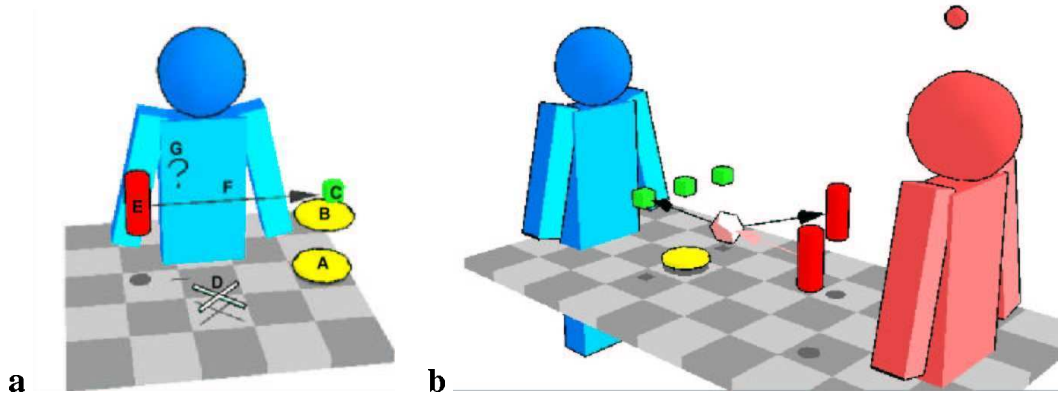


Figure 2: **(a)** Exemple de construction de l'espace de signation visualisé sous forme d'une scène $3D$ pour la phrase (traduite signe à mot): *à Toulouse (A), au cinéma Utopia (B), le film qui passe (C), jeudi 26 février à 9h30 (D), la personne (E), qui a fait le film (F), qui-est-ce ? (G).* **(b)** Exemple de construction de l'espace de signation en situation de dialogue : *(locuteur à gauche) Sur le site,il y a un espace pour le public, un pour l'administration et un pour les salariés . (locuteur à droite) Les salariés peuvent consulter leur espace et ça ne regarde pas le public* (extrait du corpus TALS : CE1.mov).

3.1 Modélisation de l'espace discursif en situation de dialogue

Les mécanismes qui régissent la construction de l'espace de signation en situation de dialogue sont identiques à ceux utilisés dans les monologues à la seule différence qu'une entité peut être évoquée par l'un ou l'autre des locuteurs. La représentation interne d'une `Entité` intègre cette information. En revanche la vérification de la cohérence de la construction de l'espace de signation se fait de façon globale. Nous avons également étendu les possibilités de l'outil interactif pour faire apparaître les deux locuteurs et leur prise de parole respectives. (fig. 2b).

3.2 Discussion

L'utilisation de la LS en situation de dialogue fait apparaître de nouveaux éléments qui ne sont actuellement pas pris en compte dans le modèle. Ces éléments concernent le contenu du discours lorsque les échanges sont au niveau "méta-linguistique", c'est-à-dire lorsque les locuteurs parlent de ce qui est dit, voire de la façon dont c'est dit, ou les éléments de "régulation" (gestes phatiques,...). Du point de vue de la réalisation de l'énoncé, le modèle ne prend pas en compte les mécanismes d'"alias" permettant la reprise, par le second locuteur, d'une entité évoquée par le premier, afin d'éviter une mauvaise interprétation d'un geste déictique.

4 Conclusion

Le modèle que nous avons proposé fournit une description, à un niveau très général, des énoncés en LS, qui peut être exploitée dans de nombreux domaines d'application : en interprétation en génération d'énoncés par exemple.

Son utilisation pour la transcription de corpus vidéo a mis en avant la nécessité de compléter cette représentation pour prendre en compte de nouveaux types de références. Des modèles

supplémentaires doivent être élaborés afin de pouvoir prendre en compte le lexique et la réalisation de l'énoncé. Sur le plan syntaxique, nous avons déjà proposé, pour l'analyse automatique de vidéos, un modèle de comportement permettant de décrire la construction de l'espace de signation en termes de séquences de gestes (Lenseigne, 2004). Pour l'introduction du lexique, nous utilisons, en attendant les résultats de recherches en cours sur les formes graphiques de la LS[2], le formalisme SignWriting[3].

Dans sa version actuelle, le modèle proposé n'en constitue pas moins un outils très intéressant qui ouvre de nombreux questionnements tant sur la linguistique de la LS que sur sa modélisation informatique. Son exploitation en tant qu'outil interactif s'est en outre avérée prometteuse tant pour analyser la LS que pour l'expliquer, voire l'enseigner.

Références

BRAFFORT A. (1996). *Reconnaissance et compréhension de geste, applicationà la langues des signes.* PhD thesis, Université Paris XI, UFR Sciences, LIMSI.

CUXAC C. (2000). *La langue des Signes française. Les voies de l'iconicité.* ISBN 2-7080-0952-4. Paris: Faits de langue, Ophrys.

HIENZ H., BAUER B. & KRAISS K. (1999). Hmm-based continuous sign language recognition using stochastic grammars. In S. BERLIN, Ed., *Lecture Notes in Artificial Intelligence : Procs 3rd Gesture Workshop'99 on Gesture and Sign-Language in Human-Computer Interaction*, p. 165–184, Gif-sur-Yvette, France: A. Braffort, R. Gherbi , S. Gibet , J. Richardson, D. Teil.

HUENERFAUTH M. (2004). Spatial representation of classifier predicates for machine translation into american sign language. In *Workshop on Representation and Processing of Sign Language, 4th Internationnal Conference on Language Ressources and Evaluation (LREC 2004)*, p. 24–31, Lisbon Portugal.

LEBOURQUE T. (1998). *Spécification et génération des gestes naturels; Application à la LSF.* PhD thesis, Université Paris XI-SUD, UFR Sciences, LIMSI.

LENSEIGNE B. (2004). *Intégration de connaissances linguistiques dans un système de vision. Application à l'étude de la langue des Signes.* PhD thesis, Université Paul Sabatier, Toulouse 3.

LIANG R. & OUHYOUNG M. (1996). A sign language recognition system using hidden markov model and context sensitive search. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology*, p. 59–66, Hongkong.

SAGAWA H., TAKEUCHI M. & OHKI M. (1997). Description and recognition methods for sign language based on gesture components. In *in Proceedings of IUI*, Orlando, Florida.

STARNER T. & PENTLAND A. (1995). *Real-Time American Sign Language Recognition From Video Using Hidden Markov Models.* Rapport interne TR-375, M.I.T Media Laboratory Perceptual Computing Section.

[2] http://lsscript.limsi.fr/
[3] http://www.signwriting.org

Pour une iconicité corporelle

Dominique Boutet

LEAPLE (UMR 8606) – Université Evry Val d'Essonne, IUT Brétigny
Château La Fontaine, Chemin de la tuilerie, 91731 Brétigny s/Orge
dboutet@noos.fr

Mots-clés : Iconicité, praxéologie, langue des signes.

Keywords: Iconicity, praxeology, sign language.

Résumé Un point central de la théorie de Christian Cuxac, soit la structuration iconique des langues des signes, permet d'inclure une iconicité qui dépend du corps. Cette iconicité-là est particulièrement productive de sens dans la mesure où des objets dépendent du mode de fabrication et/ou de manipulation. Leur représentation gestuelle dépend par conséquent du corps. Les activités gestuelles de création symbolique et technique placent le corps au centre d'un dispositif cognitif de représentation.

Abstract One of the aims of the Christian Cuxac's theory, namely the iconic structuration of sign languages, allows including an embodied iconicity. This kind of iconicity has specifically a good sense's output. Some objects depend on the modality of manufacture and/or handling's modality. So, their representation in gesture depends on the body. Gesture activities of symbolic and technical creations invest the body in the center of representation's cognitive organisation.

1 Le modèle de structuration iconique comme cadre théorique de prise en compte…

Le modèle de structuration iconique de la LSF de Christian Cuxac (Cuxac, 2000) repose sur une conception unitaire de l'espace pris comme étendue : pour lui, les trois dimensions de l'espace sont utilisées en LSF pour générer, ainsi qu'organiser des espaces sémantisés par les gestes, la mimique, les regards et les déplacements posturaux ; c'est le même espace que celui où l'on évolue, dans lequel par conséquent les objets et les phénomènes se placent ou se déroulent. L'espace tridimensionnel ainsi défini constitue donc un support, autrement dit une étendue sur laquelle s'inscrit du sens donné par la LSF. De plus, la nature de cette étendue permet une simultanéité possible des signes et elle ne semble pas contraindre la mise en forme pour les signes (Cuxac, 1998, 2000, 2003) ; au contraire, cet espace support facilite et démultiplie même certainement la combinatoire à l'œuvre dans la production gestuelle. La matérialité du geste est alors assumée par le corps qui utilise cet espace quadridimensionnel (l'espace plus le temps) à sa disposition.

2 …Du corps comme substrat

En première analyse, on devrait appréhender le corps comme un véritable substrat qui génére des différences ici et là, grâce aux conformations stabilisées de ses segments et naturellement selon une iconicité générale. De plus, l'iconicisation de l'expérience perceptivo-pratique – qui est le processus cognitif à la base des créations de signe – (Cuxac, 1998, 2000) place le corps en position de rejouer des schémas d'actions pour lesquels une iconicité non pas seulement imagique ni seulement diagrammatique mais aussi corporelle structure *a priori* les signes des très nombreux objets artefactuels manipulés (marteau, ouvre-boîte, tournevis, balai…), naturels et manipulés (pomme, sable, pierre, steack…). Ainsi, le corps devrait participer comme substrat au canon de cette iconicité (Eco, 1988). On ne trouve pas de trace de la prise en compte d'une telle iconicité même si le mécanisme général décrit dans la théorie de Christian Cuxac le permet. Je propose d'élargir la seule conception contraignante des articulations, vues comme de simples ajustements facilitateurs de la production gestuelle (Cuxac, 2000, 2004) à une prise en compte praxéologique où la proprioception complèterait la vision comme modalité structurante pour certains signes. L'intérêt d'une telle incorporation du corps dans la théorie de l'iconicité serait limité s'il ne s'agissait que d'une addition ou d'une précision. Je crois qu'elle s'étend de fait bien au-delà : elle modifie la conception de l'espace de signation autant qu'elle permet de préciser un modèle de va et vient entre les modalités visuelle et proprioceptive. Au demeurant, le corps fabrique des objets, il modèle leurs formes et, pour les langues gestuelles, il fabrique du sens.

2.1 Influences des objets artefactuels et/ou manipulés

Les objets artefactuels constituent la liste principale des objets à représenter. Leur mode de production et de manipulation, voire les deux, implique une omniprésence gestuelle souvent multiséculaire, parfois multimillénaire. À ce titre, la tendance vers leur maximum d'efficacité qu'André Leroi-Gourhan (Leroi-Gourhan, 1964) voit dans l'évolution des objets, implique une adaptation fonctionnelle de la forme et même de la matière de l'objet. En conséquence, une adaptation gestuelle s'effectue tant dans le mode de fabrication que dans l'utilisation de

l'objet façonné. Ayant ainsi évolué, certains objets n'ont plus changé de forme depuis longtemps : ils sont stabilisés tout comme leurs modes d'utilisation. Étant donné qu'une part parfois importante de leur forme dépend de la manipulation qu'on en fait, on peut supposer que la structuration de leur représentation gestuelle réside davantage dans leur mode d'utilisation que dans la saisie visuelle de leur forme.

2.2 Iconicité corporelle

Ainsi, pour ces objets l'iconicité met *a priori* plus en jeu des schémas d'action liés à leur manipulation. En l'absence de ces objets, la proprioception constitue une limitation formelle incorporée ([TOURNEVIS] tourne jusqu'au maximum de la supination, [MARTEAU] percute au maximum de l'adduction manuelle, [CAFÉ] par déplacement se dévisse et se revisse dans l'encadrement maximum donné par les extensions/flexions manuelles). La stabilisation réside souvent dans des amplitudes de mouvement contraintes par le ou les degrés de liberté en mouvement étant donné la configuration utilisée et/ou l'emplacement où l'action a généralement lieu (c'est le cas de [COFFRE] ou de [RÉFRIGÉRATEUR] mais aussi [ACCROCHER UN TABLEAU]…). L'utilisation de l'espace dépend ici des rapports que le corps entretient avec le monde et singulièrement des rapports que le corps construit avec le monde à travers des objets manufacturés.

