

Journée ATALA

Interface lexique-grammaire et lexiques syntaxiques et sémantiques

Les ressources lexicales du LADL: leur utilisation dans un contexte d'analyse syntaxique

Olivier Blanc, Matthieu Constant, Javier Sastre
Institut Gaspard Monge, laboratoire d'informatique
Université de Marne-la-Vallée

Introduction / Résumé

Dans cet exposé, nous souhaitons présenter les différentes ressources lexicales accumulées depuis une trentaine d'années au LADL par les équipes de Maurice Gross, dans plusieurs langues : dictionnaires morphosyntaxiques (mots simples, mots composés figés et semi-figés), dictionnaires syntaxiques pour les prédicats verbaux, adjectivaux et nominaux et dictionnaires de phrases figées (M. Gross, 1975, 1994; Bibliographie générale du LADL). Dans un souci d'ouverture vers la communauté, ces données sont en train d'être converties dans un format standard XML et vont être librement diffusées totalement ou partiellement.

Dans une deuxième partie, nous présentons un analyseur syntaxique alimenté par une grammaire lexicalisée compilée à partir de ces ressources. A cet effet, nous décrivons le formalisme utilisé qui est une grammaire de constituants décrits avec des automates récursifs augmentés de contraintes d'unification. Un système de pondération sur les automates permet au grammairien de donner une priorité à certains chemins par rapport à d'autres (ex. règle lexicale favorisée par rapport à une règle générale) et ainsi de choisir une analyse parmi plusieurs analyses candidates. Afin d'être utilisables, les ressources lexicales ont besoin d'être compilées dans un format directement exploitable par le parseur. Nous utilisons les méthodes décrites par E. Roche (1993).

A- Les ressources lexicales

A.1- Présentation des ressources

Les dictionnaires morphosyntaxiques se trouvent soit sous la forme de listes pour décrire des mots simples et des mots composés, soit sous la forme de graphes (ou grammaires locales) pour décrire des ensembles de variantes plus complexes d'un même lemme. Chaque entrée lexicale comprend plusieurs types d'informations : une forme fléchie, une forme canonique, une partie du discours, des informations morphologiques telles que le genre, le nombre, le temps de conjugaison, etc., des traits sémantiques génériques pour les noms (humain, concret, nom propre, toponyme, ...), plus quelques traits syntaxiques de base pour les verbes (transitif, pronom réfléchi obligatoire, ...). Une description plus fine du comportement syntaxique et de la sous-catégorisation des éléments prédicatifs est faite dans des dictionnaires syntaxiques.

Ces dictionnaires (ou tables de lexique-grammaire) décrivent les comportements syntaxiques pour chaque prédicat de manière exhaustive :

- nombre et nature des arguments (ex. complétive, infinitive, groupe nominal humain, ...)
- les prépositions appropriées
- les transformations acceptées (ex. passif, construction croisée, effacement d'un argument, ...)
- résolution de co-références (ex. entre le sujet d'une infinitive et un argument du verbe principal)

Ces propriétés sont codées sous la forme de tables (ligne : entrée lexicale ; colonne : propriété).

Il n'existe pas de système de description de sens formalisée. Cependant, les différents emplois (ou sens) pour une même valeur lexicale font l'objet de différentes entrées distinguées à partir de critères syntaxiques formels. Par exemple, le verbe *porter* comporte une quinzaine d'entrées parmi lesquelles:

Max se porte bien
Luc porte une cravate
La discussion porte sur ce sujet
Luc porte la valise dans la voiture

Il en est de même pour les noms et les adjectifs prédicatifs. Par exemple,

Max fait la fête
Max fait sa fête à Luc
 ...

Cette description de la langue est complétée par un dictionnaire de phrases figées de 30 000 entrées (français).

A.2- Normalisation et diffusion des ressources

La construction de ressources linguistiques et leur utilisation est en pleine explosion dans le domaine du Traitement Automatique des Langues. Ainsi, afin de faciliter les échanges dans la communauté, il est important de mettre en place des standards (ex. EAGLES, RNIL). Le formatage des données en XML semble être devenu une règle. Des architectures de système d'échanges des données ont même été conceptualisées (L. Romary, 2000).

