Titre

Définition formelle d'un formalisme linguistique pour le Traitement Automatique des Langues

Laboratoire

LaBRI - Equipe Méthodes Formelles - Thème Linguistique Informatique

Directeur

Lionel Clément (co-directeur)
lionel.clement@labri.fr

Directeur habilité

Christian Retoré

Description du Sujet

Motivations

Le traitement automatique des textes tout-venant se heurte régulièrement à quelques cas complexes pour lesquels il est indispensable d'avoir une bonne analyse des phrases en syntaxe et en sémantique. C'est le cas des constructions causatives et factitives, des constructions à verbes supports, des coordonnées, corrélatives et comparatives, des phénomènes d'ellipse, et de bien d'autres cas encore.

Nous pouvons constater que les techniques développées en Traitement Automatique des Langues, et plus encore en *Data Mining* se passent le plus souvent de l'analyse fine en syntaxe et en sémantique pour ne retenir que l'exploitation de méta-données ou au mieux, l'analyse en surface des textes (*Chunking, Shallow-parsing*). Nous observons que les techniques d'analyse robuste en TAL se passent également des connaissances linguistiques fines de la syntaxe et de la sémantique.

Ces observations de l'état de l'art de l'ingénierie des langues montrent la difficulté à retenir les formalismes linguistiques pour des applications informatiques. Cependant certains de ces formalismes, comme "*Lexical-Functional Grammar*" (LFG) ou "*Head-Driven Phrase Structure Grammar*" (HPSG) sont très utilisés dans la communauté pour décrire la langue et pour construire des grammaires.

La raison de cette aporie est simple: l'analyse des textes avec les grammaires de ce type est un problème NP-complet.

Objectif

L'étude sera menée dans un cadre de linguistique formelle issue d'un modèle récent largement utilisé par les linguistes comme LFG ou HPSG.

L'objectif visé est une redéfinition théorique du formalisme visant deux buts difficilement conciliables:

- Le premier est le respect du pouvoir descriptif du modèle théorique. Les phénomènes linguistiques doivent pouvoir être représentés, les principes réglant la régularité ou la variation en langue doivent s'exprimer naturellement dans le formalisme amendé avec peu de restrictions.

- Le second est l'intérêt pour le Traitement Automatique des Langues de la redéfinition du formalisme. La construction du langage engendré doit être un problème de complexité polynomiale, ou pour lesquels des heuristiques bien définis pourront être satisfaisants dans le cadre d'une application du TAL (génération automatique de texte, analyse syntaxique, calcul sémantique, etc).