L'espace de signation est alors moins une étendue qu'une mise en scène de formes gestuelles reliées et organisées **autour** d'une absence jamais dessinée, celle de l'objet. L'espace n'est pas étale, il est en creux. Il n'est plus seulement un support neutre mais participe pleinement à la représentation comme substrat y compris pour des portions de l'objet qui ne sont même pas désignées (l'abattant et la charnière du [COFFRE]) : l'espace se densifie à certains endroits. Il répond à une continuité et, pour cet exemple, à celle d'un objet-plan plutôt long, saisissable à deux mains, qui pivote autour d'un axe distant et qui exerce un jeu de contraintes sur les formes gestuelles.

3 …Du corps comme générateur de formes objectales

En même temps, les formes des objets dépendent directement de leurs interactions avec le corps ; (un tournevis présente ces proportions et cet allongement du manche en raison d'une bonne prise en main et d'une force de rotation importante et axiale de la prono-supination). On peut dire que le corps exerce son empreinte jusque sur la forme de tous les outils et de bon nombre d'objets. Il constitue dès lors un des moules phylogénétique essentiel à l'aboutissement des formes que les hommes ont produites, bien plus contraignant en tout cas que tous les embellissements stylistiques très souvent imagiques portées aux productions humaines. Le corps est donc lui aussi un substrat, il génère des formes objectales et symboliques.

En outre, il me semble qu'une prise en compte praxéologique du corps dans la théorie de l'iconicité ré-articulerait la gestualité symbolique avec la praxis et donc la technique, l'autre grande activité anthropologique. Cela permettrait de replacer l'activité cognitive de représentation au centre de la praxis (création d'objets) et de la symbolique (création de sens). Cela éviterait de ne prendre en compte que le seul référent objectal du trébuchet ou de la balance à double plateau pour les signes [BALANCE] ou [PESER], alors que derrière ce

signifiant figure certainement l'origine gestuelle de la pesée et, par la suite, dans la phylogénèse, l'origine de l'objet.

4 …Du corps comme aide au Traitement Automatique

Derrière l'objet BALANCE et son signe, on voit le générateur gestuel contemporain, voire antérieur, à la création du référent *peser*, cela quelle que soit la langue orale ou gestuelle. Il en va de même pour *percuter, frapper, gratter, couper, écraser, séparer*… L'iconicité est bien corporelle puisque le référent, l'ancrage, la primitive sont corporels. Ces schémas d'action, générateurs de nombreux objets aux lignées connues, sont-ils productifs de concepts dont on pourrait suivre aussi le lignage ? On part ainsi d'un substrat continu, le corps, qui permet de tracer des filiations gestuelles. Ce n'est pas le cas pour les signes non objectaux ou pour ceux dont le signifiant reprend la forme du référent ; leur structuration est telle que le corps n'est déjà plus qu'un support.

Les schémas d'action qui correspondent à une manipulation d'objets naturels et/ou à une empreinte dialectique exercée entre les objets et la gestuelle, mènent à penser que l'orientation, la configuration, l'emplacement, le mouvement ne répondent pas tous ensemble à une structuration de type morphémique. Une approche paramétrique n'est pas opérante pour ces signes.

Il ne semble pas légitime pour le signe [COFFRE] de distinguer la saisie, c'est-à-dire la configuration, l'orientation de la main, son emplacement et le type de courbe que provoquent les mouvements de l'avant-bras et du bras. Ce signe forme un tout sans lequel on ne peut voir à quoi il réfère. Il s'agit bien ici d'un signe qui s'organise autour de l'absence d'un objet. Il en va différemment d'autres signes qui, eux, s'organisent autour de paramètres. Dans une première approche, ces signes-là marquent la présence d'un objet [CISEAUX] ou d'une entité [PERSONNE], [ARBRE], [VACHE] non plus autour de celle-ci mais par celle-ci, même de manière partielle. Cet état de fait reste à vérifier pour l'ensemble du lexique standard.

Références

CUXAC, C. (1998), Constructions de références en Langue des Signes Française, *Sémiotiques*, Vol. 15, pp.85-105.

CUXAC, C. (2000), La Langue des Signe Française, les voies de l'iconicité, *Faits de Langues*, Vol. 15-16, pp.391.

CUXAC, C. (2003), Iconicité des langues des signes : mode d'emploi, *Cahiers de linguistique analogique*, Vol. 1, pp.237-263.

CUXAC, C. (2004), Phonétique de la LSF : une formalisation problématique, *Silexicales*, Vol. 4, pp.93-113.

ECO, U (1988), *Le signe*, Bruxelles, Labor.

LEROI-GOURHAN, A (1964), *Le geste et la parole*, Paris, Albin Michel.

Construction/déconstruction de l'espace de signation

Annie Risler

UMR SILEX – Université Lille3
Pont de Bois BP 169 59653 Villeneuve d'Ascq Cedex
annie.risler@univ-lille3.fr

Mots-clés : langue des signes, syntaxe, espace
Keywords: sign language, syntax, space

Résumé

Nous proposons d'envisager les constructions spatiales selon trois niveaux : lexical, syntaxique et textuel. Ces trois modes d'utilisation de l'espace se combinent pour former un texte narratif. Nous mettrons en évidence les paramètres formels qui permettent de savoir quel type d'espace est réalisé par un mouvement du signeur.

Abstract

Spatial construction in FSL has different effects depending on 3 levels : lexical, syntactic and textual. These three specific uses of space combine to create a narrative text. We'll put forward the different morphological parameters that enable one to identify the kind of space that a signing movement stands for.

1 Objectif

Nous proposons d'analyser la syntaxe de la langue des signes à partir de la structuration de l'espace de signation. En cela, nous persistons dans la voie qui, depuis 1998, nous amène à rechercher la trace des opérations cognitives dans le tracé des énoncés en LSF, en mettant en parallèle les mouvements réalisés par le signeur et la formalisation spatiale du sens produit. (à partir du modèle de Desclés essentiellement). Nous mettons en avant l'iconicité structurelle de la LSF, comme Cuxac ou Sallandre, mais nous la plaçons à un niveau cognitif (entre les représentations langagières et la forme en langue) et pas référentiel (entre la situation dénotée et la forme en langue). En partant de l'opposition entre signes lexicaux et signes relateurs (Risler 1998), nous avons été amenée à isoler d'abord deux types d'espace : l'espace des signes lexicaux (construction d'images) et l'espace syntaxique (construction de relations syntaxiques de nature spatiale et formelle par le mouvement et la forme des signes relateurs)

(Risler 2003). Nous avons alors proposé une décomposition des signes relateurs selon leur paramètres formels, spatiaux et cinématiques, qui ont chacun une valeur morphosyntaxique (Risler-Lejeune 2004)[1]. La poursuite de nos investigations nous amène maintenant à rajouter un troisième type d'espace : les sous-espaces textuels ou sous-parties, constitués par une organisation topologique de l'espace entre le signeur et des emplacements pertinisés, ou loci, réalisés par le buste, le mouvement des mains et le regard. Ce troisième niveau est nécessaire pour affiner le rôle syntaxique joué par l'espace et faire la part entre la syntaxe verbale (relations spatiales construites par le mouvement de chaque signe relateur) et les renvois anaphoriques et interphrastiques (pointages de loci par le regard ou l'index).

Nous commenterons en direct, un texte narratif de récit de vie, tiré du corpus LS-Colin. Nous mettrons ainsi en évidence les mécanismes de construction de l'espace. Ceci permettra de montrer que :

- Le texte se décompose en sous-parties ayant une cohérence spatiale, qui correspondent à des paragraphes. Ces sous parties s'organisent à chaque fois autour du buste du signeur et deux loci (qui seront notés loc1, loc2… loc N). Entre ces trois emplacements se répartissent les tracés des signes processifs.

- L'espace de signation n'est pas un espace scénique qui se remplit progressivement. C'est le lieu d'expression d'espaces phrastiques successifs. L'orientation des mouvements des procès répond à une logique syntaxique plus que topologique. Un même référent pourra donc être localisé en différents loci, en fonction de l'organisation spatiale de chaque sous-partie. Des séquences de signes lexicaux réalisés en espace neutre servent de transition ou d'introduction aux séquences de signes spatialisés.

- Certains loci servent de lien entre les sous parties.

- Un locus est le lieu d'une entité ou d'un événement.

2 Présentation du texte

Dans ce récit de vie, catalogué 11 septNIC1, le locuteur raconte que le 11 septembre, il s'était rendu dans un magasin de voitures. Il a trouvé tous les employés massés devant un poste de télévision, et il a eu du mal à obtenir des renseignements. De plus, le vendeur n'avait de cesse de retourner voir les images diffusées, auxquelles lui-même n'arrivait pas à donner un sens cohérent. De retour chez lui, il a branché le poste et a alors constaté que c'était toujours les mêmes images. Il a dû attendre le journal sous-titré du soir pour avoir la clé de l'histoire. Le lendemain, il a appelé un ami pour partager sa stupéfaction.

Ce texte se décompose en 4 sous-espaces textuels construits, introduits par des séquences de signes lexicaux en espace neutre.

---

[1] Ceci nous a amenée à abandonner la terminologie de Cuxac (transferts), qui ne rend pas compte de cette décomposition et à adopter une terminologie de niveau morpho-syntaxique explicitée plus loin..

1. Espace du magasin : la voiture par rapport à la télévision.

L'espace du magasin est d'abord rempli par des proformes[2] de voitures, puis il va s'organiser autour de deux loci : la télévision devant laquelle sont groupés tous les employés (loc1, devant le signeur, légèrement à droite), et une voiture repérée plus particulièrement (loc2, sur sa gauche).

Le signeur va localiser par son buste, en proforme corporelle, soit lui-même au moment des faits rapportés (le client), soit le vendeur. On aura à l'intérieur de ce sous-espace textuel une séquence en espace dialogique, entre le client et le vendeur. Chaque proforme corporelle est alors identifiée par rapport aux loci regardés.

2. Espace du magasin : aller vers une voiture en particulier et retourner vers le poste

Autour du loc2 se construit un deuxième sous-espace, dialogique, entre le vendeur et le client, regard du client tourné complètement à droite. Cette séquence se conclut par une trajectoire vers la gauche de la proforme manuelle qui réfère au corps du vendeur, suivie du regard par le signeur qui incarne alors le client.

Le point d'arrivée de la proforme manuelle du vendeur va constituer le loc3, celui de la télévision dans le magasin. C'est donc le loc1 déplacé. On retrouve la signification de ce locus par le fait qu'il serve de repère à l'attroupement déjà signé précédemment ; bien que ne se situant pas au même endroit dans l'espace de signation.

Pourquoi un tel changement ? La relation prédicative de déplacement conjoint se fait latéralement et non frontalement, les points de départ et d'arrivée des proformes manuelles s'opposent latéralement. Il fallait donc déplacer le repère. Cette réorganisation de l'espace est syntaxiquement pertinente, elle correspond aux espaces pré-sémantisés décrits par A. Millet (Millet, 1997).

3. espace de la maison : le téléviseur, et référence aux images vues dans le magasin

Après transition, il crée un loc4 devant lui, sur sa gauche, par le regard et l'emplacement de signes lexicaux. Il y place le signe [TELEVISION] et situe son corps en tant que spectateur, dans un rapport visuel avec ce locus. Dans ce sous-espace seront signées deux périodes temporelles, gardant les mêmes emplacements, mais situées de part et d'autre d'un signe de coupure du déroulement du temps.

Puis par le regard il va faire le lien entre loc3 et loc4, les regardant alternativement, furtivement, pendant qu'il signe le commentaire : [MEME - FILM]. Ce qui indique que ce qui a été vu sur le premier téléviseur est identique à ce qu'il vient de voir chez lui. La référence d'identité renvoie non pas à l'entité localisée (la télévision, mais aux événements perçus localisés dans ces loci).

4. espace de la discussion avec son ami : son interlocuteur, le poste sur lequel il a vu les nouvelles.

---

[2] Nous appelons proforme (au féminin) une configuration anaphorique placée ou déplacée entrant dans la constitution d'un signe relateur. Ce terme est repris des travaux de Engberg-Pederson 1989. Il y a des proformes manuelles et des proformes corporelles qui remplissent une fonction pronominale dans la construction verbale. Un signe relateur peut ainsi comporter jusqu'à 3 proformes simultanées (de la main G, de la main D, du buste).

Il crée un loc5 à sa droite, par le regard et la direction de ses signes, où est localisé son interlocuteur. Puis pour évoquer son propos (ce qu'il a vu la veille à la télévision), il va pointer l'index vers le loc4 tout en maintenant le regard fixé sur le loc5 . Le texte se termine par un regard sur l'interlocuteur.