Dans cette nouvelle optique, l'équipe de linguistique informatique de l'IGM, Université de Marne-la-Vallée, a mis en place une politique de normalisation et de diffusion des ressources lexicales du LADL. Des outils de conversion des ressources lexicales du format existant à un format XML ont récemment été implantés. Il existe également un éditeur de tables de lexique-grammaire en XML. Aussi, la diffusion des ressources est devenue une priorité : une partie des dictionnaires sont distribués via le logiciel libre Unitex et certaines tables du lexique-grammaire sont visualisables et téléchargeables à partir du site web du laboratoire. La diffusion des grammaires locales constitue la prochaine étape logique de cette politique.

B- Un exemple d'utilisation: un analyseur syntaxique alimenté par ces ressources

B.1- Formalisme

Historiquement, au LADL, la mise en place de systèmes d'analyse syntaxique automatique de textes avec grammaires lexicalisées a commencé avec M. Salkoff (1973), A. Abeillé (1991) [LTAG], M.

Mohri (1993) et E. Roche (1993) [transducteurs]. Le système INTEX (M. Silberztein, 1993) a permis de mettre en place une plate-forme commune à la communauté du lexique-grammaire. La représentation des grammaires sous la forme de graphes et de réseaux récursifs de transitions (RTN) a alors été adoptée pour l'analyse de textes. Ces graphes sont des automates à états finis dont les transitions peuvent être étiquetées soit par des éléments lexicaux, des références à des ensembles lexicaux à travers l'utilisation d'un dictionnaire morphosyntaxique et des références à des sous-graphes. Bien que ce formalisme ne soit pas le plus puissant (cf. TAG, HPSG), sa simplicité a permis une large utilisation pour l'analyse de textes dans des domaines spécialisés, avec succès (T. Nakamura, 2004).

Dans cet exposé, nous proposons de présenter un analyseur syntaxique utilisant une grammaire lexicalisée au formalisme plus évolué. Afin de palier aux limites théoriques inhérentes aux RTN, nous avons fait évoluer nos grammaires vers des grammaires à structures de traits. Nous conservons la simplicité initiale du système de graphes, tout en augmentant les grammaires de contraintes d'unification ; notre formalisme se rapproche ainsi du modèle PATR développé par S. Shieber (1986) avec la différence que les règles de réécriture sont remplacées par des descriptions linguistiques représentées par des automates finis. Ces derniers permettent de mettre très simplement en relation les différentes possibilités de réalisation de chaque constituant de la grammaire. Les contraintes d'unification permettent de résoudre de manière homogène différents phénomènes linguistiques tels que les contraintes d'accord, les phénomènes d'extraction et les dépendances non bornées ou encore la résolution de certaines co-références.

Dans un but plus applicatif, nous avons mis en place un système de pondération artisanale sur les automates permettant de donner des notes (ou scores) aux analyses obtenues. Le grammairien pourra, par exemple, favoriser, parmi un ensemble d'analyses candidates, les analyses où tous les arguments du prédicat sont identifiés (problème de l'attachement prépositionnel) ou encore favoriser des analyses de façon ad hoc à partir d'observations empiriques.

A titre d'exemple, le graphe de la figure 1 présente différentes réalisations de phrases ayant pour prédicat principal le verbe *empêcher* dans son emploi décrit dans la table 12 du lexique-grammaire (Gross, 1975).

La partie gauche décrit la possibilité d'avoir un sujet sous forme de groupe nominal, complétive ou infinitive :

(Lea|Que Lea ait quitté Max|sortir en boîte) empêche Luc de dormir.

L'étiquette <:V> dans la partie centrale fait référence à un sous graphe qui décrit le noyau verbal de la phrase (c'est-à-dire le verbe éventuellement modifié par des adverbes ou des auxiliaires modaux et aspectuels).

Enfin la partie droite présente les différentes réalisations du complément N1 pour le verbe *empêcher* : groupe nominal prédicatif ou complétive au subjonctif.

La ligne du bas décrit la possibilité de monter le sujet de la complétive en position d'objet direct, qui est une transformation acceptée par tous les verbes de la table 12.

La neige empêche que les gens sortent = La neige empêche les gens de sortir.

