Introduction

"Il y a en même temps une mauvaise adaptation du monde au langage, et de l'être pensant au monde."
Jean Tardieu

Ebauche pour un traitement automatique de la construction du sens

Le travail présenté ici s'inscrit dans le cadre du développement de dispositifs pour le TALN qui permettent une analyse critique des conjectures dont les analyseurs automatiques étudiés donnent une image.

Au commencement, une histoire de traces

L'origine de ce travail est né à la suite de l'utilisation d'analyseurs automatiques au cours de mon année de DEA : XTAG (Paroubek & al. 1992) et LN-2-3 (Zweigenbaum 1992). L'utilisation de ces analyseurs dans le cadre d'un séminaire sur les formalismes grammaticaux et certains outils informatiques associés a fait apparaître quelques insatisfactions. Ces logiciels ne fournissent pas d'interface satisfaisante entre l'utilisateur et le processus d'analyse en cours. L'utilisateur n'a aucun moyen d'intervenir au cours de l'analyse. Quand l'analyse réussit, l'analyseur propose une représentation graphique de la dérivation construite. Par contre quand elle échoue, la réponse rendue par l'analyseur est loin d'être satisfaisante. Dans le cas de XTAG, cette réponse peut d'ailleurs résulter soit de la donnée d'une phrase mal construite, soit de la donnée d'une phrase contenant des erreurs orthographiques ou un item lexical inconnu par l'analyseur. GASPAR (Fleury 1992) constitue une approche pour l'étude des problèmes de représentation des connaissances dans un analyseur et pour le développement de processus capables de renseigner l'utilisateur sur l'activité de l'analyseur au cours d'une analyse. Une partie de ce travail a permis de proposer une représentation possible des connaissances syntaxiques reposant sur les cadres offerts par la programmation à objets. Ce travail a aussi permis de plus de mettre en valeur le problème crucial mais non trivial qui est celui de construire un méta-langage cohérent pour exploiter les ressources mises en jeu lors d'une analyse et qui réponde aux besoins de l'utilisateur.

Dans le cadre du travail présenté ici, notre tâche a consisté à poursuivre l'entreprise permettant de rendre "lisible" les activités d'une analyse : on verra que cet objectif est atteint dans le dispositif `Gaspar` présenté ici. Cette lisibilité doit permettre aux utilisateurs des outils de l'analyse de visualiser/manipuler les entités présentes au cours d'une analyse sous la forme "linguistique" qu'ils connaissent avant d'avoir recours au dispositif qu'ils vont utiliser.

On doit donc disposer d'outils qui permettent une identification des dispositifs techniques mis en place pour l'analyse et les phénomènes linguistiques représentés.

Le Traitement Automatique du Langage Naturel (TALN) : du dur désir de représenter

Le travail du linguiste est une recherche d'équilibre entre différentes contraintes et il **est condamné à rester provisoire** : les résultats sont toujours remis en question par la prise en compte de nouvelles informations (Milner 1989). Le TALN est donc une entreprise dont les résultats ne peuvent être que progressifs. On est encore loin d'atteindre un Traitement Automatique en Langage Naturel intégral. Les travaux existants ou à venir constituent, semble-t-il, une infime partie du parcours à accomplir pour atteindre ce but. Il n'est d'ailleurs pas certain que ce but ultime soit envisageable en raison justement de ce que *"la donnée minimale de langue, qui est une phrase, est toujours trop complexe par rapport à la proposition minimale de linguistique"* (Milner 1989). La difficulté principale consiste à cerner l'objet de la science linguistique et a fortiori du TALN. L'incertitude quant à la réussite de cette tâche ne permet pas d'entrevoir une issue définitive et proche pour la mise en place d'un traitement automatique sur un objet parfaitement déterminé. Il semble bien d'ailleurs qu'il soit impossible de résoudre les problèmes liés au traitement automatique en ne restreignant pas son champ d'application.

De la réalité des représentations : du renoncement à tout prévoir à l'expérimentation linguistique.

"Il se voyait clairement en artisan, en fabriquant, en 'fabbro' des déductions, en menuisier des propositions, des corrolaires, des 'scollies'…".
Jacques Roubaud, Mathématique:, Editions du Seuil, 1997.