3 Les paramètres de la construction de l'espace

Le regard est le premier indicateur du type d'espace construit : espace lexical, espace syntaxique, espace textuel... Mais la direction du regard concourt à la valeur de l'espace en relation avec l'emplacement des mains :

- les mains en espace neutre, le regard sur l'interlocuteur : espace lexical, signes lexicaux de transition ou d'explicitation non placés, valeur énonciative du regard.
- les mains en locus (espace non neutre), appuyés par le regard sur le locus : espace syntaxique, signes relateurs ou signes lexicaux dirigés ou placés.
- les mains en espace neutre, le regard sur un locus : espace textuel, par localisation d'entités ou d'événements en ce locus.

On note aussi des relations entre les valeurs anaphoriques des mains et du buste : soit les mains du signeur réfèrent aux mains du personnage dont il prend la proforme corporelle, elles sont incluses dans la proforme corporelle ; soit ses mains sont proformes d'autres éléments que le buste (comme par exemple dans : *il regarde les employés massés devant la télévision* : où son buste est proforme du client alors que ses mains sont proformes des personnes devant la télévision qui a été localisée précédemment au loc1) ; soit encore main et buste réfèrent parallèlement au même individu (comme dans : *il vient avec moi* : où le signeur incarne le client avec un mouvement du buste vers le côté, doublé d'une proforme manuelle qui réfère aussi au corps du client qui se déplace en même temps que la proforme de l'autre main qui réfère au vendeur .

Il est donc fondamental d'isoler ces paramètres de regard, main, buste, mais aussi de les mettre en perspective les uns avec les autres. La grille de transcription utilisée (Millet, Bras, Risler, 2002) met en évidence la différence entre signes lexicaux et relateurs et fait apparaître la création des loci, ceci par le jeu des différentes lignes de la partition et du choix des items notés :

- la valeur du buste, proforme ou énonciateur
- la direction et la qualité du regard
- la valeur formelle des mains : proforme manuelle, proforme corporelle, configuration conventionnelle, pointage
- la valeur cinématique des mains : cinématique : mouvement interne ou trajectoire
- la valeur spatiale des mains : arrangements entre mains et en fonction des loci
- la mimique.

Cette liste d'items, pensons-nous, devrait être exploitable par du traitement d'image. Nous pensons en effet pouvoir déterminer des combinaisons morpho-syntaxiques de paramètres formels qui concourent à la construction du sens.

Références

CUXAC C. (2002) La LSF, les voies de l'iconicité, *Faits de Langue* n°15-16, Ophrys

DESCLES JP. (1990) *Langages applicatifs, langues naturelles et cognition*, Hermes

ENGBERG – PEDERSON E. (1989) Proformes en morphologie, syntaxe et discours, in *Etudes européennes en langue des signes*, Irsa, Bruxelles

MILLET A. (1997) Statut du mouvement dans les langues gestuelles, in *Langues gestuelles, quels enjeux pour les sourds ?, Lidil* n°15, Grenoble

MILLET A., BRAS G., RISLER A. (2002) Projet *Emergence, rapport intermédiaire*, Grenoble

RISLER A. (1998) L'iconicité en Langue des Signes et les procédés d'imagerie à la base de la définition notionnelle des catégories grammaticales de nom et verbe, in *Cahiers de C.I.S.L* n°13, Toulouse, p. 121-136

RISLER (2003) Point de vue cognitiviste sur les espaces créés en LSF : espace lexical, espace syntaxique , in Lidil n°26, Grenoble

RISLER A., LEJEUNE F. (2004) Traces des opérations langagières et des représentations sémantico-cognitives dans la forme verbale en LSF, in *Silexicales 4*, Lille

SALLANDRE M-A. (2003) Le rôle de l'espace dans l'émergence de l'iconicité (imagique et diagrammatique) en LSF, *Colloque Espace, ENS* Paris, février 2003

Annexe : extrait de la grille de transcription.

translitération	*Il me donne qqchose*	*Il repart vite*	*Je m'interroge*	*écran*	*groupe*	*Ils regardent X*	*Je regarde X et m'interroge*
Buste	Pr client	Pr client	Pr client		Pr client Vers loc3	Pr vendeur	Pr client Recul puis se tourne vers int
M Gauche	Main vendeur	Pr vendeur		index	Pr vendeurs	Pr vendeurs	
A Droite	Main vendeur		???conv	index	Pr vendeurs	Pr vendeurs	??? conv
I Mouvt	Vers buste	Trajectoire G	Interne G	tracé			
N Espace	Depuis loc2	De loc2 à loc 3		En loc3	loc 3	loc3	
Regard	Sur mains——main G--						
		De loc2à loc3—sur loc3-———————————————int,					sur loc3, int——
Mimique			interrogative				interrogative

Verbes et actants en Langues des Signes Française

Loïc Kervajan (1), Emilie Guimier De Neef (2), Jean Véronis (3)

(1, 3) DELIC – Université de Provence
29, avenue Robert Schuman 13621 Aix-en-Provence Cedex 1 France
loickervajan@wanadoo.fr, jean.veronis@up.univ-mrs.fr
(2) Tech/Easy/LN – France Télécom Division R&D Technopole Anticipa
2, avenue Pierre Marzin 22307 Lannion Cedex
emilie.guimierdeneef@rd.francetelecom.com

Mots-clés : Langue des Signes Française, typologie des verbes, classes nominales, accord verbes-actants.

Keywords : French Signs Language, verb typology, nominal classes, verb-actors agreement.

Résumé Cet article s'intéresse à la relation entre les verbes et les noms en LSF. L'approche proposée provient des résultats d'une expérience conduite au laboratoire "Langues Naturelles", Division Recherche et Développement (DRD/tech/easy/LN) de France Télécom au sein duquel a été développé l'analyseur et générateur syntaxique, TiLT (**T**raitement **Li**nguistique de **T**extes). Une approche de l'accord entre les verbes et leurs actants sera présentée à travers un exemple après une description des objets particuliers à partir desquels se construisent les verbes.

Abstract In this paper, we propose an approach to represent the relations between verbs and nouns in FSL. This proposition comes from the results of an experiment conducted in the "Langues Naturelles" laboratory, Division R.&D. (DRD/tech/easy/LN) of France Télécom using TiLT (**Li**nguistic **T**reatment of **T**exts), an automatic syntactical analyser and generator. An approach of agreement between verbs and actors will be presented through one example after a description of the particular objects from the which verbs are constructed.

1 Introduction

La présente communication est le résultat d'un travail mené à l'occasion d'un stage de DESS, point de départ d'une thèse prévue pour 2007. L'objectif du stage était la modélisation du lexique de la Langue des Signes Française (LSF) sur le modèle de ce qui se fait en traitement automatique des langues orales. Le travail de modélisation nous a demandé d'approfondir certains mécanismes autour de la relation entre le verbe et ses actants. Une typologie des verbes s'en est dégagée et a permis de prototyper une mini-grammaire de la LSF.

2 Concepts élémentaires

Au-delà des objets habituels décrits par les grammaires traditionnelles – noms, verbes, adjectifs, etc. – nous pouvons remarquer trois objets particuliers utilisés par les langues des signes (LS) : les classificateurs, les loci et les objets non-discontinus. Même si les deux premiers semblent nouveaux, leur expression se retrouve facilement dans les langues orales (LO). Le seul point qui les rend spécifiques est leur manière d'utiliser l'espace pour organiser les différents éléments du discours. En effet, c'est davantage l'espace physique comme support des relations syntaxiques et sémantiques qui émerge comme étant spécifique aux LS.
Le classificateur a longtemps été présenté comme la principale caractéristique de la LSF (Moody, 1983). Il est désormais identifié comme un simple pronom, mais à l'instar de Cuxac (Cuxac, 2000) nous préférons le terme de ***proforme*** en référence aux paramètres de configuration manuelle utilisés pour l'élaboration des structures de grande iconicité (SGI).
Les loci (Vercaigne, Pinsonneault, 1996) sont des parties réservées de l'espace. Le locus, véritable objet linguistique, doit être intégré dans les règles syntaxiques. Il permet d'articuler certains actants autour du verbe Il peut être conventionnel comme le pronom personnel ou contextuel et être rattaché à son référentiel par pointage au moment de la construction de l'espace grammatical.
Les objets non-discontinus apparaissent comme particulièrement inhabituels en linguistique. En effet, alors que les LO décrivent le monde d'une façon catégorielle, les LS peuvent quant à elles participer d'une vision continue du monde par le biais d'un nombre (potentiellement) infini d'acteurs, par le biais d'une manière de qualifier – adjectifs[1], adverbes – ou encore par celui des SGI qui appartiennent à un continuum descriptif (Cuxac, 2000). Ces trois modalités ne sont pas forcément indépendantes et peuvent se recouper puisqu'il est possible d'utiliser les SGI pour qualifier.
Par ailleurs, de la même manière qu'en swahili, les LS présentent une organisation des noms en classes nominales : êtres humains, animaux, objets préhensibles, objets animés… A chaque classe correspond un certain nombre de proformes, véritables pronoms infixés sur les verbes avec lesquels les noms sont en interaction. Le nom et sa proforme ne se confondent pas. Il s'agit bien de deux gestes distincts. Mais une fois la référence de la proforme connue, cette dernière devient autonome.

3 Typologie des verbes

L'usage et la littérature nous parlent des verbes directionnels[2] ou non directionnels. Cette simple distinction s'est révélée insuffisante dès que nous avons eu à décrire la langue d'un point de vu micro-syntaxique pour son implémentation. C'est ainsi que s'est mise en place une typologie dont la codification s'inscrit dans une perspective de modélisation pour le TAL. Chaque verbe est ainsi associé à un code morphologique composé d'une part de "V_" pour signifier la catégorie de verbe et d'autre part de quatre lettres représentant quatre actants potentiels : (1) le sujet, (2) l'objet1, (3) l'objet2 et (4) le lieu. L'objet1 et l'objet2 pourraient correspondre respectivement au COI et COD de la grammaire traditionnelle du français. Mais nous avons préféré nous dégager de ces notions pour rester ouvert à une approche propre à la micro-syntaxe de la LSF. Par exemple, « je te donne quelque chose » est en français très

[1] Il s'agit des qualificatifs de forme en opposition aux adjectifs de couleurs, par exemple, qui appartiennent au discontinu.

[2] "Directionnel" est d'ailleurs un terme imprécis dans la mesure où il n'indique pas sur quel objet se construit la directionnalité : sur un lieu (partir à Paris) ? Sur une personne (téléphoner à quelqu'un) ? Sur le sujet et l'objet1 (suivre quelqu'un) ?

différent de « je vais de Marseille à Paris » alors qu'en LSF les deux verbes donner et aller se construisent physiquement de la même façon : un locus de départ et un locus d'arrivée reliés par une configuration manuelle posée sur un mouvement.
Chacun des actants peut générer un accord en : **L** pour locus, **P** pour proforme en **B** pour locus et proforme en simultané (bi), **X** pour l'invariant. Chacune de ces lettres occupe, derrière le "V_", le rang de l'actant qu'elle représente. Ainsi, la typologie élaborée nous permet-elle de distinguer 15 types de verbes, dont par exemple :

[dire] V_LLXX : accord en loci du sujet et de l'objet1
[suivre] V_BBXX : accord en proformes – humains – et en leurs loci respectifs.
[poser] V_XXPL : accord en proforme de l'objet à poser et en locus du lieu de pose.

Cette typologie demande à être affinée et confrontée à un corpus ainsi qu'à la compréhension de la part des locuteurs sourds. De même, elle devra intégrer d'une manière pertinente les observations des auteurs ayant déjà proposé de différencier les verbes en fonction de leurs réalisations (Lejeune, 2004), (Lejeune, Risler, 2004).

4 Relation entre le nom et le verbe

Afin d'intégrer la notion d'accord entre les noms et les verbes, nous proposons une représentation basée sur la flexion des verbes en fonction des proformes nominales et des loci. Prenons pour exemple « je te donne un verre ». Nous avons besoin des deux entrées lexicales de « verre » et de « donner », le reste appartenant à la grammaire :
Entrée lexicale pour « verre » :

[verre]{c, boo…}
où [verre] est le lemme et {c, boo…} les proformes qui peuvent être utilisées pour représenter le lemme.

Entrée lexicale pour « donner » :

[donner]-{je, tu, il, il2nd, ilP^{teur} }-{ je, tu, il, il2nd, ilP^{teur} }-{c, o, bco, boo, pince… }[1]
où [donner] est le lemme – ou le radical – et les parenthèses représentent respectivement la combinatoire possible entre :
- les loci possibles de début de réalisation du verbe pour le sujet : préfixe verbal
- les loci possibles de fin de réalisation du verbe pour l'objet1 : suffixe verbal
- les configurations admises par le verbe pour l'objet2 : infixe verbal.

