Origines et choix

"C'est un fait que ce que nous exprimons en paroles, couchons sur le papier, est dix fois plus bête que ce que nous pensons, et cependant nous acceptons, comme les grands écrivains, de passer pour beaucoup plus bêtes que nous ne sommes et commettons ce non-sens de dire quelque chose, de le coucher sur le papier, d'exprimer une opinion, de défendre une orientation, de prendre partie pour une idée,"
Thomas Bernhard, *Dans les hauteurs*.

Objectif

Affiner les connaissances à utiliser lors de l'analyse automatique : architectures d'analyseurs et paradigmes computationnels.

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. Il ne s'agit en aucun cas ici de construire une théorie. En particulier, puisque nous allons beaucoup parler de prototype, notre tâche n'est pas de prendre appui sur la théorie des prototypes des cognitivistes pour théoriser des faits de langue. Nous souhaitons au contraire confronter certains paradigmes informatiques aux problèmes que pose le traitement automatique de la construction du sens puis construire des outils en réponse aux problèmes rencontrés. On utilise donc le prototype (dans le cadre de la programmation à prototypes) comme un outil de représentation de certains faits linguistiques dans la mesure où nous pensons qu'il peut répondre à certains problèmes que posent ces faits de langue. Cet outil de représentation conduit à construire des structures de représentation simples et ajustables pour rendre compte justement des problèmes d'ajustements qui sont à l'oeuvre dans la construction du sens dans le langage naturel.

1. Pourquoi des classifications mouvantes ?

"Tous les objets du monde ont ceci en commun d'être et de n'être que la permanence provisoire de certains changements."
Jacques Roubaud, "La boucle", Editions du Seuil 1993.

Le traitement automatique du langage naturel doit passer par une phase de représentation des éléments linguistiques contrainte à figer les savoirs à modéliser, de présumer de ce que les savoirs représentés peuvent réaliser. Un examen des comportements des mots révèle des variations qu'il semble difficile de fixer dans des structures de représentations statiques. En langue, ce qui est considéré comme un objet n'est pas stable et la manière d'être d'un objet ne l'est pas non plus (Bachimont 1995).

Prenons trois exemples:

- L'exemple de `moulin`.

Dans le Petit Robert, on trouve associées à `moulin` les définitions suivantes:

Appareil servant à broyer, à moudre le grain des céréales.
Par extension, établissement qui utilise cet appareil.
Bâtiment où ce type de machine est installé.
Moulin servant à battre, à piler, à extraire le suc par pression.
Sens figuré : moulin à paroles.
Moulin à prières : cylindre renfermant des bandes de papier recouvertes de formules sacrées pour acquérir les mérites attachés à la répétition de cette formule.
Sens familier actuel, désigne un moteur de voiture ou d'avion, `"faire tourner son moulin"`, `"emballer son moulin"`.

Cette simple énumération ne rend pas compte des nuances de sens que l'on rencontre effectivement sur notre corpus (-> annexe corpus moulin, -> partie 1 chapitre 1). Les emplois métaphoriques étendent l'interprétation de `moulin` dans des séquences comme : `moulin à paroles`, `moulin à ennui` ou `moulin à thèmes anglais`. De plus les énumérations précédentes ne sont ni exhaustives ni stabilisées et ne sauraient l'être : un domaine notionel est toujours ouvert, soumis à des variations ou à des créations, et ne peut donc pas s'assimiler à un inventaire fini (Franckel & Lebaud 1992).

- L'exemple de `livre`.

Un livre peut être considéré comme un `Objet-Physique`. Un livre peut aussi être appréhendé via le contenu qui le compose. Il est aussi un élément constitutif du processus de publication ou de commercialisation attaché à la réalisation de ce livre. La "polémique" entre D. Kayser et M. Kleiber autour de `livre` met d'ailleurs en évidence les difficultés inhérentes à la description sémantique des mots (Kayser 1987,1989a, Kleiber & Riegel 1989,1991).

- L'exemple de `antialcoolique`.

Danielle Corbin (Corbin 1990) montre que les règles dérivationnelles du français permettent de construire quatorze mots `antialcoolique`. Et grâce au principe d'associativité[1], chacun des mots construits possède une structure morphologique et donc une interprétation sémantique (celle-ci étant différente des autres mots). Même si les mots construits possibles ne sont pas tous attestés, *"ils ne témoignent pas seulement d'un goût immodéré de l'auteur pour les jeux de*

[1] Principe d'associativité : à une pluralité de structures morphologiques correspond biunivoquement une pluralité d'interprétations sémantiques (Corbin 1990).

langue, ils ont surtout une valeur méthodologique et théorique. D'une part ils permettent de mesurer en grandeur réelle la différence entre le lexique conventionnel et le lexique dérivationnel, dans lequel tout locuteur peut puiser, pour peu qu'il en ait besoin et/ou envie, d'autre part ils sont l'outil dont dispose le linguiste pour trouver et faire apparaitre les régularités qui sous-tendent le lexique construit" (Corbin 1990). (Corbin & al. 1993) montrent d'ailleurs que les sens a priori figurés apparaissant dans les mots construits peuvent avoir des origines diverses. Ces sens résultent soit de règles sémantiques appliquées à différentes strates de la construction du mot, soit d'opérations sémantiques associées à un suffixe particulier ou encore le produit nécessaire d'une combinaison de contraintes culturelles et linguistiques (Corbin & al. 1993).