Vouloir entreprendre par exemple de résoudre les problèmes liés à une analyse automatique intégrale du langage naturel sur un objet aussi hétérogène et complexe semble quelque peu déraisonnable dans l'état actuel de nos connaissances sur l'objet de la science linguistique, à moins de construire des dispositifs incomplets et qui ne couvrent pas l'intégralité des phénomènes linguistiques du langage naturel. Une couverture raisonnable de la syntaxe et du lexique, pour une analyse automatique, demande des moyens considérables. De plus cela nécessite un grain très fin de représentation (pour les modèles correspondants) et parallèlement des moyens considérables pour stocker et traiter des informations extrêmement complexes (Pitrat 1992).
Un dispositif propose une représentation matérielle des phénomènes traités. Les remarques faites précédemment permettent tout de suite de préciser qu'il s'agit d'étudier des dispositifs existants ou à venir qui ne constituent pas des monstres de technicité non manipulables/gérables par des individus qui ne sont pas de purs techniciens (Marandin 1993). Dans notre cadre de travail, les dispositifs mis en place ou étudiés sont constitués par des outils informatiques développés dans le cadre de la programmation à objets. Il s'agira d'établir qu'on a bien une adéquation forte entre les opérations de connaissance que ces outils permettent et les représentations de la langue. Et dans tous les cas, on sera amené à entreprendre une analyse critique des techniques utilisées pour la mise au point du dispositif (Marandin 1993) puisque le recours à une technique est déterminé par le choix de la meilleure référence à un moment donné capable de fournir le dispositif le plus clair et le plus distinct pour la représentation. Ce travail est indispensable dans le cadre de la construction de dispositif informatique développé à l'aide de langages de haut niveau si l'on veut satisfaire une des conditions du statut du dispositif choisi à savoir : fournir le dispositif le plus clair et le

plus distinct pour la représentation et ainsi permettre d'évaluer la cohérence entre le dispositif et les hypothèses théoriques suivies. On reprendra pour cela le travail présenté dans GASPAR (Fleury 1992) en tentant de résoudre ou d'améliorer les (pré-)résultats qui y étaient annoncés.

Quels outils informatiques pour la représentation des connaissances en TALN?

Ce travail vise à une réflexion sur la mise au point de la connaissance linguistique à utiliser pour l'analyse automatique. Celle-ci implique une adaptation, un affinage des connaissances linquistiques à utiliser lors de l'analyse. Le TALN pose aussi le problèmes du type d'architectures, du type de paradigmes computationnels, à utiliser ou à mettre en place pour définir un dispositif digne de produire la meilleure réponse au problème de représentation matérielle d'une activité linguistique. En particulier, le choix d'une représentation modulaire ou non des connaissances linguistiques est fondamentale pour le TALN. D'autant plus que l'utilisation de langages informatiques de haut niveau à architecture réflexive permet dans le cas d'une présentation modulaire des connaissances linguistiques d'envisager de résoudre des problèmes fondamentaux pour l'analyse automatique. A l'inverse, la non-modularité peut se révéler être une entrave à la progression de dispositifs cohérents et puissants (Habert & Fleury 1993a).
Les travaux menés en TALN depuis une quinzaine d'années ont longtemps privilégié les modèles de représentation fondées sur les taxonomies de classes. Depuis la fin des années 1980, les représentations hiérarchisées des connaissances se sont généralisées (Habert 1995):

- 1. La place croissante donnée au lexique dans les formalismes d'unification a nécessité l'utilisation d'outils pour gérer les redondances (DATR (Gazdar & Evans 1990), TAGs (Schaber & Vijay-Shanker 1992), etc.).

- 2. Les développements continus et parallèles des applications utilisant les langages à objets et des réseaux sémantiques ont convergé vers une représentation hiérarchisée des connaissances.

- 3. Enfin, les types de hiérarchie utilisés par les langages à objets et ceux utilisés en linguistique se sont rejoints sur un modèle de raisonnement non-monotone.