L'accord entre le nom et le verbe devient là une évidence. Pour produire une phrase complète, il suffit d'antéposer un objet2 ayant une proforme commune avec le verbe, "c" pour notre exemple :

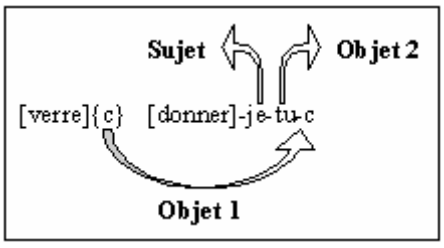


Figure 1 : exemple d'accord pour le verbe [donner]

Le nom est signé, puis sa proforme ("c") est utilisée en tant qu'objet indépendant pour achever la construction du verbe positionné sur ses loci ("je" et "tu"). La flexion verbale obtenue est : [donner]-je-tu-c.

[1] "c" et "o" sont les lettres dactylologiques, "bco" et "boo" sont respectivement les configurations "bec_de_canard_ouvert" et "bec_d_oiseau_ouvert,"

5 Conclusion

L'ordre syntaxique de la « phrase » en LSF ne se réduit pas à un enchaînement canonique de type Temps-Lieu-Quoi-Qui-Verbe. Les verbes invariants donneront plutôt un ordre SVO – rêver : [je] [rêver] *quoi*[1]. D'autres obligeront une antéposition de l'objet2 pour se construire (OSV) – arroser : [plante](locusA) [je][arroser](locusA). Une fois cette observation faite, il s'agit de mettre en place les moyens de la reconnaissance ou de la génération d'un certain type d'énoncés. Et, c'est dans le but d'élaborer une grammaire automatique que nous avons procédé à l'établissement de notre typologie des verbes à partir de laquelle TiLT peut produire ses premiers arbres de dépendance grammaticale en LSF.
Rappelons que le but poursuivi n'est pas la reproduction à l'identique de ce que nous propose la LSF mais la manipulation d'énoncés artificiels intelligibles par les sourds. De même que le français écrit ne reproduit pas toutes les possibilités de l'oral, nous admettrons qu'un système automatique manipulant la LSF le fasse de manière imparfaite, en se situant pour ses débuts encore bien loin de l'approche iconiciste (Cuxac, 2000) et des applications techniques qui en découlent (Risler, 1998). De même, les multiples variations des relations possibles entre agents et patients ne sont pas encore prises en compte, attendu qu'en terme de génération d'énoncé il est envisageable d'effectuer un choix parmi les possibles, le tout étant que le message soit compris. Pour autant, nous chercherons, au cours de la thèse, à sortir du cadre strict de micro-syntaxe en intégrant les aspects de linguistique cognitive développés par Cuxac et repris par Lejeune (Lejeune, 2004), notamment en ce qui concerne l'approche globale du geste au niveau du rapport sens/forme et du regard. Il faudra également se rapprocher des travaux sur l'espace de signation (Lenseigne, 2004) pour identifier les limites de responsabilité pour la gestion des référentiels : entre le module grammatical et celui de la mise en forme – avatar –, lequel décide de l'appropriation et du renouvellement de l'espace ?

Références

CUXAC C. (2000), La Langue des Signes Française, Les Voies de l'Iconicité, *Faits de Langues*, Vol 15-16, Paris, Ophrys.

LEJEUNE F., (2004), *Analyse sémantico-cognitive d'énoncés en Langue des Signes Française pour une génération automatique de séquences gestuelles*, thèse de doctorat, Paris-XI.

LEJEUNE F., RISLER A. (2004), Traces des opérations langagières et des représentations sémantico-cognitives dans la forme verbale en LSF, Actes des *Journées d'études sur la LSF* : La linguistique de la LSF : recheches actuelles, Silexicales.

LENSEIGNE B. (2004), *Intégration de connaissances linguistiques dans un système de vision. Application à l'étude de la langue des Signes*, thèse de doctorat, IRIT-UPS, Toulouse.

MOODY B. (1983), *Dictionnaire de LSF, Tome 1 : « Introduction à l'histoire et à la grammaire de la langue des signes »*, Vincennes, Ellipse - IVT.

RISLER A., (1998), L'iconicité en langue des signes et les procédés d'imagerie à la base de la définition notionnelle des catégories grammaticales de nom et verbe, *Cahiers du Centre interdisciplinaire des Sciences du langage*, Vol. 13, pp. 121-135.

[1] Nous avons mis, comme habituellement, les gestes précis entre crochets, « quoi » est en italique parce qu'il ne s'agit pas là du geste « [quoi] » mais de ce dont je rêve.

VERCAINGNE-MÉNARD A., PINSONNEAULT D. (1996), L'établissement de la référence en LSQ : les loci spatiaux et digitaux, in C. Dubuisson et D. Bouchard (dir.) : Spécificités de la recherche linguistique sur les langues signées, *Les cahiers scientifiques de l'Acfas*, Vol. 89, pp. 61-74, Montréal.

Problèmes et méthodes pour l'analyse d'énoncés en LSF

Balvet Antonio (1), Sallandre Marie Anne (2)

(1) UMR 8528 Silex – Université Lille 3
Université Lille 3, BP 60149 - 59653 Villeneuve d'Ascq Cedex
antonio.balvet@univ-lille3.fr

(2) UMR 7023 SFL – Université Paris 8
15 rue Catulienne, 93200 Saint-Denis
sallandre@yahoo.com

Mots-clés : LSF, transcription, traitement automatique des langues, linguistique de corpus
Keywords: LSF, transcription, analysis, natural language processing, corpus linguistics

Résumé Après avoir rappelé les principales caractéristiques de la construction des énoncés en LSF, nous abordons les questions liées à leur traitement automatique. Nous tentons ainsi de préciser dans quelle mesure les outils et méthodes aujourd'hui disponibles pour la modélisation et l'analyse automatique des langues naturelles apparaissent compatibles avec le cadre théorique et méthodologique ici adopté, centré sur les Structures de Grande Iconicité.

Abstract In this paper, we address the issue of the formal treatment of French Sign Language, in a functionalist and enunciativist framework. We advocate for a copus-driven approach and we emphasize the need for alternative frameworks for NLP, alongside classical generativist symbolic-computational ones.

1 Introduction

La LSF pose de nombreux défis tant pour sa description que pour sa modélisation, y compris par des moyens automatiques. En effet, cette langue construit l'espace comme un système linguistique, ce qui implique de passer par des transcriptions réalisées par des linguistes disposant d'une bonne connaissance de la langue. Du point de vue descriptif, l'approche de type énonciative et cognitive, retenue par (Cuxac, 2000) et (Sallandre, 2003), qui constituera la base des travaux ici exposés, semble difficilement conciliable avec les approches classiques en TALN, centrées sur l'élaboration de systèmes de règles formelles. En effet, le fonctionnement synthétique[1] et la nécessité d'intégrer le contexte élargi[2], mis en évidence par

[1] Intégration de plusieurs paramètres : signes manuels, mimiques faciales, posture du corps.

[2] Co-texte, mais également situation d'énonciation, attentes des interlocuteurs.

ces travaux, apparaissent comme une source de difficultés majeures pour toute tentative d'analyse automatique des énoncés en LSF par le biais de règles symboliques. Dans cet exposé, nous examinons, tout d'abord, les conditions d'un traitement formel des énoncés en LSF. Ensuite, nous présentons quelques outils et résultats d'exploration de corpus de transcription, afin d'en faire émerger des régularités au niveau syntagmatique.

2 Exploration de corpus de transcriptions en LSF pour le TALS

2.1 Quelques propriétés saillantes de la construction des énoncés en LSF

Le modèle proposé par (Cuxac, 2000) se situe dans une perspective sémiogénétique et considère l'iconicité (référentielle) non comme seul outil de description de la langue mais comme principe organisateur. Une bifurcation fonctionnelle a ainsi été postulée et détermine deux pôles entre lesquels le va-et-vient est constant. Elle se compose d'une part des structures de grande iconicité (SGI), qui *donne à voir tout en disant*, d'autre part des signes standard (SS), sans visée illustrative, qui *disent* seulement. Les SGI se structurent à partir des trois principaux transferts : transferts de taille et de forme, transferts de situation et transferts de personne. Ces SGI, contrairement aux signes standard[3], ont peu fait l'objet de descriptions formelles et encore moins de modélisations.

Le modèle de Cuxac a été confronté à un corpus de référence : LS-COLIN[4], présenté dans (Sallandre, 2003), constitué de trois discours de genres variés, sur une durée totale d'une heure et cinq minutes. L'analyse de ce corpus montre que plus des 2/3 des unités sont effectués avec une visée illustrative dans les deux narrations (3/4 pour le premier récit, 2/3 pour le deuxième) et environ 1/3 dans le genre explicatif. En raison de son importance, de la diversité des genres qui le composent et du nombre de locuteurs enregistrés, ce corpus constituera la base des observations et réflexions consignées dans le présent exposé, centré par conséquent sur les SGI plus que sur les SS.

3 Quels fondements méthodologiques et théoriques pour le TALS ?

Les questions habituellement traitées en TALN sont centrées sur les combinaisons possibles et impossibles de suites de symboles (mots, constituants), déterminées par l'application systématique de règles logiques. Or, le cadre ici choisi pour traiter des énoncés en LSF met l'accent sur des paramètres énonciatifs, ainsi que sur l'importance du contexte, tant dans l'exercice de la compétence linguistique des locuteurs que dans la pratique de transcription. Une accommodation est donc nécessaire entre les présupposés théoriques sous-jacents au TALN d'inspiration générativiste et le modèle de Cuxac.

---

[3] Voir (Lejeune, 2004) pour une analyse d'énoncés utilisant la Grammaire Applicative et Cognitive de Desclés.

[4] Action Cognitique 2000, LACO 39 (Université Paris 8, LIMSI-CNRS, IRIT), 39 discours, 13 locuteurs adultes.

3.1 Quelques difficultés pratiques et théoriques pour le TALS

Le degré de grammaticalité des phrases constitue habituellement le fondement de toute description formelle en TALN. Or, en LSF, la notion syntaxique de phrase, ainsi que celle de mot (typographique), paraissent peu pertinentes. Reste la question des conditions de bonne formation des énoncés, pour une langue dont la syntaxe semble guidée essentiellement par des principes d'optimisation de contraintes cognitives[5]. Il convient donc de déterminer à quelles conditions un énoncé sera refusé par un locuteur natif de la LSF. La méthode traditionnelle d'investigation en la matière, basée essentiellement sur l'introspection et l'intuition linguistique du chercheur, apparaît ici insuffisante, étant donné l'état actuel des connaissances. Un début de solution apparaît passer l'étude de corpus d'énoncés attestés, tel que LS-COLIN.

Par ailleurs, en TALN, l'un des prérequis de toute modélisation est la caractérisation des énoncés, dans les termes des grammaires formelles définies par N. Chomsky, caractérisées par deux ensembles d'éléments (auxquels s'ajoutent les règles de réécriture) : le Vocabulaire Non Terminal (i.e. catégories syntaxiques) constitué d'un nombre fini d'éléments, et le Vocabulaire Terminal (i.e. unités lexicales) constitué d'un nombre potentiellement infini d'éléments. Une des premières questions à résoudre pour la LSF serait donc celle du VNT. En effet, en LSF, l'identification des unités n'est souvent possible qu'en contexte, et en adoptant un point de vue fonctionnel. De ce fait, il semble difficile de proposer un système de conditions nécessaires et suffisantes pour identifier les éléments du VNT, en-dehors des catégories de grande iconicité dont une taxinomie est fournie dans (Sallandre, 2003). Ainsi, dans l'approche adoptée ici, il est possible que les éléments candidats pour un VNT de la LSF ne constituent pas une liste fermée. De plus, dans le cadre adopté ici, les problèmes centraux du sens, du contexte et de la multilinéarité, restent encore insuffisamment formalisés en TALN.

Ceci nous incite à aborder le problème du TALS d'abord par le biais des contraintes de combinaison des unités appartenant aux deux grands pôles de la bifurcation posée par Cuxac : SS et SGI. La caractérisation formelle des énoncés en LSF, par exemple sous la forme de grammaires hors contexte probabilistes, nous paraît pouvoir passer par une accumulation de données quantitatives et systématiques sur les possibilités de choix d'unités appartenant à chacun des pôles SS ou SGI. Les observations ici faites militent donc pour une exploration outillée de corpus attestés, comme préalable à toute tentative de TALS.

4 Méthodes et outils pour l'exploration de corpus en LSF

Nous présentons ici quelques résultats préliminaires de l'exploration de corpus de transcription d'énoncés en LSF tirés de LS-COLIN, en vue d'en faire émerger des régularités au niveau syntagmatique. Nous déterminons quels paramètres semblent les plus adaptés dans la perspective ici adoptée, puis nous discutons quelques résultats d'exploration de corpus de transcription, à l'aide des outils évoqués (i.e. n-grammes, concordances).