Les mots indiquent des accès

(Cadiot & Nemo 1997a) montrent qu'il est illusoire de vouloir caractériser les mots par leurs propriétés intrinsèques : il est donc illusoire de classer un objet par ses propriétés physiques immédiates . Ce n'est pas tant les propriétés physiques d'un objet qui permet de le définir mais plutôt les liens ou rapports que l'on établit avec ces objets. De même, il est illusoire de vouloir lister tous les sens possibles qu'il est possible d'attacher à un mot. Hors contexte, l'expression linguistique, ne portant pas en elle-même les prescriptions interprétatives permettant de lui associer un sens unique et non ambigü, perd les prescriptions interprétatives du contexte et devient donc équivoque. La pluralité des déterminations en contexte ne se généralise pas en une signification hors contexte univoque (Bachimont 1995).

Donner du jeu aux représentations des descriptions linguistiques

La langue évolue en permanence et les résultats acquis par la description linguistique sont toujours remis en question par la prise en compte de nouvelles informations. Il en va de même pour les comportements lexicaux : les savoirs associés aux mots ne sont pas donnés une fois pour toute. Ces derniers peuvent bouger et remettre en cause des représentations construites à un moment donné. En suivant ce constat, notre travail vise à décrire puis à représenter des conditions nécessaires pour des sens possibles et non des conditions nécessaires pour des sens effectifs. Puisqu'**il n'est pas satisfaisant de se contenter d'un modèle apriorique** pour construire une représentation des comportements des unités lexicales, notre travail de représentation s'inscrit dans **une approche expérimentale** qui ne présume pas complètement des choses à représenter. L'hypothèse suivie dans ce travail pour la représentation des informations linguistiques consiste à ne pas prédéterminer de manière figée ni les structures définies pour cette représentation ni leurs classements. Nous convenons avec (Haton & al. 1991) que *"la construction d'une hiérarchie est un processus incrémental"* et qu'une hiérarchie *"évolue et s'améliore en fonction des résultats obtenus jusqu'à ce qu'une certaine forme de stabilité soit atteinte"* . Notre démarche met en place **un processus de représentation évolutif** : les structures de représentation construites devront pouvoir être affinées dès que de nouveaux savoirs seront mis à jour. On peut en effet considérer que les comportements des mots ne sont pas tous prédéfinis mais que ceux-ci "émergent" dans le contexte dans lequel ces mots "agissent". Il n'est donc pas raisonnable de les considérer comme acquis par définition. Il s'agit au contraire de les mettre en lumière ainsi que les corrélations multiples qui existent entre les mots dans un flot continu de discours. Notre démarche consiste en quelque sorte à "faire émerger" les comportements des mots puis à les représenter ou à affiner les représentations existantes et enfin à classer les structures de représentation construites.

2. Modèles pour la construction du sens : critiques et discussion

> *"Le sens d'un mot en langue naturelle est peut-être le meilleur exemple de tous les éléments qui peuplent notre monde naturel : il faut bien connaître une langue pour appréhender les sens multiples d'un mot, et le même mot contribue par ailleurs à définir le sens de tous les autres mots. Aucun aspect de notre monde naturel et vivant ne peut être classifié à partir de délimitations nettes : on ne peut en faire un domaine dont on tracerait la carte."*
> (Varela 1986)