Dans le même temps, de nombreuses recherches sur la catégorisation (Rosch 1975-76-77-78) ont été menées en termes de prototypes. Les développements linguistiques de cette approche prototypique sont restés au niveau programmatique et n'ont pas débouché sur des projets descriptifs importants. La prédominance des approches fondées sur les taxonomies de classes restent de mise pour le TALN. Cela s'explique en partie par le fait que les classements utilisés en TALN reposent sur des années de tradition. De plus, le TALN se concentre principalement sur des domaines de savoir pour lesquels il y a le plus de recouvrement. Les domaines de savoir où la catégorisation est mouvante (non figée) se confrontent à deux types d'organisation : un mode apriorique (les taxonomies de classes), et un mode inductif (celui des prototypes). La volonté de modéliser les connaissances utilisées par le locuteur dans l'interprétation de séquences énonciatives nouvelles pose le problème crucial du mode de catégorisation le plus adéquat.

Quels objets pour représenter la mouvance ?

Pour illustrer ces phénomènes, on étudie dans un premier temps une partie privilégiée des formes dénominatives complexes, les `N1àN2` et plus particulièrement celles où `N1 = moulin` dans deux cadres de représentation:

- On présente tout d'abord un cadre de représentation (Habert & Fabre 1993) où l'on utilise des catégories prédéfinies et des modèles d'interprétation. Pour cela, une approche lexicale hiérarchisée est mise en place pour représenter les arguments possibles d'un prédicat sous-jacent. L'utilisation d'une hiérarchie de règles sert dans ce cadre à modéliser l'affinement progressif d'une interprétation.

- On utilise ensuite un cadre privilégiant une approche prototypique qui permet une classification évolutive des séquences visées.

Dans un deuxième temps, on montre comment la Programmation à Prototypes (i.e. PàP) (Liebermann 1986, Ungar & al. 1988, Self Group 1993-1996) permet de rendre compte d'une représentation dynamique des savoirs linguistiques et d'un classement automatique de ces savoirs (en utilisant le mécanisme de la délégation[1]). La PàP permet de représenter dynamiquement par apprentissage et par affinements successifs des unités lexicales en utilisant des informations recueillies sur corpus et elle permet aussi d'amorcer un début de classement de ces unités lexicales sur la base des contraintes syntaxiques attachées aux unités lexicales en construisant des hiérarchies locales de comportements partagés.

Description du travail

Plan de la thèse

Cette thèse comprend trois parties.

Partie 1 : De la réalité

Chapitre 1

Ce chapitre introduit la nécessité de construire des représentations évolutives pour modéliser les faits de langue. Les faits de langue sont évolutifs et non figés. Les processus de représentation de ces faits de langue doivent donc tenir compte de ces évolutions potentielles. Le travail de représentation doit construire des ébauches. Il faut ensuite ajuster les représentations initialement définies en tenant compte des informations que l'analyse dévoile. Pour la construction du sens, il s'agit de décrire au mieux les mécanismes qui opèrent pour réaliser cette mise en action de mécanismes complexes si on veut s'approcher au plus près de ces mécanismes.

Chapitre 2

Ce chapitre présente certains modèles qui traitent de la construction du sens et qui sont globalement cohérents par rapport à nos hypothèses de travail. Il vise surtout à mettre en lumière les difficultés rencontrées pour cette modélisation dans un domaine de savoirs mouvants. La construction du sens s'effectue par touches successives. De nombreux savoirs réalisent cette construction du sens, et ils interagissent de manière complexe. Les mécanismes qui participent à l'interprétation agissent sur les savoirs présents dans le dit, mais ils sont aussi capables de reconstruire des savoirs matériellement absents de cette chaîne du dit.

[1] Comportements partagés "localement" par un ensemble de prototypes (-> partie 2 chapitre 3) (Liebermann 1986).

Chapitre 3

Ce chapitre présente deux types d'outils disponibles pour mettre en oeuvre des processus de représentation (la Programmation à Objet dans une taxonomie de classes et la Programmation à Prototypes) et vise à établir une comparaison critique du pouvoir expressif de ces deux cadres de représentation. Dans un cas, il est plus facile de représenter des savoirs établis donc de fixer ces savoirs dans des hiérarchies. Dans l'autre cas, le processus de représentation est contraint par des savoirs non connus à l'avance : il est donc plus difficile de présumer de ces savoirs et de leur mode de structuration et d'organisation. Il s'agit donc d'évaluer les apports pertinents de chacun de ces cadres pour la représentation du langage naturel, ce dernier articulant très bien ces deux aspects de la manipulation des connaissances.