---

[5] Ex. : ordre privilégié Contenant>Contenu. Pour plus de détails, voir également (Lejeune, 2004).

4.1 Analyse de concordances tirées des corpus de transcription

Les observations relatées ici reposent sur l'analyse de concordances réalisées sur le champ « étiquetage linéaire » des transcriptions tirées de (Sallandre, 2003). En effet, ce champ contient l'ensemble des unités identifiées, qu'elles soient du type standard ou non, il constitue donc le lieu d'observation privilégié pour les alternances entre unités standard et non standard. Les concordances permettent l'analyse des contraintes distributionnelles des unités linguistiques, elles peuvent être réalisées grâce à une base de 5-grammes construits à partir de la fusion des étiquettes linéaires de toutes les transcriptions, production par production, de LS-COLIN. La base de données ainsi constituée permet, par exemple, d'aborder la description synthétique des contraintes s'exerçant sur les unités et segments pertinents ci-dessous.

4.1.1 Début et fin de récit (Histoire du Cheval)

Pour l'*Histoire du Cheval*, seuls 2 locuteurs sur 13 utilisent des unités de la catégorie transferts (TTF) dès le début de leur production. Pour ce récit, 11 locuteurs sur 13 suivent un schéma narratif d'introduction du thème par des SS. Ce comportement va dans le même sens que l'intuition des transcripteurs, qui ont isolé le schéma : [introduction de thème en SS] suivi de [SGI]. L'examen des contextes immédiats de la marque de fin de récit montre que seuls 3 locuteurs sur 13 terminent leur récit par des unités appartenant aux SGI.

4.1.2 Contextes gauche et droit de Signe Standard (Histoire du Cheval)

Dans la plupart des cas, les SS sont enchaînés en séquences de plusieurs unités, puis, dans l'ordre décroissant, les SGI (TP, TTF, etc.), puis les Pointages. Un SS semble être avant tout un introducteur d'un autre SS, puis d'une SGI (i.e. TP, TTF, TS, et d'autres sous-types). Par comparaison entre les contextes gauche et droit, on peut souligner que, pour le récit concerné, les SS sont plus souvent introduits par un pointage qu'ils n'introduisent un pointage.

5 Conclusion

Comme nous l'avons évoqué dans le présent exposé, le TALS se heurte à des difficultés, tant pour la description que pour la modélisation des énoncés. Ces difficultés sont essentiellement liées à la dynamique énonciative, aux problèmes d'interprétation en contexte et à l'intégration synchronisée de plusieurs niveaux linguistiques à l'œuvre en LSF. Une des avancées que nous attendons de l'interaction entre linguistique formelle, linguistique de corpus et LSF est, précisément, de poser la nécessité d'un réel traitement du contexte, de la sémantique et de la dynamique conversationnelle tant pour les LS que pour les autres LN, qui ne peut, à notre sens, que passer par un travail sur corpus attestés et par l'exploration de nouveaux cadres formels non nécessairement logiques, telles que les Grammaires de Construction Radicales de (Croft, 2001) ou la Théorie de l'Optimalité (Prince & Smolensky, 1993). Nous proposons donc, comme étape préalable à un TALS la mise en œuvre et le développement d'outils adaptés pour le travail d'exploration de corpus de transcription en LSF, en ayant en tête l'élargissement à d'autres langues, voire d'autres domaines d'exploration. Une partie de ces outils est en cours d'élaboration et de validation, ils ont vocation à être intégrés à la plate-

forme CoPT[6] (Corpus Processing Tools) dont la constitution a été, en partie, dictée par les explorations sur corpus de transcription en LSF ici relatées.

Références

CROFT W. (2001), *Radical Construction Grammar*, Oxford University Press.

CUXAC C. (1996), *Fonctions et structures de l'iconicité des langues des signes*, thèse de Doctorat d'Etat, Université Paris V.

CUXAC C. (2000), La Langue des Signes Française; les Voies de l'Iconicité, *Faits de Langues* n°15-16, Paris: Ophrys.

LEJEUNE F. (2004), *Analyse sémantico-cognitive d'énoncés en Langue des Signes Française pour une génération automatique de séquences gestuelles*, thèse de doctorat, Université Paris XI.

PRINCE A., SMOLENSKY P. (1993), *Optimality Theory, Constraint interaction in generative grammar*, Technical Report, ROA.

SALLANDRE M.-A. (2003), *Les unités du discours en Langue des Signes Française. Tentative de catégorisation dans le cadre d'une grammaire de l'iconicité*, thèse de doctorat, Université Paris VIII.

[6] Voir le descriptif de la plate-forme, ainsi que quelques modules à l'adresse : http://copt.sourceforge.net.

Système d'annotation et de segmentation de gestes de communication capturés

Alexis Héloir, Sylvie Gibet, Nicolas Courty, Mickaël Raynaud

Samsara Valoria – UBS
BP 573, 56017 Vannes Cedex
prenom.nom@univ-ubs.fr

Mots-clés : Annotation, Synthèse du Geste, Capture de Mouvements, Segmentation, Indexation

Keywords : Annotation, Gesture Synthesis, Motion Capture, Segmentation, Indexation

Résumé Nous proposons un outil permettant l'annotation de gestes préalablement acquis par des méthodes récentes de capture du mouvement. La segmentation et l'annotation peuvent être réalisées selon différents niveaux de représentation du geste. Cet outil est conçu pour être utilisé aussi bien par des signeurs non informaticiens que par des spécialistes de l'animation par ordinateur. Dans le cadre de notre projet, cet outil a pour vocation la création d'une base de données de représentations numériques de gestes capturés capable d'alimenter des méthodes originales de synthèse et de spécification de gestes de communication. Nous pensons qu'il peut également s'avérer utile dans le contexte de l'analyse linguistique des langues signées.

Abstract We present an annotation tool handling motion capture gestures captured via recent techniques. Segmentation and annotation can be carried out according various representation levels. This tool is meant to be used by computer graphics experts as well as signing individuals with no computer oriented prerequisites. The main goal of this tool, according to our objectives, is to set up a coherent data base of signals. Data provided should support our investigation on designing original methods for specification and synthesis of expressive gestures. We also believe that such a tool can prove to be useful in the context of linguistic analysis of signed languages.

1 Introduction

Les méthodes d'analyse et de segmentation du geste permettent d'indexer des bases de données de mouvement en introduisant des éléments structurels contribuant à enrichir les modèles de synthèse et de spécification du mouvement existants. Cette étape d'analyse nécessite des outils appropriés. Plusieurs projets ont mis au point ou sont en train de réaliser des logiciels d'annotation et de transcription de séquences gestuelles. Ces outils ont pour principal flux de données des séquences vidéo pouvant être enrichies de signaux tels que la direction du regard (Wittenburg, 2002) ou les configurations manuelles (Koizumi, 2003).

Bien que certains logiciels proposent des traitements évolués tels que la reconstruction de postures par analyse d'images (Dalle, 2001), à notre connaissance, aucune des solutions existantes n'offre la possibilité de travailler directement sur des représentations numériques du mouvement telles qu'elles sont obtenues par l'intermédiaire des outils de capture récents. Ce résumé décrit le logiciel de segmentation et d'annotation de mouvements capturés que nous sommes en train de développer en présentant dans un premier temps les signaux pris en compte et l'intérêt de tels signaux dans le cadre de nos recherches et dans le cadre de la linguistique. Nous décrivons ensuite la structure et le format des annotations. Nous concluons par les objectifs et les perspectives de notre travail.

2 Description des signaux acquis

Nous disposons d'un système d'acquisition composé d'un système optique Vicon réalisant la capture des mouvements corporels et de l'expression faciale synchronisés avec une paire de gants de données. Ce système permet de capturer l'évolution de la majorité des articulateurs entrant en jeu dans la langue des signes (corps, visage et mains). En sortie du système d'acquisition, nous obtenons des vecteurs de positions (angulaires ou cartésiennes) caractérisant l'évolution de chaque marqueur de mouvement au cours du temps (figure 1). Ces données sont structurées selon un squelette articulaire modélisant le corps et les mains de l'acteur, ainsi qu'un treillis relatif aux marqueurs disposés sur son visage.

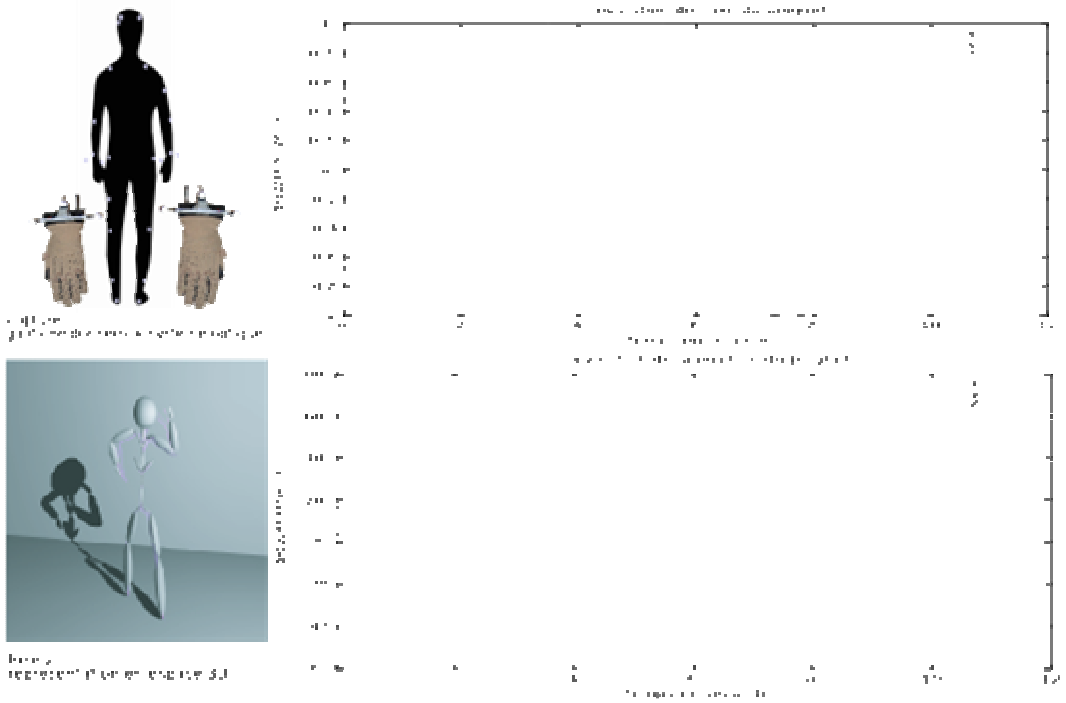


Figure 1 : signaux numériques capturés

L'outil d'annotation est en cours de développement. Un précédent prototype nous a permis d'évaluer sa faisabilité et de valider certains choix d'interface. Outre les fonctionnalités propres à la plupart des logiciels d'annotation existants, notre outil propose une représentation animée du mouvement dans un espace tridimensionnel permettant à l'utilisateur de choisir son point de vue. Cette représentation est accompagnée d'une vidéo de l'acteur enregistrée pendant la séquence de capture et synchronisée avec la visualisation dans l'espace

tridimensionnel. Enfin, notre outil permet l'affichage des trajectoires des signaux numériques éventuellement inclus dans les pistes d'annotation sous la forme de courbes intégrées aux lignes de temps des pistes d'annotation.
L'utilisation de la représentations numérique du geste offre la possibilité d'effectuer des traitements automatiques pour la segmentation (Barbic, 2004), facilitant ainsi le processus de segmentation manuelle et d'annotation.
Ce type de représentation permet également la création d'une base de données permettant des recherches multi-critères (spatialisation du geste, dynamique du geste, analyse en composantes principales, etc.).

3 Structure et format des annotations

La structure des fichiers annotés obéit à un format de données basé XML et est conçue pour répondre aux préoccupations de différents utilisateurs, signeurs, linguistes ou informaticiens. Le format d'annotation est composé de quatre sections (Figure 2).