2.1. Savoirs linguistiques et heuristiques : comment représenter la mouvance ?

L'étude de certains modèles sémantiques (Fuchs et Victorri 1990, Pustejovsky 1991, Briscoe et Copestake 1994, Kayser 1987-1992, Rastier 1987-1992) nous oriente vers une représentation de la construction du sens qui se réalise sous forme de processus qui agissent en interaction. Notre travail s'inscrit fondamentalement dans la lignée de ces différentes approches, même si nous nous montrons critique sur certains aspects de ces modèles. Ces modèles visent tous à concevoir la construction du sens comme la mise en action de mécanismes. Ces approches s'écartent d'un appariement de termes et de sens. Elles ne consistent pas non plus à recourir uniquement à de l'induction ou à du raisonnement. Il s'agit plutôt d'établir un compromis entre des informations figées et des heuristiques qui peuvent enrichir ces informations. Interpréter c'est donc parcourir les chemins possibles à partir d'un savoir minimum que des pistes de sens permettent d'affiner. De même, dans les situations de communication, les savoirs linguistiques à l'oeuvre s'ajustent en tenant compte des incidences multiples que ces situations induisent. Ce qui est en jeu, c'est la confrontation entre la représentation de savoirs minimaux et les mécanismes qui permettent d'étendre ces savoirs en fonction des réalités contextuelles. Il s'agit pour certains d'organiser les savoirs de manière hiérarchique de telle sorte que des processus attachés à ce type de représentation permettent de reconstruire le sens. Pour les autres, il s'agit au contraire de décrire plus précisément les mécanismes qui réalisent la construction du sens. Dans le domaine particulier des séquences `moulin à N2`, la construction d'une telle hiérarchie de savoirs n'est pas triviale et pousse à réfléchir sur l'organisation des connaissances dans un tel cadre. Il est important de souligner que ce n'est pas tant le problème de construire une hiérarchie qui importe le plus[2], mais bien de donner les moyens de faire évoluer cette hiérarchie. Plus précisément, dans le domaine restreint étudié, l'interprétation du patron `moulin à N2` met en lumière les difficultés à organiser les savoirs sur des formes qui autorisent des extensions de sens : faut-il pour cela considérer que ces extensions de sens doivent être prises en compte dans une hiérarchie de concepts, ou bien dans une hiérarchie lexicale? De plus quel est le lien de dépendance entre ces types de hiérarchies? Les entrées lexicales se réduisent-elles à l'addition d'informations héritées d'une hiérarchie conceptuelle?

2.2. Ce que ces modèles peuvent (ne peuvent pas) faire ?

2.2.1. Kayser et la profondeur variable

L'approche suivie par D. Kayser vise à établir un équilibre entre des savoirs établis (non modifiables par inférence) et du raisonnement (les processus qui affinent par inférence les savoirs établis). La profondeur variable suppose en effet une hiérarchie de concepts aux noeuds de laquelle sont attachées des connaissances, des rôles qui permettent des inférences. Cela structure le savoir sur le monde effectivement mis en oeuvre. Cette hiérarchie est le support de l'inférence. Le savoir sur le monde est structuré et cette organisation guide le travail inférentiel. Ce type d'approche permet tous les remodelages possibles et met clairement en avant la difficulté qu'il y a à articuler les différents niveaux de savoirs manipulés. Le problème étant bien entendu de rendre compte dans la description des savoirs linguitiques de

[2] Ce n'est d'ailleurs pas le plus difficile, avec la programmation à objets tout est objet, i.e. on peut tout représenter dans une taxinomie de classes hiérarchisées.

leurs interactions multiples et de leurs potentielles évolutions respectives. La mise en place d'articulations dynamiques entre les savoirs décrits n'est pas un problème qui se laisse résoudre facilement surtout si les structures de représentation utilisées ne permettent pas de tenir compte de tels ajustements toujours possibles. De plus une approche conceptuelle à la Kayser est difficilement compatible avec le type d'outils de représentation que nous utilisons. Il semble en effet difficile de penser et de représenter une articulation cohérente des connaissances conceptuelles et des connaissances lexicales manipulées dans le modèle de représentation du langage naturel proposé ici. De bons outils réalisant cette articulation ne sont pas encore disponibles en raison justement de la difficulté à la mettre en place.

2.2.2. Le lexique génératif de Pustejowsky

De même, le lexique génératif de Pustejovsky suppose une hiérarchie aux noeuds de laquelle sont attachés des connaissances, des scénarios qui permettent des inférences. Là encore ce type de représentation[3] reste en deçà des possibles de réalisation d'une entité donnée. Puisque les faits de langue sont soumis à des évolutions permanentes, leur représentation se trouve en permanence remise en question et il est difficile de prévoir tous les scénarios possibles à attacher aux mots.

2.2.3. Les approches Objet pour le TALN

Si on choisit de représenter les savoirs linguistiques dans une taxonomie de classes, on se heurte de front au modèle de représentation sous forme d'objets dans lequel "un objet est (principalement) ce qu'il est". Les outils pour la représentation dans une telle approche ont tendance à figer les choses. Quand il s'agit de représenter une matière organique, les problèmes surgissent rapidement. Si on accepte de penser en termes d'objet, c'est-à-dire d'admettre que tout est réductible en un ensemble de propriétés définitoires stables, une telle approche se justifie. Malheureusement, on n'en saura jamais assez sur les faits de langue. Le langage crée en permanence. Un fait de langue peut prendre le contre-pied de résultats établis, bousculer les règles définitoires. On est donc contraint à des esquisses de représentation.

