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.