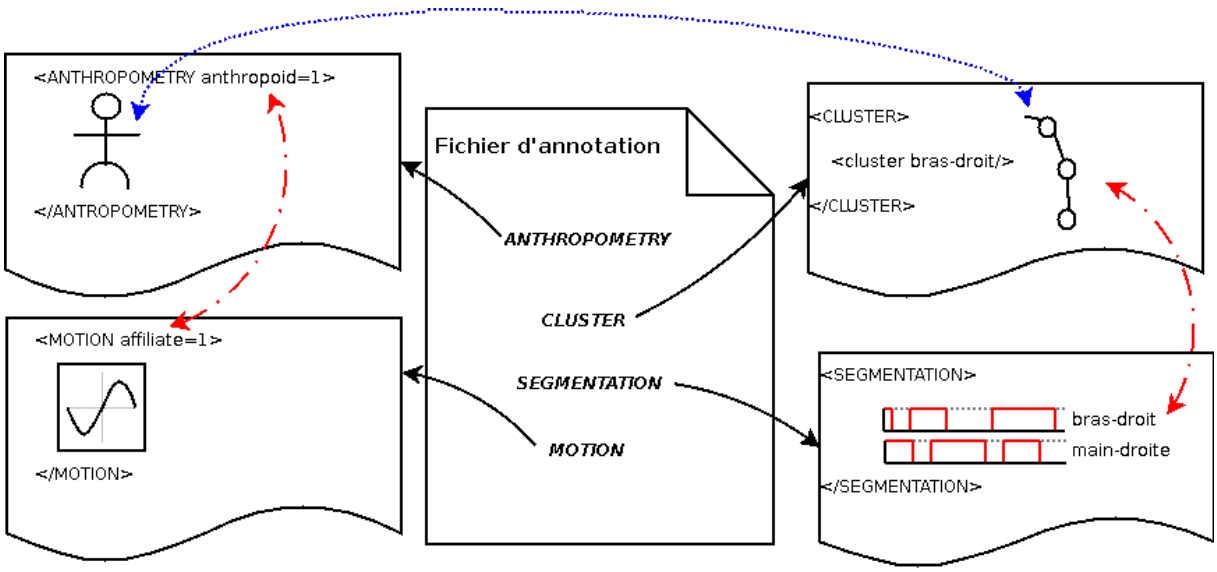


Figure 2 : Structure des annotations

La section ANTHROPOMETRY décrit de manière hiérarchique la morphologie du squelette articulaire de l'acteur (corps et mains pour le moment).
A cette anthropométrie, sont affiliés des signaux capturés et définis dans la seconde partie intitulée MOTION.
La section CLUSTER permet de regrouper en sous-structures hiérarchiques signifiantes (bras gauche, main droite, etc) des éléments de la hiérarchie décrite dans la première section. Ces clusters constituent le support à l'annotation.
Enfin, la dernière partie, intitulée SEGMENTATION contient les pistes d'annotation. Ces pistes sont spécialisées selon l'annotation que l'on désire effectuer. Une segmentation de type signal permet de décomposer les mouvements élémentaires en phases caractéristiques ; une segmentation de plus haut niveau permet d 'étiqueter les gestes suivant leurs paramètres formationnels (configurations manuelles et digitales, primitives de mouvement, emplacement et orientation). De nouveaux types peuvent être définis pour permettre l'extension des possibilités de l'annotation et l'adaptation à d'autres catégories de mouvements.

4 Perspectives

Un outil de segmentation et d'annotation du geste de communication a été brièvement présenté. Il se base principalement sur des représentations numériques du mouvement obtenues par des appareils de capture récents. L'accent a été mis sur l'ergonomie du logiciel ainsi que sur la souplesse des structures d'annotation utilisées. Les données annotées et segmentées obtenues nous permettront d'étendre les langages de spécification du geste (Lebourque, 99). Nous nous appuierons également sur ces données pour enrichir les méthodes de synthèse déjà utilisées (Gibet, 2001) et explorer les méthodes récentes de synthèse par graphes de transition (Kovar, 2002). Bien que ce logiciel soit dédié à l'annotation de données pour la synthèse, nous pensons qu'il peut s'avérer utile pour l'analyse linguistique du geste, notamment pour étudier les relations de spatialisation propres à la langue des signes (Lejeune 2001).

Remerciements

Travail réalisé dans le cadre du projet Signe, financé par la région Bretagne (Réf. B/1042/2004/SIGNE) et du projet RobEA HuGEx, financé par le département STIC du CNRS. La capture des données présentée dans cet article a été rendue possible grâce au LPBEM de Rennes 2, au LESP de Toulon et du Var et au LINC de Paris 8.

Références

WITTENBURG P., LEVINSON ST., KITA S., BRUGMAN H.(2002), Multimodal Annotations in Gesture and Sign Language Studies, *Int. Conf. on Language Resources and Evaluation.*

KOIZUMI A., SAGAWA H., TAKEUCHI M.(2003), An Annotated Japanese Sign Language Corpus,. Actes de *Meeting of the Association for Computational Linguistics (ACL).*

DALLE P., LENSEIGNE F., HUDELOT C.(2001), Apport d'un système d'analyse d'images à l'étude de la Langue des Signes, Actes de *Journées d'études Recherches Sur Les Langues Des Signes.*

BARBIC J., SAFONOVA A., PAN JY.,FALOUTSOS C., HODGINS J., POLLARD N.(2004), Segmenting Motion Capture Data into Distinct Behaviors, Actes de *conf. on Graphics interface*, 185–194.

LEBOURQUE T., GIBET S.(1999), High level specification and control of communication gestures : the GESSYCA System, Actes de *Computer Animation*, 24-36.

GIBET S., LEBOURQUE T., MARTEAU PF.(2001), High level Specification and Animation of Communicative Gestures, *journal of Visual Languages and Computing*, Vol. 12, pp.657-687.

KOVAR L., GLEICHER M., PIGHIN F.(2002), Motion graphs, Actes de *conf. on Computer graphics and interactive techniques*, 473-482.

LEJEUNE F., BRAFFORT A., DESCLÉS JP.(2001), Study on Semantic Representations of French Sign Language Sentences. *int. Gesture Workshop*, 197-201.

Using SignWriting as a Phonetic Notation System

Vívian Bonow Boeira, Luis Roberto Volz de Oliveira,
Diogo Souza Madeira, Antônio Carlos da Rocha

Escola de Informática, Universidade Católica de Pelotas, Brazil.
Email: rocha@atlas.ucpel.tche.br

Mots-clés : Langues des signes, SignWriting, transcription phonétique
Keywords: Sign languages, SignWriting, phonetic transcriptions

Résumé Cet article décrit une expérience d'utilisation du système SignWriting comme une notation phonétique pour ce type de langue. On a utilisé le système SignWriting pour faire la transcription phonétique d'une vidéo portant une conversation en Langue de Signes Française (LSF). La transcription a été faite au niveau phonétique, et non pas aux niveaux phonologique ou lexical. Le fait que les paramètres phonétiques des langues de signes ont un caractère universel a permis que la transcription soit faite sans que les transcripteurs connaissent la LSF. L'article présente initialement une brève description du système SignWriting. Après, il introduit les principaux paramètres phonétiques qui ont été utilisés dans la transcription. Finalement, il présente et décrit les transcriptions de quelques signes du dialogue.

Abstract This paper describes an experiment in the use of the sign language notation system, called SignWriting, as a phonetic notation system for that kind of language. We have used the SignWriting system for the phonetic transcription of a video containing a conversation in French Sign Language (LSF). The fact that the phonetic parameters have a universal character allowed the transcription without the transcribers knowing LSF. The paper presents initially a brief description of the SignWriting system. Then it introduces the main phonetic parameters that were used in the transcription. Finally, it presents and describes the transcriptions of some of the signs of the dialogue.

1 Introduction

Transcribing sign language vídeos into a notation system is a technical activity, that requires training and ability. A part from the visual accuracy needed to grasp the details of gestures, a precise understanding of the semantics of the notation system is also necessary. Whenever the notation system uses specially created symbols, with contrived semantics derived from an underlying linguistic account of the sign languages, theoretical understanding of that linguistic account is also often necessary. This paper aims to show that the SignWriting system (Sutton, 2005) can be put into use as a phonetic notation system that requires almost

no linguistic theoretical background from the transcribers, since the visual characteristics of the system immediately give them the phonetic intuition behing the symbols of the system. We do that by showing the transcription of the FSL dialogue contained in the video **CE1.mov** that belongs to the sample corpus offered to the participants of the TALS 2005 workshop. To emphasize the almost blind condition in which the transcription work was done, we note that the transcribers based their work only on their knowledge of the set of phonetic parameters they analyzed in the signs, and on their knowledge of SignWriting: they do not know LSF. Of course, this makes of the obtained transcription a very rough and preliminary result, that could be immensely improved in many ways by someone with a proper knowledge of LSF. Anyway, the idea is to show the simplicity, ease of writing and reading, and adequate faithfulness of SignWriting as a phonetic transcription system.

2 The SignWriting system

The SignWriting system (Sutton, 2005) was created as visual notation system for sign languages. It's aim is not the technical transcription of sign languages, from the point of view of their linguistic studies. On the contrary, its principal aim is allowing the daily writing of sign languages, so that Deaf people can write texts in sign languages in the same way that hearing people can write texts in oral languages. The SignWriting system was inspired by a choreographic notation system called DanceWriting, and both are subsystems of a general movement writing system called Sutton Movement Writing (Sutton, 2005). The original choreographic foundation of the system was enriched with linguistic information (mainly phonetic information) indispensable for a clear transcription of sign languages. The result is a notation system that, although directed to the non-technical daily writing in sign languages, is able to give precise phonetic transcriptions of such languages.

3 The phonetic parameters

The SignWriting system is a set of graphical symbols for the representation of the various features that characterize the phonetic description of signs. The symbols of SignWriting belong to a larger set of symbols create for the detailed representation of general body movements, not only signs, called IMWA (International Movement Writing Alphabet) (Sutton, 2005). IMWA is organized in categories, as shown in Fig. 1. Each category is divided into groups of symbols. For instance, the category Hand includes symbol groups for representing, among others, hand configurations with the various selected features, for instance, index finger, middle finger, groups of fingers, etc., in various shapes (straight, bent, curved, etc.). Hand orientation is indicated by decorations applied to symbols, like coloring, rotations, etc. Positions of hands are expressed by reading conventions, together with symbols for contacts, etc. Hand movements are represented by a variety of arrows. Also, a rich set of symbols exist for the representation of aspects of facial expressions, including mouthing, eyebrows, eyegaze, etc.

Hand	Hand configurations
Movement	Arrows and other symbols, for the representation of movements
Face	Detailed representation of facial expressions
Head	Head movement, positions and location
Upper Body	Upper body movement, positions and location
Full Body	Limbs, limb location and full-body gestures
Space	Planes, room location and group patterns
Punctuation	Symbols for indicating prosodic features of the movements

Figure 1: Symbol categories.

4 The phonetic transcription of some signs of the dialogue

The full transcription of the dialogue in the **CE1.mov** video is available at http://gmc.ucpel.tche.br/TALS2005. Figure 2 shows a simple phonetic transcription of first sign of the dialogue.

Figure 2: First sign of the dialogue in video **CE1.mov.**

The transcription focuses on the hand configurations and the touch between the finger and the hand. Closer and more careful examination of the sign would tell if the choices made in the transcription are adequate, or if they need to be trimmed. For instance, the right hand was represented as a flat hand with spread fingers. It could as well be represented as a flat hand with fingers together. Figure 3 shows two other simple phonetic transcriptions.

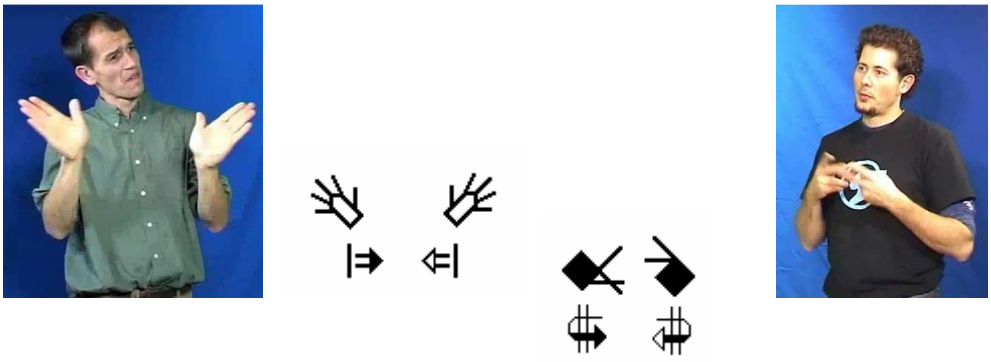


Figure 3: Simultaneous signs in the dialogue.

Again, the focus is on the hand configurations. Note the choice for representing a V and a P hand, in the sign of the second signaler. The fact that we have focused on hand configurations is due to short time available and inexperience in transcribing sign videos, that make difficult for us to grasp other features, like face expression. SignWriting can represent face expressions, and also many other sign features, like body shifting, etc. Also, we have chosen to write in the so-called "receptive" point of view, that of looking at a signer. For daily writing, it is usually preferred to write in the "expressive" point of view, that of the signer.