- Les classes de mots de Gross

G.Gross (Gross 1994, Le Pesant 1994) travaille dans le cadre théorique d'un lexique composé de phrases et non de mots isolés et propose un modèle pour le traitement informatisé du lexique à des fins notamment de traduction automatique. Il construit pour cela des classes de noms clairement spécifiées. Elles sont définies non pas "dans l'absolu", mais toujours en fonction d'un opérateur (plus précisément en fonction d'une construction syntaxique). Notre travail de classification vise lui aussi à classer les mots en prenant appui au départ sur des savoirs syntaxiques associés aux mots. Il ne s'agit pas de préconstruire une classification générale en dehors de toutes réalités textuelles mais de reconstruire automatiquement cette classification des mots en tenant compte justement de réalisations textuelles dans lesquelles le processus de classification doit tout d'abord retrouver les comportements lexicaux puis les classer : les classes de mots ne sont pas données de manière définitive mais construites sur corpus, dans la mesure où ces comportements sont soumis à des variations quand on passe d'un corpus[4] à un autre. Le travail sur les sous-langages révèle que les classements généraux[5] que l'on peut établir sur la langue ne sont pas toujours pertinents[6]. L'utilisation des mots dans des domaines

[3] Pustejowsky manipule un treillis de types figés.

[4] *"Two findings repeatedly come out of corpus-based studies: first, individual linguistic features are distributed differently across registers, and second, the same (or simular) linguistic features can have different functions in different registers"* (Biber 1993).

[5] *"Corpus-based research shows that our intuitions about lexicals patterns for grammatical structures, for many words there is no general pattern of use that holds across the whole language; rather, different word senses and collocational patterns are strongly preferred in different registers"* (Biber 1993).

[6] *"The collocational patterns (confirmed by concordance listings) identify two surprising fact about (these) words : 1) 'certain' is commonly used to mark uncertainty rather than certainty; and, 2) certainty is rarely expressed in social science at all. Thus, the most common collocations for*

de spécialités particuliers peut induire des comportements linguistiques nouveaux qui mettent justement en lumière un nouveau type de rapport avec l'objet visé dans ce domaine particulier (Cadiot & Nemo 1997). Si notre travail s'éloigne sur ce point de l'approche suivie par Gross, il en retient la reprise de l'approche harrissienne (-> partie 1 chapitre 1) (Daladier 1990, Dachelet 1994, Sager & al. 1987, Habert & Nazarenko 1996) : en fait la détermination de classes d'opérateurs et d'opérandes par le fonctionnement linguistique. Il ne s'agit donc pas d'une sémantique ontologique reposant sur des existants hors de toute réalisation linguistique.

- Hiérarchies rigides pour le TALN

Même si un choix de représentation qui manipule la notion d'objet dans une taxonomie de classes est adéquat pour fixer un domaine de savoir établi dans une structure hiérarchisée, il s'avère parfois délicat de construire puis de prendre en compte les évolutions de telles représentations hiérarchiques quand celles-ci décrivent des domaines de savoir non stabilisés. Dans une telle approche, la mise en contexte des instances de classes peut révéler la nécessité d'une mise à jour de la structure initialement adoptée pour la définition des classes. Dans le cas de la construction d'une hiérarchie lexicale pour laquelle on souhaite inscrire dans la définition des entrées lexicales une articulation forme-sens, une telle représentation doit affronter le problème complexe de la confrontation d'un savoir établi et d'un savoir à venir. La prise en compte de nouvelles informations, de résultats issus de l'analyse, de configurations énonciatives particulières... peut conduire à modifier, à affiner la représentation initiale. Dans le cadre de notre travail sur la modélisation des configurations nominales du type `moulin à N2`, une telle évolution sur le savoir initial est à l'oeuvre sur des formes du type `moulin à paroles` pour laquelle `moulin` devient `+humain`. Et c'est justement le travail de mise à jour d'une hiérarchie qui pose problème. Plus on structure un domaine de savoir, plus les liens de dépendance entre les noeuds du graphe d'héritage se complexifient, et il devient ainsi délicat de modifier cette structure hiérarchique. (Godard & Jayez 1994) propose le multitypage des entrées lexicales (`livre : objet matériel & objet informationnel`) pour prendre en compte les différentes réalisations sémantiques d'une entrée lexicale. On est confronté ici à une difficulté liée à la mise en place de hiérarchies rigides qu'il est souvent difficile de faire évoluer : le multitypage (de même que la coercion de type à la Pustejowsky ou les règles lexicales de Briscoe) s'accommodent d'un graphe des types figé. Vouloir classer les choses de manière figée pose donc des problèmes difficilement compatibles avec une nécessaire évolution dans les représentations construites des savoirs linguistiques à un moment donné. Il ne s'agit pas ici de dire que la représentation sous forme hiérarchique est inconcevable mais d'adapter les structures hiérarchiques aux processus qui sont à l'oeuvre dans la construction du sens des faits linguistiques. Ce travail vise aussi à établir une comparaison critique du pouvoir expressif de deux cadres de représentation : la Programmation à Objet dans une taxonomie de classes et la Programmation à Prototypes. 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'agira donc aussi 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.