Les contraintes d'unification sous les boîtes permettent :

- d'identifier les arguments N0 et N1 du prédicat et de vérifier que leur nature est compatible avec les contraintes de sous-catégorisation,
- d'imposer l'accord entre le verbe et son sujet,
- et de résoudre certaines co-références en rétablissant le sujet des infinitives.

B.2- Construction d'une grammaire lexicalisée

Nous sommes actuellement en train de construire une grammaire lexicalisée pour le français suivant le formalisme décrit précédemment, générée semi-automatiquement à partir des tables du lexique-grammaire ; nos travaux sont en partie inspirés de ceux d'Anne Abeillé (2002).

Pour chaque élément prédicatif, nous décrivons les différentes réalisations des constituants syntaxiques (phrase, infinitive, phrase privée d'un argument, groupe nominal pour les noms prédicatifs, etc.) dont il est le noyau.

Afin d'automatiser ce procédé, nous construisons manuellement pour chaque table une meta-grammaire constituée d'un ensemble de graphes paramétrés ; cette grammaire paramétrée consiste en la grammaire d'une entrée fictive de la table qui vérifierait toutes les propriétés qui y sont encodées. Chaque chemin de la grammaire est identifié par un paramètre référant à la propriété correspondante dans la table.

A partir de cet ensemble de graphes paramétrés, on génère automatiquement pour chaque entrée une grammaire spécialisée dans laquelle seuls les chemins correspondant aux propriétés vérifiées sont conservés.

C'est selon ce procédé que le graphe de la figure 1 a été généré à partir du graphe paramétré présenté dans la figure 2.

En l'état actuel de nos travaux, la couverture de la grammaire est faible. Nos résultats sont donc partiels mais, dans l'ensemble, encourageants.

Références

Abeillé, Anne, 1991, *Une grammaire lexicalisée d'arbres adjoints pour le français : application à l'analyse automatique*, thèse de doctorat, Paris, Université Paris 7.

Abeillé, Anne, 2002. *Une grammaire électronique du français*, CNRS Editions, Paris.

Constant, Matthieu, 2003, *Grammaires locales pour l'analyse automatique de textes : Méthodes de construction et outils de gestion*, Thèse de doctorat, Université de Marne la Vallée.

Gross, Maurice, 1975, *Méthodes en syntaxe*, Hermann, Paris.

Gross Maurice, 1994, Constructing Lexicon-grammars, In *Computational Approaches to the Lexicon*, Atkins and Zampolli (eds.), Oxford Univ. Press, pp. 213-263

Mohri, Mehryar, 1993, *Analyse et représentation par automates de structures syntaxiques composées : applications aux complétives*, Thèse de Doctorat, Paris, Université Paris 7.

Nakamura, Takuya, 2004, "Analyse automatique d'un discours spécialisé au moyen de grammaires locales", *Le poids des mots : Actes des 7èmes Journées internationales d'analyse statistique des données textuelles*, Purnelle G., Fairon C. et Dister A. (eds.), UCL Presse universitaire de Louvain, pp. 837-847.

Pollard C. and I.A. Sag (1994), *Head-Driven Phrase Structure Grammar*, University of Chicago Press and CSLI Publications.

Roche, Emmanuel, 1993, *Analyse syntaxique transformationnelle du français par transducteurs et lexique-grammaire*, Thèse de Doctorat, Paris, Université Paris 7.

Romary, Laurent, 2000, *Outils d'accès à des ressources linguistiques*, In J.M. Pierrel (ed.), *Ingénierie des Langues*, Hermes Science, Paris

Salkoff, Morris. 1973, *Une grammaire en chaîne du français. Analyse distributionnelle*, Paris: Dunod.

Schabes, Yves, Anne Abeillé and Aravind K. Joshi, 1988, *Parsing strategies with 'lexicalized' grammars: Application to tree adjoining grammars*, In *Proceedings of the 12 International Conference on Computational Linguistics (COLING'88)*, Budapest, Hungary, August 1988.

Shieber, Stuart, 1986, *An introduction to unification-based theories of grammar*, CSLI, University of Chicago Press.

Silberztein, Max D., 1993, *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX*, Paris, Masson, 234 p.