5 Remarks on the transcription

The full transcription is available at the site http://gmc.ucpel.tche.br/TALS2005. The fact that we don't know FSL prevents us to understand the dialogue and give a translation for it. Also, it prevents us from reaching a phonological representation of the signs in the dialogue, since we have no direct clue about phonological equivalences of different phonetic features. But the fact that the phonetic features of sign languages have a universal character helped us in producing a phonetic transcription that seems quite acceptable, as a preliminary result (that needs, of course, to be reviewed by native FSL signers). More important, the reading of the transcription is quite easy, after a short acquaintance with the SignWriting system, which makes the transcription potentially useful to an audience wider than that composed only of expert in linguistics, if it is looked as a written presentation of the dialogue, in a writing system for daily use. Due to our ignorance of FSL , a persistent difficulty during the transcription was the clear separation of signs and phrases, to allow for a neat presentation of individual signs. Such separation was not always possible, due to the many phonological and prosodic processes operating on the signs. So, we had often to arbitrarly separate complex gestures into individual parts, in order to present such gestures as sequences of individual "signs". We are sure that many of the separations that we guessed may simply be wrong.

6 Conclusion

The SignWriting system was originally proposed as a daily writing system for sign languages. The main kind of linguistic information incorporated in SignWriting is of the phonetic kind. So, the writing in SignWriting tends to be of a phonetic kind, thus its natural application to phonetic transcriptions. The simplicity of the graphical symbolism of the system, makes it quite easy to read. Moreover, its frame-based structured (organized around the so-called "sign boxes"), makes it suitable for the transcription of dialogues, as exemplified by the full dialogue transcription in http://gmc.ucpel.tche.br/TALS2005. Other video transcriptions are available on the Internet, at (Sutton, 2005). On the other hand, we want to emphasize that even if it happens that as a technical device for phonetic transcription the SignWriting system needs to be improved, its solid foundation in the idea of **movement writing** (Sutton, 2005) makes of that system a strong candidate not only for such purpose but also – and most important – for its original purpose of being a writing system for daily use.

Acknowledgments

This work is partially supported by FAPERGS and CNPq. Valerie Sutton continuously shared with us her knowledge about movement writing, and also her enthusiasm for the subject.

Reference

SUTTON, V. (2005), The SignWriting System. La Jolla, California, Center for Sutton Movement Writing. Homepage: http://www.signwriting.org.

Semantic Searching for SignWriting

Steven Aerts (1), Bart Braem (1),
Katrien Van Mulders (2), Kristof De Weerdt (2)
(1) Université d'Anvers
Middelheimlaan 1, B 2020, Anvers, Belgique
{bart.braem, steven.aerts}@ua.ac.be
(2) Université de Gand
Rozier 44, B 9000 Gand, Belgique
{katrien.vanmulders, kristof.deweerdt}@ugent.be

Mots-clefs : analyse de signes, reconnaissance automatique, SignWriting

Keywords: sign analysis, automatic recognition, SignWriting

Résumé Dans cet article nous vous pésentons les résultats de recherches d'un système en ligne de langue des signes indépendant d'annotations manuelles et basé sur SignWriting. La recherche se fait d'une façon intuitives mais flexibles. Les résultats sont classés d'après leur relevance. Le système est en ce moment utilisé pour le dictionnaire de langue des signes flamand contenant plus de 7000 signes.

Abstract In this paper we present the development results of an online sign searching system independent of manual annotations based on SignWriting. Lookup is done on an intuitive yet flexible basis and results are ordered by relevance. The system is currently active for the Flemish Sign Language dictionary containing over 7000 signs.

1 Introduction

We have developed an online database driven dictionary system named `Dixit` currently powering the Flemish Sign Language Dictionary with over 7000 signs(Aerts, Braem, De Weerdt, Van Mulders, 2004-2005; Verweire, 2005). These signs were collected by researchers of the university of Ghent. `Dixit` can convert SignWriting signs typed with SignWriter DOS (Gleaves, Sutton, 1985-2004) to SWML-D. We developed SWML-D as an XML-based representation language for SignWriting dictionaries based on SWML (da Rocha Costa, Dimuro, 2005). The `Dixit` database is modelled on the hierarchical SWML-D structure in order to contain exactly the same information.
The SignWriting system itself is a practical visual writing system for sign languages, composed of a set of intuitive graphical-scadhematic symbols and simple rules for combining them to represent signs(da Rocha Costa, Dimuro, 2005). It was invented by Valerie Sutton inspired by her already created choreographic writing, called DanceWriting (Sutton, 1996-2005; Sutton, 1996-2005). SignWriting symbols represent the body parts involved and the movements and

face-expressions made when producing signs. We based `Dixit` on SignWriting because it is understandable for people who have never seen it before.
In this paper we discuss the internals of our intuitive, user-friendly, yet powerful search by sign system for SignWriting. We started off from the idea that all information is automatically extracted from the signs, without manually enriching them with external information. This machine learning approach overcomes tedious finetuning which is required in other proposed systems (da Rocha Costa, Dimuro, Freitas, 2004).

2 The manual approach

Searching for the meaning of a sign in a database manually enriched with extra semantic information is common practice. It usually consists of selecting the type and direction of the movement, the location on the body where the sign is made and finally the hand form. This information needs to be added to each individual sign (Wilcox Scheibman, Wood, Cokely, Stokoe, 2000), which causes a big slowdown when composing a dictionary with thousands of signs.

3 Semantic view on SignWriting

The first consideration to make is which information can be computed out of SignWriting signs. Movements distinction is reasonable as movements are represented by different symbols. Although trivial for humans determining the moving body part is impossible to implement without an extensive physiological model. The same logic is valid when detecting three-dimensional direction of the movement out of a two-dimensional representation.
When a body part is touched SignWriting gives a very good clue on the zone touched. However when no body part is touched, the location can only be extracted from the symbols by considering the most likely positions. SignWriting depicts a large number of touch variations but most of the time users want more general selection. Thus we allow users to look for the five major SignWriting touch groups (touch, grasp, in-between, strike, brush contact and rub contact).
One very nice property of SignWriting is the accurate and intuitive distinction between hand forms. This is the main feature we search by.

4 Dissection of a search

Everything starts with the user specifying the hand form, which of the five body zones that hand touches (head, torso, arms, legs and hands) and the way in which they are touched, see figure 1. The system can now easily rule out signs that do not contain the selected hand forms, touches if any and body parts. This results in a rather small set of probable signs which can be evaluated thoroughly. This evaluation starts with determining the contact zones and discarding signs without the requested zones. If multiple zones are involved, they have to be matched with the touching body parts.
We have parametrized the semantic goodness of a match to allow ordering by relevance. This goodness-measure is based on the summation of the product of the Euclidean distance between the touch and the middle of the corresponding body zone. We introduced additional improve-

ments based on user feedback: a simple sign will be ranked higher when compared to more complicated signs, very bad matches are dropped to avoid confusion.
The resulting ordering by relevance does not correspond to the SignWriting ordering(Butler, 2001; Sutton, 2004), because for that to be possible information about the dominant hand would be necessary which is impossible to compute without the physiological model. Notice that this search requires no advanced SignWriting knowledge from our users, which turned out positive as it is a publicly available system mostly used as a reference worked by SignWriting novices.

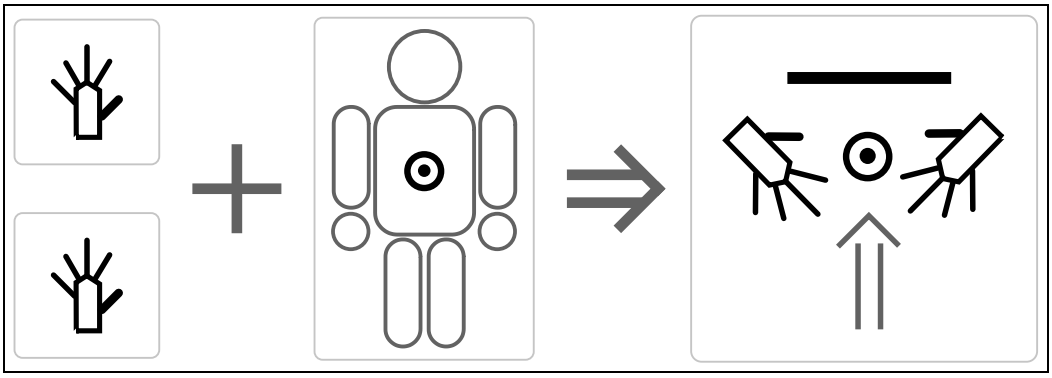


Figure 1: Search for *two opened hands rub torso* returns the Flemish sign for "to enjoy".

5 Performance

Selecting signs with the right symbols and contacts is a database issue where performance is not at stake. Determining the body parts is harder but also be done in the database: in general one thick black horizontal line stands for the chest, whereas two lines can depict the legs and hips.
Calculating the goodness-measure is done over a very limited number of matching symbols and contacts: a sign containing four contacts is extremely rare[1] and will most probably never be searched for by a user. Thus the number of comparisons will be low and will not affect the global performance by one order-of-magnitude. Webserver statistics proved that most queries to take less than 0.5 seconds.

6 Future work

Currently some of the goodness measure parameters are fixed numbers that were finetuned manually while developing. The nature of this problem lends itself perfectly for mapping to the structure of a neural network which would allow an automatic optimization of these parameters (Cohen, Schapire, Singer, 1998). Currently `Dixit` powers the Flemish Sign Language Dictionary however it is highly portable and sign language independent. In the near future we plan to open source `Dixit` to allow wider usage in the Deaf Community, both as a scientific framework for sign language research and as an advanced and intuitive sign language dictionary (Braem, 2000).

[1] We counted 35 signs with 4 or 5 contacts out of 7460 signs.

7 Conclusion

Searching with `Dixit` is intuitive even for a user with very basic SignWriting knowledge. Its friendliness is very high, as proved by user reactions and daily usage. We showed that applying a goodness measure combined with a broadened search makes it a powerful sign language tool. The real strength of the search system lies in the use of the very well specified SignWriting hand forms, which compensates for the vague movements. Because of the use of databases, SWML-D and relatively simple calculations, the method presented allows effecient lookups. Most importantly, the Deaf Community and its researchers will benefit by this new search method since it allows for easier dictionary-searching and linguistic research.

Références

AERTS S., BRAEM B., DE WEERDT K., VAN MULDERS K. (2004-2005), Woordenboek Vlaamse Gebarentaal, `http://gebaren.ugent.be/`.

DA ROCHA COSTA A.C., PEREIRA DIMURO G. (2001), A SignWriting-Based Approach to Sign Language Processing, *The SignWriting Journal*, Vol. 0. 1.

VERWEIRE E. (2005), Vlaamse gebaren gedigitaliseerd (Flemish signs digitalised), *EOS magazine in cooperation with Scientific American*, Vol. 22, 1.

DA ROCHA COSTA A.C., PEREIRA DIMURO G. (2003), SignWriting and SWML: Paving the Way to Sign Language Processing, *TALN 2003 (Batz-sur-mer)*, `http://www10.org/cdrom/posters/p1011/`.

BUTLE C. (2001), An Ordering System for SignWriting, *The SignWriting Journal*, Vol. 0. 1, 1, `http://sw-journal.ucpel.tche.br/number1/article1.htm`.

SUTTON V. (2004), Sutton's SignSpelling Guidelines 2004, *SignWriting Library Database*, 0-914336-83-5.

GLEAVES R., SUTTON V.(1985-2004), SignWriter DOS, `http://www.signwriting.org/`.

SUTTON V. (1996-2005), SignWriting: Read, write, type Sign Languages, `http://www.signwriting.org/`.

SUTTON V.(1996-2005), DanceWriting: Read and Write Dance, `http://www.dancewriting.org/`.

COHEN W.W., SCHAPIRE R.E., SINGER Y. (1998),Learning to Order Things, *Advances in Neural Information Processing System*, Vol. 10.

DA ROCHA COSTA A.C., PEREIRA DIMURO G., BALDEZ DE FREITAS J. (1998), A Sign Matching Technique To Support Searches In Sign Language Texts, *Proceedings of the LREC 2004 workshop for Representation and Processing of Sign Language* .

BOYES BRAEM P. (2000),Sign Language Text Transcriptions and Analyses Using Microsoft Excel, *Center for Sign Language Research, Basel, Switzerland* .

WILCOX S., SCHEIBMAN J., WOOD D., COKELY D., STOKOE W.C. (2000),Multimedia dictionary of American Sign Language, *Proceedings of the first annual ACM conference on Assistive technologies table of contents* 0-89791-649-2.