2.2.4. Rastier et la sémantique différentielle

Dans l'approche sémantique suivie par Rastier, c'est le global (i.e. les textes) qui détermine le local. Un mot ne se définit pas par rapport à des états de choses ou à des états mentaux mais par rapport à ses contextes : un mot (occurrence) ne se définit que dans et par un contexte et reçoit des déterminations par le texte. Le sens d'un mot ne lui est donc pas permanent, il est

'certain' in social science reflect a kind of vagueness, marking a referent as possibly being known (by someone) but not specified in the text (e.g., a certain kind of..., in certain cases..., there certain indications that...). These collocations are relatively rare in fiction. In contrast, those collocational patterns for 'certain' that directly state that someone or something is certain - he/she/they/you/it + BE + certain and i/we + BE + certain - are extremely rare in social science but relatively common in fiction" (Biber 1993).

toujours le produit d'une interprétation fondée ou non sur des inférences et si le sens résulte d'une interprétation, il convient pour en rendre compte de décrire au mieux les parcours interprétatifs et les contraintes linguistiques sur ces parcours. De fait, placer les mots sous l'autorité d'un type revient à les décontextualiser et à les détextualiser et le rapport du type à l'occurrence s'éclaire si l'on considère que les types ne préexistent pas aux occurrences mais sont reconstruits à partir d'elles (Rastier 1995). Nous verrons (-> partie 2 chapitres 5, 6) que notre travail est contraint par une décontextualisation des faits de langue décrits, ce qui a priori nous éloigne de ces positions de fond. Il ne s'agit pas d'un choix théorique mais d'une condition de travail initiale minimale pour illustrer le traitement de certaines séquences nominales. En fait, dans l'état actuel de notre travail, les déterminations globales que les mots peuvent recevoir des textes dans lesquels ils s'inscrivent ne sont pas accessibles.

- Dans la première partie de ce travail (-> partie 2 chapitres 5, 6), on étudie des séquences nominales du type `moulin à N2` dont l'interprétation est principalement déterminée par le local (i.e. les composants de ces séquences) : ce sont les valeurs locales des composants qui déterminent les valeurs globables des composés étudiés. Cependant dans le cadre particulier étudié ici, il semble possible de faire converger une détermination du local sur le global et une détermination du global sur le local. L'absence (provisoire) de telles contraintes globales sur la production des séquences analysées force pour le moment l'analyse à donner plus de poids aux informations sémantiques locales (en particulier celles attachées au `N2` dans le composé) et donc à l'orienter dans une version compositionnelle : ce n'est pas un choix théorique de fait. Les savoirs lexicaux représentés étant ajustables dynamiquement, il reste possible d'affiner ces interprétations si des processus décrivant le rôle des textes sur ce type de séquence sont disponibles : soit en récupérant des contraintes sémantiques liées au développement de l'analyse, soit en tenant compte des contraintes générales que les séquences analysées révèlent. Ce dernier point n'est pas traité dans cette première phase de notre travail mais la question reste ouverte. (Fabre 1996)[7] montre d'ailleurs qu'il est possible de combiner des informations générales et des informations spécifiques pour une interprétation des noms composés. Nous souhaitons à terme pouvoir confronter les informations sémantiques disponibles, aussi bien contraindre les opérations que le global impose sur le local et réciproquement. Cette dernière exigence pose d'ailleurs le problème de savoir de quelle manière représenter les contraintes globales sur un texte dans un formalisme qui manipule des prototypes.
- Dans la deuxième partie de ce travail (-> partie 2 chapitres 7, 8), on travaille sur des savoirs extraits à partir de corpus. Là encore, l'impact des textes scrutés sur les éléments constitutifs de ces textes n'est pas complètement formalisé. En particulier, on ne tient compte pour le classement des mots représentés que des descriptions syntaxico-sémantiques qui leur sont associées. Si les savoirs syntaxiques correspondent à des réalisations avérées sur les corpus étudiés, certaines valeurs sémantiques utilisées sont quant à elles des prédéfinitions non issues du travail d'extraction de savoirs sur ces corpus. Là encore si on dispose de processus capables de décrire les liens entre les types ou les genres des textes sur les occurrences des mots rencontrés, il reste possible d'ajuster les mécanismes de représentation définis pour en tenir compte.

2.2.5. Continuité et TALN

La notion de continuité[8] pour décrire les processus sémantiques n'est pas une nécessité absolue : (Kayser 1995) montre qu'il est possible avec des modèles discrets de représenter une continuité conceptuelle. Mais l'utilisation de la continuité pour la modélisation des processus sémantiques (Fuchs & Victorri 1994) met surtout l'accent sur les difficultés rencontrées pour

7 *"The aim of our research is to define as precisely as possible the border line between what can be regularly described with general linguistic mechanisms, and what has to do with subregular or irregular phenomena which depend on corpus charactericts." /.../ "This work describes a domain-independent model for the interpretation of nominal compounds; it shows how general knowledge and domain-specific information may be combined for the interpretation of nominal compounds"* (Fabre 1996).

