

Rapport d'avancement DFD - 2^e année de thèse

Yannick PARMENTIER
Projet Langue Et Dialogue

12 septembre 2005

Sujet de thèse

Titre : « Développement d'une architecture logicielle pour le traitement sémantique des langues naturelles à partir de Grammaires d'Arbres Adjoints ».

Problématique : A l'heure actuelle, il n'existe pas de grammaire d'arbres adjoints (*Tree Adjoining Grammars - TAG*) à large couverture permettant d'annoter une phrase non seulement avec sa structure syntaxique mais également avec sa représentation sémantique (i.e. une représentation du sens de cette phrase, généralement une formule appartenant à un langage logique).

Ce sujet de thèse vise à spécifier, développer et évaluer des outils informatiques permettant le traitement sémantique basé sur les TAGs. Plus précisément, nous souhaitons mettre en place une architecture permettant notamment :

- (1) la production semi-automatique d'une grammaire TAG à portée sémantique pour le Français,
- (2) la validation de cette grammaire,
- (3) la construction automatique, par le biais de cette grammaire, d'une représentation du sens d'énoncés en langue naturelle.

Travail Réalisé

En 1^{ère} année de thèse, mes travaux se sont concentrés sur le développement d'un *compilateur de méta-grammaire* à portée sémantique. L'intérêt principal d'un tel outil est de pouvoir produire automatiquement une grammaire de taille relativement importante à partir d'une description réduite, tout en évitant les problèmes d'écriture et de maintenance.

Ce travail a donné lieu à l'enregistrement du logiciel multi-plateforme XMG¹ auprès de l'INRIA, sous licence CeCILL. En outre, suite aux commentaires reçus d'utilisateurs potentiels, nous avons rédigé une documentation relativement complète du logiciel².

L'un des enjeux de la 2^{nde} année de thèse était de réaliser un interfaçage entre XMG et un analyseur syntaxique, dans le but non seulement de valider les grammaires produites mais également de mettre en place un procédé de construction sémantique basé sur l'analyse syntaxique.

Le premier point, la validation de grammaires produites automatiquement, a été réalisé dans le cadre de la thèse de Benoît Crabbé au LORIA.

Le second point a donné lieu au développement de méthodes et outils spécifiques pour réaliser une construction sémantique à partir du résultat de l'analyse syntaxique dans le cadre des grammaires d'arbres adjoints. Cette construction sémantique peut être réalisée soit (a) pendant l'analyse syntaxique en annotant les arbres de la grammaire avec des informations sémantiques qui seront unifiées lors des dérivations, ou (b) après l'analyse syntaxique en exploitant la forêt d'analyse et un lexique sémantique issu de la grammaire produite automatiquement. Ces deux procédés de construction sémantique sont en cours de développement. Plus précisément dans la 2^{nde} approche, j'ai développé des programmes permettant la combinaison de l'information sémantique contenue dans les grammaires TAG avec les informations issues du lexique syntaxique (i.e. prédicats et rôles thématiques) afin de construire un lexique sémantique. Ce lexique est ensuite utilisé accompagné de la forêt d'analyse par un programme

¹ disponible librement à l'adresse <http://sourcesup.cru.fr/xmg>.

² visible à l'adresse <http://wiki.loria.fr/XMG/Documentation>

de construction sémantique développé par Claire Gardent pour construire une formule de sémantique plate représentant le sens de l'énoncé.

Perspectives de travail pour la 3^e année de thèse

Actuellement, j'ai réalisé un prototype de module de construction sémantique qui fonctionne avec une grammaire et un lexique de taille moyennes. Le prochain objectif consiste à évaluer la robustesse du système sur une suite de tests telle que le TSNLP et de comparer les résultats obtenus au moyen d'un générateur³.

Le principe de cette évaluation est le suivant :

- utiliser le module de construction sémantique couplé avec un analyseur syntaxique pour construire la représentation sémantique d'un certain nombre d'énoncés,
- transmettre ces représentations à un générateur muni des mêmes grammaires et lexiques,
- vérifier la présence de l'énoncé de départ dans l'ensemble des structures syntaxiques produites par le générateur. Dans le cas contraire, nous avons mis en évidence (a) une erreur dans le module de construction sémantique ou (b) une erreur dans les ressources utilisées (grammaire - lexique).

En outre voici une première ébauche du plan de mon mémoire de thèse :

1. Production de ressources adaptées à la construction sémantique au moyen de Grammaires d'Arbres Adjoints : interface syntaxe/sémantique dans les Méta-Grammaires d'arbres adjoints
2. Deux procédés de construction sémantique pour grammaire d'arbres adjoints de taille réelle : pendant l'analyse syntaxique ou a posteriori à partir de la forêt de dérivation.

Publications

- XMG : a Multi-formalism Metagrammatical Framework. Yannick Parmentier and Joseph Le Roux. *Student Session of the 17th European Summer School in Logic, Language and Information (ESSLLI)*, Edinburgh, august 2005.
- Large scale semantic construction for Tree Adjoining Grammars. Claire Gardent and Yannick Parmentier. *Logical Aspect of Computational Linguistics (LACL)*, Bordeaux, april 2005.
- The Metagrammar Compiler : An NLP Application with a Multi-paradigm Architecture. Denys Duchier, Joseph Le Roux and Yannick Parmentier. *2nd international Oz / Mozart conference*, Charleroi, october 2004.
- XMG : Un Compilateur de Méta-Grammaires Extensible. Denys Duchier, Joseph Le Roux et Yannick Parmentier. *12^e conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Dourdan, juin 2005.

Formations suivies dans le cadre du Département de Formation Doctorale

- participation à la 16^e Université d'Été en Logique, Langage et traitement de l'Information : *ESSLLI 2004*, Nancy : cours sur les algorithmes d'analyse syntaxique tabulaires.
- participation à la 15^e Université d'Été en Logique, Langage et traitement de l'Information : *ESSLLI 2003*, Vienne : cours sur les grammaires d'unification et la syntaxe formelle (*model-theoretic syntax*).
- module PRTAL 3 du DEA d'Informatique de Lorraine (année 2003/04) : introduction aux algorithmes d'analyse syntaxique pour différents formalismes, parmi lesquels grammaires hors-contexte et grammaires d'arbres adjoints.
- module PRTAL 4 du DEA d'Informatique de Lorraine (année 2003/04) : introduction au traitement automatique de la sémantique des langues naturelles (étude des représentations logiques et des mécanismes de composition sémantique).
- module d'Anglais niveau 1, dispensé par le CRELENS (année 2003/04) : présentation écrite et orale de travaux scientifiques.

³Un générateur est un outil prenant en entrée une formule sémantique et une grammaire et produisant en sortie un ensemble de réalisations syntaxiques correspondantes.