Variations dans la représentation écrite d'un signe en Signwriting

Guylhem Aznar (1), Patrice Dalle (2)

Laboratoire IRIT, Équipe TCI –UPS, 118 route de Narbonne, 31062 Toulouse
(1) aznar@irit.fr (2) dalle@irit.fr

Mots-clés: unicode, XML, SWML, méta-données, langue des signes, SignWriting
Keywords:unicode, XML, SWML, metadata, variation, sign language, SignWriting

Résumé La langue des signes peut être transcrite selon un formalisme d'écriture. Le plus connu et le plus utilisé se nomme SignWriting: un signe y est transcrit en symboles, qui constituent un signe SignWriting. Le choix de ces symboles ainsi que leur disposition spatiale dans le signe SignWriting sont responsables d'une perte de bijectivité: pour un signe, plusieurs symboles peuvent être utilisés. Ceci pose des problèmes pour l'informatisation des documents rédigés en SignWriting. Le problème se pose vraisemblablement pour tout formalisme d'écriture bidimensionnel composé. Les solutions proposées permettent de traiter ces documents comme des documents standards.

Abstract Following a writing formalism, sign language can be transcribed. The best known and the most widely used is called SignWriting: a sign is transcribed into symbols, which constitue a SignWriting sign. The choice of these symbols along with their spatial disposition within the sign are responsible for a loss of bijectivity: for one sign, different symbols can be used. This is causing problems for computer support of documents written in SignWriting, a problem likely present in every bidimensionnal writing formalism . Solutions are proposed here to manage such documents as standard documents.

1 Le formalisme SignWriting (SW)

1.1 Définition: signe, signe SW, symboles, cellules

SW (Sutton, Gleaves, 1995) fait partie d'un ensemble de systèmes conçus pour transcrire n'importe quel mouvement. Un signe y est transcrit en signe SignWriting, composé de symboles positionnés dans des cellules. Les symboles correspondent à des positions statiques ou dynamiques des différentes parties du corps. Ces symboles sont particulièrement nombreux: 425 dans la version 2003, répartis en 60 groupes selon 10 catégories. Chacun peut avoir 4 représentations, 6 remplissages et 16 orientations spatiales différentes. Par rapport à d'autres systèmes graphiques de la langue des signes, comme HamNoSys (Prillwitz et al 1987), plus orientés vers la transcription, SW semble paradoxalement plus adapté pour écrire la langue des signes, sans pour autant avoir été initialement conçu dans ce but. Toutefois, SW présente un problème, lié à sa complexité, pour être informatisé. Plusieurs dizaines de milliers de symboles uniques sont utilisables, dans une infinité de combinaisons pour composer un signe, selon leur position bidimensionnelle respective. Cette trop grande variabilité nécessite

un encodage se basant sur la composition. Si plusieurs encodages se sont succédés, SWML (DaRocha, Dimuro,2003) est le plus utilisé actuellement. Des logiciels tel SW-Edit permettent de saisir en SWML. Un encodage basé sur Unicode a été proposé pour pallier à certains problèmes (Aznar, Dalle, 2004) .

2 Problème: variation des signes SignWriting

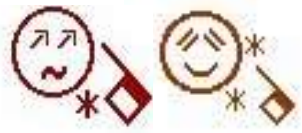
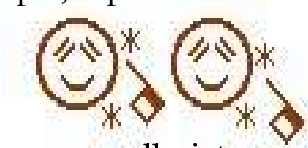
Toutefois, une classe de problèmes a persisté malgré les changements d'encodage: la variabilité des symboles utilisés dans un signe SW. Ainsi, il n'y a pas de bijectivité entre un signe et le signe SW qui le représente, car la transcription peut se faire de plusieurs manières.

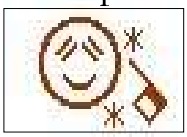
2.1 Variabilité inter personnelle: liée au choix des symboles: bouche ~ ou

Le choix des symboles peut se concevoir comme une variabilité inter-personnelle liée à la très grande permissivité de SignWriting qui permet de transcrire un signe de plusieurs manières. Prenons par exemple le signe correspondant à « sourd », identique en langue des signes française et américaine. Plus de 7 variations existent, selon le choix des symboles. Chaque variante est totalement interchangeable: elles ont toutes strictement le même sens.

2.2 Variabilité intra-personnelle: liée au positionnement manuel:  ou 

À chaque itération d'écriture d'une variante donnée d'un signe, les symboles seront très légèrement déplacés, du fait de l'imprécision d'une manipulation manuelle (à la souris...) dans un plan bidimensionnel. Ces modifications sont difficilement perceptibles, mais importantes pour le système informatique puisque la position exacte des signes de base est enregistrée. Même si cette dernière ne varie pas, la position relative du symbole dans la cellule peut varier.



Figures 3, 4, 5: variabilité intra-personnelle, inter-personnelle, et inter-personnelle de cellule

La variabilité peut se problématiser en trois étapes : choix des symboles (inter-personelle), puis positionnement de ces symboles les uns par rapport aux autres (intra-personelle), puis positionnement du signe final dans la cellule (inter-personelle de cellule). Dans le premier cas, les symboles changent, dans le deuxième cas, leurs cordonnées par rapport aux autres changent, dans le dernier cas, leurs coordonnées par rapport aux autres ne changent pas.

2.3 Rôle de la saisie par l'utilisateur et de la gestion informatique

L'utilisateur des différents logiciels permettant de saisir des signes SW procède toujours de la même manière: il choisit le groupe des symboles, puis celui qu'il veut rentrer, pour ensuite le positionner sur le plan bidimensionnel constitué par la cellule SW. Si certains logiciels proposent des dictionnaires permettant de saisir directement un signe, l'aide à la saisie n'est pas encore gérée: l'utilisateur doit choisir et positionner son symbole. La recherche d'un signe SW dans un document ou un dictionnaire constitue donc un problème de recherche à part entière. Les solutions proposées reposent sur des formules demandant soit un nombre de paramètres liés à la dynamique du signe comme la forme de la main, la direction du mouvement, le type de contact évenuel (Aerts et al, 2004), soit la décomposition en symboles de ce signe (DaRocha et al, 2004). Le problème est encore plus important pour une analyse lexico-sémantique (Huenerfauth, 2002) de documents en SW: quelque soit l'encodage, dans la

mesure où un même signe peut être représenté par plusieurs signes SW, toutes les équivalences seraient à considérer pour neutraliser l'effet des variations amenées par le formalisme d'écriture, et ne s'intéresser qu'au signe initialement transcrit.

2.4 Partage du document SignWriting

Si chacun peut transcrire les signes à sa manière, les avantages liés à la numérisation pourraient être limités par les possibles incompréhensions amenées par l'utilisation de variantes peu fréquentes ou exagérées. Toutefois, de nombreux travaux ont démontré le rôle des aires motrices du cerveau dans la reconnaissance de mouvements. SW transcrivant la dynamique du mouvement, il est très probable que la reconnaissance des signes fasse appel aux aires motrices du cerveau, et non aux aires liées à la compréhension, comme il a été déjà démontré pour la reconnaissance de l'écriture manuscrite (Knoblich et al, 2002). Comprendre un signe même si ses symboles varient ne serait donc que ralenti, et non rendu impossible.

3 Solution proposée: rajout de métadonnées à la saisie

La solution la plus simple consisterait à discrétiser les coordonnées de la cellule où est réalisé le signe SW, ou sinon à choisir un encodage simplifiant les comparaisons de symboles (Aznar, Dalle, 2004). Toutefois ces solutions ne permettraient pas de supprimer la variabilité inter-personnelle, et n'abaisseraient, sans la supprimer, que la variabilité intra-personnelle. La solution donc proposée est un agent de saisie, proposant et positionnant automatiquement les symboles dès que le signe est reconnu, puis sauvegardant à la fois le signe SW tel qu'édité et le signe SW standard, mis en métadonnée. Le signe édité est affiché pour faciliter sa modification éventuelle, le signe standard étant utilisé pour toute autre opération. Cette solution nécessite d'étudier des corpus de documents SW pour en extraire manuellement les signe édités. Ensuite, ils doivent être comparés aux signes standards correspondants tels qu'issus d'un dictionnaire SW (Roald 2004). Enfin, le format d'encodage doit supporter des métadonnées, pour encoder et gérer à la fois la version réalisée et la version standardisée.

3.1 Inconvénients: difficulté logicielle, nécessité d'un dictionnaire

Le principal inconvénient est la difficulté de reconnaissance d'un symbole – aussi bien en cours de saisie pour proposer les signes de bases le composant, qu'en fin de saisie pour proposer un symbole standard si nécessaire. Une méthode logicielle robuste sera nécessaire. De plus, cette solution nécessite un dictionnaire recommandant des formes standards pour les symboles SW. Les dictionnaires existants ne se prêtent pas tous à cette manipulation, dans la mesure où les symboles correspondant à des signes sont parfois stockés sous forme de simples images bitmap. Enfin, une adaptation des logiciels utilisateurs sera à prévoir, pour supporter un encodage gérant la dualité entre symbole présenté / symbole utilisé (métadonnée).

3.2 Avantages: réutilisabilité dans d'autres buts et d'autres formalismes

La saisie est suivie d'une éventuelle interaction avec l'utilisateur,ce qui lui offre l'opportunité de saisir des métadonnées complétant ainsi ce signe. Ces dernières peuvent amener diverses précisions sur le signe, approche souhaitée (Crasborn et al, 2004) mais encore non normalisée. Ces métadonnées pourraient être utilisées afin de simplifier le problème de recherche dans un texte. Pour l'étude linguistique de documents SW, un autre avantage se présente: les versions et les positionnements des symboles seraient conservés au sein des documents. Un corpus SW ainsi constitué pourrait être utilisé dans la réalisation du système de prédiction à l'entrée, ou

pour toute autre étude statistique sur SW. De plus, le support des métadonnées dans les documents SW pourrait être étendu afin de supporter leur annotation dans un but pédagogique (Schilit et al, 1998). Enfin, comme le problème semble se poser pour toute écriture bidimensionnelle, l'approche duale serait réutilisable dans un autre formalisme ou pour l'annotation de vidéos selon plusieurs langues et formalismes.

4 Conclusion

L'approche ici proposée permet de déplacer la complexité de la phase de reconnaissance vers l'étape de saisie, afin d'alléger l'exécution de fonctions de recherche ou de manipulations des documents SW. Elle permet surtout d'enrichir de métadonnées le signe édité. La possibilité d'utilisation à posteriori de ces métadonnées est aussi très intéressante. La principale difficulté est la réalisation de ce système, ne pouvant s'appuyer sur aucune métadonnée préexisante.

Références

Sutton V., Gleaves R. (1995), *SignWriter - The world's first sign language processor*, La Jolla , Ed. Center for Sutton Movement Writing

Prillwitz, S. et al. (1987), *HamNoSys. Hamburg Notation System for Sign Languages. An introduction*, Hamburg, Ed. Zentrum für Deutsche Gebärdensprache.

Da Rocha A., Dimuro G. (2003), *SignWriting and SWML: Paving the way to sign language processing*, Actes de TALN Workshop In O. Streiter, editor

Aznar G, Dalle P. (2004), *Computer Support for SignWriting Written Form of Sign Language,* Actes du Workshop RPSL, LREC 2004, p109-110

Aerts S., Braem B.,Van Mulders K., De Weerdt L. (2004), *Searching SignWriting Signs*, Actes du Workshop RPSL, LREC 2004, p79-81

Da Rocha A., Dimuro G., De Freitas J. (2004), *A sign matching technique to support searches in sign language texts*, Actes du Workshop RPSL, LREC 2004, p32-34

Huenerfauth M. (2002) *Natural Language Generation and Machine Translation for ASL, CIS-899 Independent Study Report*, Philadelphia, University of Pennsylvania

Knoblich G., Seigerschmidt E., Flach R., et Prinz, W. (2002). *Authorship effects in the prediction of handwriting strokes.* Q. J. Exp. Psychol. A 55, 10271046.

Roald I., (2004) *Making Dictionaries of Technical Signs: from Paper and Glue through SW-DOS to SignBank*, Actes du Workshop RPSL, LREC 2004, p75-78

Crasborn O., Kooij E., Broeder D., Brugman H. (2004) *Sharing sign language corpora online:proposals for transcription & metadata* ,Actes du Workshop RPSL,LREC 2004,p20-23

Schilit B., Golovchinsky G. et Price M. (1998), *Beyond Paper: Supporting Active Reading with Free Form Digital Ink Annotations*, Actes de ACM CHI 98, v.1 249-256

EASy

EQueR

DEFT

Langues Peu Dotées

TALS