8 Cette notion vise à *"représenter la signification en termes non plus en termes de valeurs discrètes et disjointes mais en termes d'un continuum de valeurs"* (Fuchs & Le Goffic 1985). La formalisation mise en place pour représenter ce continuum de valeurs se situe dans le cadre de la géométrie différentielle.

mettre en oeuvre (dans un dispositif) les phénomènes linguistiques ainsi décrits. En effet le point important soulevé par ce type d'approche est clairement énoncé dans (Victorri 1994) : *"/./ representing continuity on a machine is all but a simple problem"*, et cette question ne se laisse pas résoudre facilement. La mise en oeuvre d'un dispositif informatique pour représenter ce type de phénomène rejoint d'ailleurs un point important de notre travail : la notion de **contrôle** et la **méta-représentation** (-> partie 3 chapitre 9). Si le connexionnisme par exemple est capable de rendre compte des phénomènes de la continuité, *"it has a drawback that prevents it from playing in continuous modeling the same role as classical IA tools play in discrete modeling. This flaw is related to an important notion developped in IA : the notion of control. A connectionist network remains 'a black box' which does not allow much reasoning about its functioning"*. Or, notre travail vise aussi à mettre en avant la nécessité de disposer de processus de contrôle pour le TALN. Si les descriptions/processus linguistiques à représenter (de manière discrète ou continue) sont potentiellement évolutifs, il est impératif de pouvoir évaluer les évolutions réalisées et donc de disposer d'outils de contrôle. De plus, si les processus sémantiques à l'oeuvre au cours de l'interprétation sont capables de créer de nouvelles connaissances, les dispositifs à mettre en place sont contraints non seulement de travailler sur de nombreuses connaissances mais aussi d'en reconstruire. On sent bien d'ailleurs que dans certaines phases d'une analyse automatique par exemple, si on pouvait disposer d'un méta-regard sur le déroulement de l'analyse et fournir des informations supplémentaires aux moments critiques de celle-ci, il serait possible d'infléchir le cours de l'analyse. Les dispositifs informatiques doivent donc impérativement prendre en compte ces phénomènes à travers les systèmes de représentation qu'ils développent. C'est en ce sens qu'il convient de les interroger pour que les modèles qu'ils représentent puissent être considérés comme des représentations empiriquement correctes.

2.3. Problèmes posés par ce type d'approche

2.3.1. Représenter, c'est figer un peu.

- Comment représenter la base de savoirs initiaux pour tenir compte de ce que l'on a affaire à des informations évolutives et non figées ? Comment enrichir ces savoirs établis quand de nouveaux types de descriptions linguistiques révèlent une certaine stabilité ?

Si les descriptions linguistiques se bornent à refléter un savoir partagé sur le monde, en particulier si le lexique est organisé (de manière hiérarchique ou non) sur la base de connaissances (le monde objectif), on perd une grande partie des mécanismes à l'oeuvre dans le langage naturel. L'interprétation permet en effet la mise en oeuvre de processus sémantiques qui créent de nouvelles acceptations et de nouveaux emplois non dérivables de la connaissance encyclopédique. Dans ce cas les parcours (arborescents ou non) n'apprennent rien que l'on ne sache déjà (Rastier 1995). Figer les savoirs linguistiques dans des taxonomies figées et non évolutives ne convient guère à la structure sémantique des langues et assez mal à celle des discours de spécialité (Biber 1993) : il n'est pas certain que les différents niveaux de l'arbre appartiennent à une même pratique (Rastier 1995). Un travail de représentation des faits de langue doit donc s'attacher à rendre les savoirs de base évolutifs en sachant qu'un ajustement est toujours potentiellement à venir sur ces savoirs.

2.3.2. "Ajuster les processus d'ajustements" : approches de la réflexivité

L'utilisation d'heuristiques pour étendre les savoirs initiaux représentés dans des configurations particulières pose quant à elle au moins deux types de problèmes.

- Comment représenter le fonctionnement des heuristiques pour rendre compte à la fois de leurs évolutions potentielles et de leurs applications sélectives (éviter de surgénérer) ?

Un dispositif pour le TALN est contraint d'enrichir ou d'affiner les processus inductifs mis en place (il en va de même pour les savoirs représentés). Dans la mesure où il est impossible de lister et de prévoir tous les scénarios possibles à attacher aux mots, il est nécessaire de pouvoir ajuster les processus définis au cas où telle heuristique doit justement être remodelée. On est donc confronté au problème fondamental qui consiste à travailler sur un nombre important de connaissances interconnectées, toujours en mouvement, remodelées par des processus évolutifs et capables de s'enrichir.