Sites

Bibliographie générale du LADL, <http://infolingu.univ-mlv.fr>

EAGLES, <http://www.ilc.cnr.it/EAGLES96/home.html>

RNIL, <http://atoll.inria.fr/RN>

ANNEXE

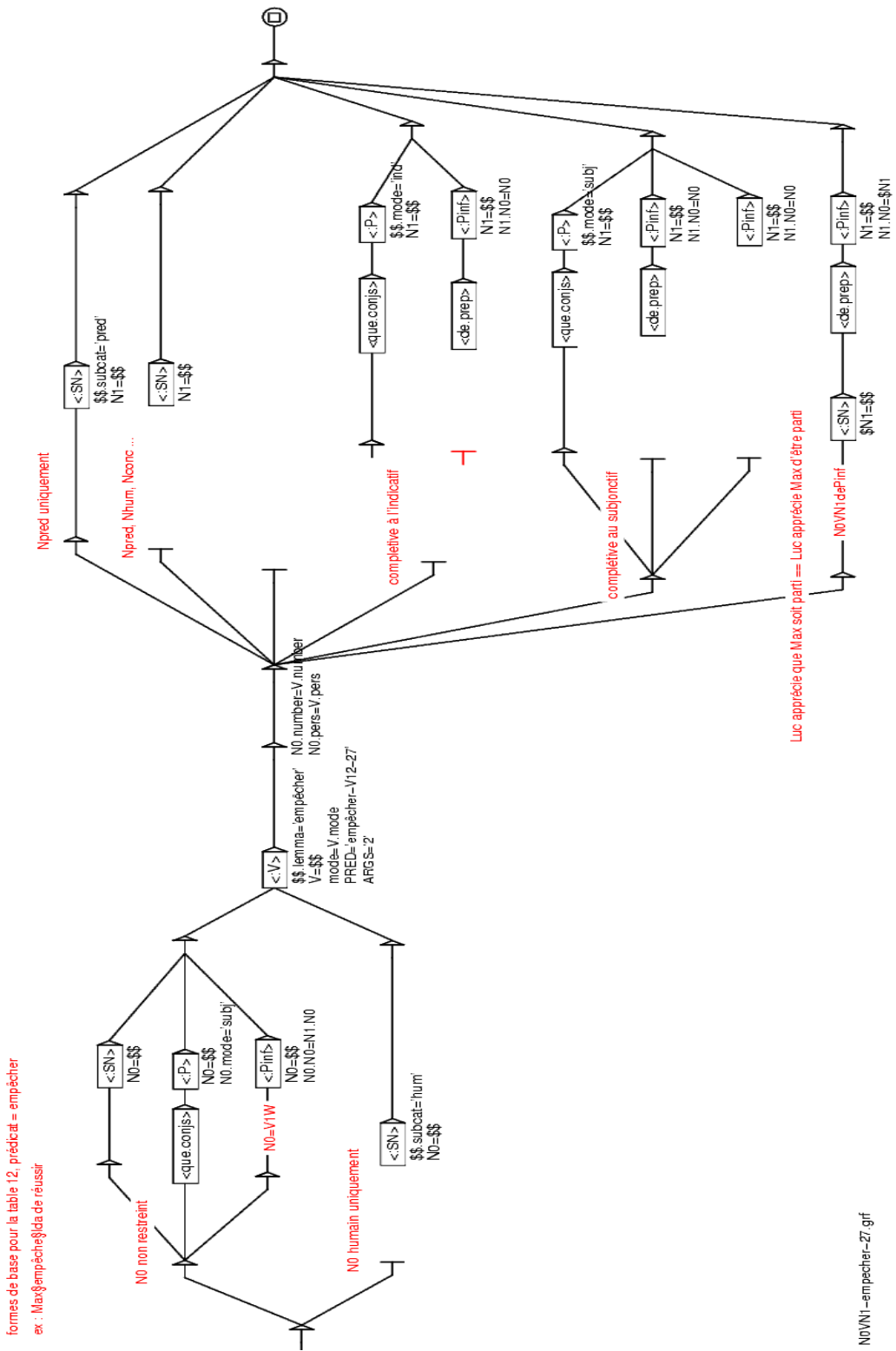


Figure 1 - Formes de base pour le prédicat *empêcher*