Chapitre 4

Ce chapitre insiste tout d'abord sur le fait que le TALN est confronté à des problèmes non triviaux quand il s'agit de donner une représentation matérielle de certains faits de langue en particulier quand il s'agit de construire un dispositif informatique qui tente de résoudre ou d'illustrer les problèmes liés aux traitements automatiques de ces faits de langue. Le traitement automatique doit passer par une phase de représentation des éléments linguistiques manipulés. Cette phase de représentation est contrainte de figer les savoirs à modéliser, de présumer de ce que les savoirs représentés peuvent réaliser.

Ce chapitre introduit ensuite la démarche de représentation qui sera suivie dans ce travail via la programmation à prototypes. Il présente enfin les premiers choix faits pour une représentation de la construction du sens des séquences nominales du type `moulin à N2`

Partie 2 : Du possible

Un cadre particulier de représentation : les prototypes.

Le choix d'une représentation à l'aide de prototypes vise à illustrer le problème de l'évolution des représentations que le traitement automatique de la construction du sens requiert. Une telle approche de représentation privilégie la définition des éléments représentés de manière contextuelle puis leur spécialisation en tenant compte des évolutions contextuelles rencontrées. En termes de représentation, les prototypes permettent donc d'envisager une définition minimale du savoir linguistique qu'il est ensuite possible d'affiner. Il conviendra ensuite de mettre en lumière les mécanismes qui permettent d'affiner ces savoirs initialement représentés.

Chapitre 5

Ce chapitre est une introduction pour la mise en place d'une représentation de savoirs linguistiques avec des prototypes dans la PàP et en particulier pour la mise au point d'outils pour la construction du sens des séquences `moulin à N2`. On prend appui ici sur un modèle semantique particulier présenté dans le chapitre 1.

Chapitre 6

Ce chapitre poursuit le travail de mise au point du dispositif qui traite de la construction du sens des séquences `moulin à N2`. La démarche de représentation suivie amène à reformuler le problème traité et à proposer une nouvelle solution pour la représentation des savoirs manipulés.

Chapitre 7

Ce chapitre étend le processus de représentation de savoirs linguistiques. Il ne s'agit plus de restreindre le processus de représentation à certains types de séquences nominales mais de prendre appui sur des corpus (de taille importante) pour représenter les savoirs linguistiques à l'oeuvre sur les mots de ces corpus. Cette nouvelle approche utilise des savoirs extraits sur corpus pour construire automatiquement des structures décrivant ces savoirs puis classe les savoirs ainsi representés. Ce travail de représentation est conditionné par une phase d'extraction de savoirs sur corpus realisé par des analyseurs syntaxiques traditionnels.

Chapitre 8

Ce chapitre présente les différents outils mis en place pour réaliser les objectifs établis dans le chapitre précédent. Il met aussi en avant les limites des processus mis en place et le travail qu'il reste à definir pour affiner les résultats construits.

Partie 3 : De la "Science Fiction"

Ce qu'il reste à faire : de la méta-connaissance à la réflexivité.

Chapitre 9

Les résultats obtenus dans les chapitres précédents ont mis en avant la nécessité de disposer de processus qui réalisent des méta-contrôles sur les opérations mises en oeuvre dans les dispositifs construits. Tout d'abord, ces processus doivent permettre de tracer les mécanismes fins et complexes mis en oeuvre. Ensuite, les problèmes posés mettent en avant un point fondamental pour la résolution d'un traitement automatique des faits de langue (aussi bien pour la construction du sens que pour le classement des savoirs linguistiques). Un programme de TALN doit pouvoir travailler sur de nombreuses connaissances, sur des méta-connaissances de ces dernières tout en reconstruisant dynamiquement de nouvelles connaissances. Ce dernier point n'est évidemment pas résolu ici mais est illustré dans notre cadre linguistique. Il n'est d'ailleurs pas encore accessible. On n'en sait pas encore assez sur les mécanismes qui permettent de reconstruire des connaissances. Il convient toutefois de poursuivre le travail de recherche sur l'utilisation des méta-connaissances pour le traitement automatique du langage naturel, quitte à limiter le champ d'étude sur des phénomènes restreints qui permettent de préciser les tâches encore à effectuer.