2.4. PàP : une nouvelle approche pour représenter la mouvance

2.4.1. Choix et perspectives

Notre travail vise à ne pas partir de représentations préconstruites et classées hiérarchiquement mais à reconstruire automatiquement des représentations de ces savoirs associés aux mots, en partant de savoirs peu déterminés qu'il s'agira ensuite d'affiner en tenant compte des informations disponibles. Il convient ici d'insister sur le point suivant. Il ne s'agit pas pour nous de nier en bloc toute forme de savoirs stabilisés dans ces processus de représentation automatiques des descriptions linguistiques. Les savoirs généraux en langue ne manquent pas. Les dictionnaires constituent d'ailleurs une trace de la fixation sur le moyen terme de ces savoirs. Il ne s'agit donc pas de tout réinventer. Notre travail vise au contraire à prendre appui sur des connaissances très générales (on utilise au départ par exemple un cadre de sous catégorisation approximatif) qui sont affinées voire remodelées et changées au gré des observations rencontrées de la même manière que les savoirs linguistiques évoluent (les dictionnaires sont mis à jour périodiquement...). Il s'agit ainsi pour nous de tendre vers un modèle cohérent qui ajuste les descriptions linguistiques dès qu'une certaine stabilité apparait. De même, dès qu'un certain type de comportement se révèle pertinent, on ajuste les savoirs définis pour en tenir compte. Notre objectif est double:

- 1. Extraire ces savoirs sur des corpus donnés et construire dynamiquement des structures de représentations évolutives.
- 2. Repérer les similitudes de comportements entre les unités lexicales rencontrées sur les corpus visés. On amorce ensuite une classification évolutive des unités de langue représentées.

2.4.2. Que fait la PàP ?

La représentation prototypique est un représentation progressive, en devenir.

La PàP (Liebermann 1986, Self Group 1987-1996) peut être considéré comme un outil privilégié pour la représentation de domaines de savoirs pour lesquels on ne dispose pas de catégories bien précises. La PàP conduit à penser différemment de la PàO pour construire une représentation informatique d'un certain domaine de connaissances. Il ne s'agit pas de partir d'une somme de connaissances figées et connues par avance mais de construire progressivement les entités informatiques suivant les connaissances dont on dispose sur le domaine visé. Si les savoirs à représenter ne sont pas connus de manière définitive, il est possible de commencer le processus de représentation en utilisant les savoirs déjà recensés puis d'affiner dynamiquement les objets construits dès que de nouveaux savoirs sont disponibles. A la limite, cette démarche de représentation n'a pas à se préoccuper de la forme terminale des objets à construire. Elle peut démarrer à partir d'une structure sous-déterminée qui sera mise à jour dès que des informations sont disponibles. **Cette flexibilité dans la PàP rend possible une mise en oeuvre du processus de représentation sur des domaines de connaissances évolutives**. Le choix des

prototypes semble cohérent avec la volonté de représenter des opérations dynamiques d'interprétation. Les prototypes sont malléables : ils se construisent contextuellement et leur spécialisation se définit suivant les évolutions contextuelles. Ils doivent commencer par fixer un savoir minimal - qu'il est possible d'attacher à une entrée lexicale - puis finir par étendre ce noyau de sens dans les directions permises par les configurations interprétatives rencontrées. La PàP met en place une représentation de savoirs que l'on peut considérer comme peu organisée au départ. Les savoirs représentés sont répartis sur une myriade de structures peu connectées entre elles à priori. **La PàP ne rejette pas pour autant toute idée de classification, bien au contraire**. Elle met d'ailleurs en place un système d'héritage simple basé sur la délégation qui permet de factoriser localement des comportements[9] partagés. Il reste donc possible d'affiner la représentation du domaine de savoirs étudié en reclassant les structures construites sur la base des "généralisations" potentielles que ces structures peuvent induire.

2.4.3. PàP et TALN : un certain point de vue

Notre utilisation de la PàP n'est pas celle qui est habituellement de mise avec ce type de langage. Notre approche met en avant la représentation des connaissances associées aux faits de langue avec la PàP alors que celle-ci a d'abord été conçue comme une boite à outils pour la réalisation d'interfaces graphiques. De plus notre travail ne s'appuie pas sur une conception cognitiviste du prototype. Il s'agit plutôt de le considérer comme un mode, un moyen de représentation adéquat pour décrire certains phénomènes du domaine visé. Ce qui est en jeu, c'est la confrontation dans la représentation entre une généralisation minimale et une évolution potentielle de ces représentations.

3. Objectifs

Notre travail se développe autour des thèmes suivants:

3.1. Travail sur `moulin`

Dans cette première partie de notre travail, on cherche à représenter la construction du sens des configurations nominales du type `moulin à N2`. Si le sens n'est pas donné, sa représentation ne peut pas être le résultat d'un simple codage. La représentation des phénomènes de la construction du sens met en avant la difficulté d'organiser les savoirs à l'oeuvre dans les énoncés. Pour réaliser ce premier objectif, on veut mettre en place une représentation matérielle de la construction du sens de ces configurations qui permette la mise en action de processus travaillant en interaction. Pour cela, on construit un dispositif informatique qui manipule des prototypes. Ce dernier doit identifier, classer et décrire ces formes pour construire une taxonomie ouverte et évolutive de cette classe particulière de composés.

