

Rapport d'avancement - 1^{ère} année de thèse

Yannick PARMENTIER
Projet Langue Et Dialogue

27 juillet 2004

Sujet de thèse

Titre : « *Développement d'une architecture logicielle pour le traitement sémantique de la langue naturelle à partir de Grammaires d'Arbres Adjoints* ».

Problématique : Pour développer des applications intégrant un raisonnement sémantique (telles que la tâche de question / réponse, ou celle plus restreinte d'extraction d'information,), il est important de disposer d'outils robustes, parmi lesquels nous comptons une grammaire.

Plusieurs formalismes grammaticaux décrivent la syntaxe de la langue naturelle, cependant nombreux sont ceux qui présentent certains inconvénients quant à leur emploi en traitement automatique des langues (TAL). Découvertes dans les années 70, les grammaires d'arbres adjoints (ou grammaires TAG) se sont révélées adaptées à un traitement informatique de par leurs qualités, dont principalement :

- ce sont des grammaires légèrement contextuelles,
- les expressions engendrées par ces grammaires sont analysables en un temps polynomial (Voir [6] pour une introduction détaillée).

A l'heure actuelle, il n'existe pas de grammaire d'arbres adjoints 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 l'étude de l'interface syntaxe / sémantique dans le cas des grammaires TAG, et plus particulièrement au développement d'outils informatiques pour le traitement sémantique basé sur les TAGs.

Travail Réalisé

Pour pallier aux problèmes d'écriture et de maintenance de grosses grammaires, un procédé de génération semi-automatique de grammaires a été proposé à la fin des années 1990 à l'Université Paris 7 : la compilation de méta-grammaires ([2]). Bien que la méthode proposée comporte des inconvénients (non monotonie par exemple), celle-ci a ouvert la voie à de nombreux travaux sur les méta-grammaires. Cela a été le cas au sein du projet Langue Et Dialogue, qui a lui même proposé un nouveau mécanisme de compilation de méta-grammaires (voir [4]).

Mon travail est dans la lignée de ces travaux, et s'articule plus précisément autour de ceux de Benoît Crabbé, Doctorant de l'équipe Langue et Dialogue. Celui-ci propose une nouvelle méthodologie pour la gestion de grammaires à larges couvertures ([3]).

Après avoir étudié les approches existantes dans le domaine des méta-grammaires, j'ai participé activement au développement d'une implémentation du procédé de B. Crabbé en collaboration avec Denys Duchier et Joseph Le Roux de l'équipe Calligramme¹.

A l'heure actuelle, nous avons implémenté un nouveau compilateur de méta-grammaires, qui permet notamment la génération de grammaires TAG à portée sémantique (suivant le format décrit dans [5]). Ce compilateur a été développé au moyen de techniques utilisées dans le domaine de la programmation logique, telles que la Warren Abstract Machine (voir [1]), et est codé en langage Oz.

La version dont nous disposons est stable, et est utilisé par B. Crabbé pour valider ses théories. Les premiers résultats sont encourageants puisque nous parvenons à générer une grammaire du français contenant plus de 3 000 arbres.

En outre nous travaillons actuellement à l'extension de notre outil à d'autres formalismes grammaticaux tels que les grammaires d'interaction.

Perspectives de travail pour la 2nde année de thèse

Afin de compléter l'architecture logicielle pour le traitement sémantique de la langue naturelle correspondant à mon sujet de thèse, j'ai commencé l'étude de l'interfaçage entre notre compilateur de méta-grammaires et des analyseurs syntaxiques (système DyALog et analyseur LLP2 principalement).

L'interfaçage sur lequel je travaille, consiste en la récupération de grammaires TAG produites semi-automatiquement par notre compilateur de méta-grammaires pour les transmettre au système DyALog afin d'obtenir un outil permettant l'analyse syntaxique et sémantique d'expressions en langue naturelle. Il y a deux difficultés à surmonter dans cette opération :

1. la première a trait au format de représentation des données,
2. la seconde consiste en l'extension du système DyALog afin d'intégrer un support du traitement sémantique.

Concernant le 1^{er} point, nous allons utiliser le langage XML, qui est un standard reconnu dans les problématiques d'échange de données.

A ce jour, nous avons réalisé un premier interfaçage entre DyALog et une grammaire générée automatiquement par notre outil (visible à l'adresse <http://atoll.inria.fr/parserdemo>).

Concernant le 2nd point, cela demande une étude plus approfondie du système DyALog et fait parti des travaux à mener prochainement.

Enfin, nous avons pu interfacé une grammaire du français à large couverture avec l'analyseur LLP2 développé au sein de l'équipe Langue et Dialogue par Azim Roussanally. Cela grâce aux travaux d'Azim Roussanally et de Jérôme Perrin (stagiaire de Maîtrise MIAGE).

¹Un groupe de travail sur le thème des méta-grammaires a été mis en place en février 2003 sous l'impulsion de Denys Duchier (Projet Calligramme) et Claire Gardent (Projet LED).

Publications

- *The Metagrammar Compiler : An NLP Application with a Multi-paradigm Architecture*. Soumis à la 2nde conférence internationale Mozart / Oz, Charleroi, octobre 2004.
- *Un compilateur de méta-grammaires*. Séminaire quasi-annuel de l'équipe Langue Et Dialogue (5–7 mai 2004, Madine)².

Formations suivies dans le cadre du DFD

- participation à la 15^e Université d'Eté en Logique, Linguistique et traitement de l'Information : *ESSLLI 2003*, Vienne.
- stage de *formation en Anglais niveau 1* dispensé par le *CRELENS* en février et juin 2004 (30 heures de formation).
- Modules PRTAL 3 et PRTAL 4 du DEA Informatique de Lorraine de l'Université Henri Poincaré – Nancy 1 (2003–2004).

Activités annexes

Parallèlement à mes activités premières de recherche, je suis moniteur en informatique à l'Université Henri Poincaré–Nancy 1 (UFR STMIA).

Je fais également parti du Comité Local d'Organisation de la 16^e Université d'Eté sur la Logique, Linguistique et le traitement de l'Information (*ESSLLI 2004*) qui se tiendra à Nancy du 9 au 20 août prochain.

Références

- [1] H. Ait-Kaci. Warren's abstract machine : A tutorial reconstruction. In K. Furukawa, editor, *Logic Programming : Proc. of the Eighth International Conference*, page 939. MIT Press, Cambridge, MA, 1991.
- [2] M.H. Candito. *Représentation modulaire et paramétrable de grammaires électroniques lexicalisées : application au français et à l'italien*. PhD thesis, Université Paris 7, 1999.
- [3] B. Crabbé. Lexical classes for structuring the lexicon of a tag. In *Proceedings of the Lorraine/Saarland workshop on Prospects and Advances in the Syntax/Semantics Interface*, 2003.
- [4] B. Gaiffe, B. Crabbé, and A. Roussanally. A new metagrammar compiler. In *Proceedings of the 6th International Workshop on Tree Adjoining Grammars and Related Frameworks (TAG+6), Venice*, 2002.
- [5] C. Gardent and L. Kallmeyer. Semantic construction in ftag. In *Proceedings of the 10th meeting of the European Chapter of the Association for Computational Linguistics, Budapest*, 2003.
- [6] A. Joshi and Y. Schabes. Tree-adjoining grammars. In G. Rozenberg and A. Salomaa, editors, *Handbook of Formal Languages*, volume 3, pages 69 – 124. Springer, Berlin, New York, 1997.

²transparents accessibles à l'adresse <http://www.loria.fr/equipes/led/actu.php#11>.