3.2. Un dispositif pour la représentation et le classement automatiques des mots et de leurs comportements

Représentation des connaissances à partir de savoirs extraits sur corpus

[9] Utiliser un langage de programmation pour modéliser des objets du monde réel conduit à faire référence à deux types de comportements. D'une part les comportements de l'objet réel décrit, dans notre cadre de travail, cela concerne par exemple les comportements syntaxiques ou sémantiques des mots. D'autre part les comportements de l'objet informatique qui sert de support à l'objet décrit, par exemple des opérations qui permettent d'afficher ou d'imprimer l'objet informatique visé (Lemaire & Rechenmann 1995). La factorisation des comportements partagés possible avec la PàP peut concerner ces deux niveaux de description.

Notre démarche vise ensuite à mettre en lumière dans des domaines restreints les comportements des mots ainsi que les corrélations multiples qui existent entre les mots dans un flot continu de discours. Elle consiste en quelque sorte à "faire-émerger" ces comportements des mots puis à les représenter ou affiner les représentations existantes. L'hypothèse suivie ici est qu'il y a peu de sens à vouloir faire de l'acquisition sémantique en dehors d'un sous-langage. L'apprentissage sur corpus vise à repérer les arbres élémentaires de dépendance entre mots (relations opérateur-opérandes) et les contraintes sur les combinaisons de ces arbres. Nous envisageons de construire des représentations informatiques des mots et d'attacher à ces objets leurs comportements syntaxiques et sémantiques en ne prédéterminant pas complètement les savoirs à représenter. Notre travail s'inscrit dans cette reprise de l'approche harrissienne (Sager & al. 1987, Habert & Nazarenko 1996) : en fait la détermination de classes d'opérateurs et d'opérandes par le fonctionnement linguistique. Pour cette représentation nous utilisons le corpus MENELAS (Zweigenbaum 1994) à partir duquel sont extraits des savoirs pour construire des structures de représentations lexicales sous la forme de prototypes. Les savoirs extraits sont des arbres d'analyse fournis par un outil d'extraction terminologique (Bourigault 1993), ces arbres étant ensuite simplifiés dans le but de déterminer les arbres minimaux qu'il est possible d'associer aux entrées lexicales (Habert & Nazarenko 1996).

Du lexique jaillit un réseau de prototypes

Notre tâche vise à s'élever du particulier au général et à procéder à des regroupements pour y découvrir les faces cachées derrière les détails. Puisque la hiérarchie est un artifice, utile d'ailleurs pour classer les choses, nous ne nous imposons pas un classement prédéfini des savoirs à représenter, mais nous choisissons une démarche qui tente de reconstruire ce classement en tenant compte des informations délivrées par le travail d'extraction de savoirs à partir de corpus. Ce classement sur les entités représentées vise à la contruction de réseaux entre les prototypes définis à la manière d'un parcours entre ces prototypes pour y trouver le chemin interprétatif adéquat. Le réseau mis en place ne construit pas pour autant une représentation du sens attaché aux unités lexicales ou au structures syntaxiques représentées, il doit proposer des chemins interprétatifs qui ne constituent que des amorces d'interprétation devant être affinées par un travail d'interprétation plus fin. Notre approche vise donc à définir une démarche en spirale : construire des représentations par affinements successifs avec projections de savoirs soit constitués soit établis par ailleurs; puis ajustements et affinements des représentations à chaque nouvelle étape.

3.3. Un méta-niveau pour le Traitement Automatique du Langage Naturel

Puisque notre approche se situe dans un cadre de modélisation des connaissances par la construction de représentations évolutives, il est nécessaire de disposer d'éléments "dégagés" de ces représentations et qui permettent d'évaluer ce qui est en jeu pour ensuite permettre une prise en compte statique/dynamique des éléments pertinents susceptibles de faire évoluer les représentations initiales. Notre activité de production de faits de langue dispose de mécanismes de contrôle sur ce qui est produit : leur but principal est d'assurer une cohérence linguistique sur ce qui est produit (et notamment de produire une articulation forme-sens adéquate pour la compréhension). La production linguistique est faite de corrections, de reprises, de commentaires; les signes émis ne sont généralement pas de simples médiateurs physiques entre le "vouloir-dire" et le "dire construit" mais peuvent aussi matérialiser des ruptures, des précisions, des raffinements,... qui s'interposent dans le trajet de cette production (Authiez-Revuz 1989). On est ainsi confronté à des processus qui agissent à un méta-niveau. Ils accompagnent la formation du dire et les mécanismes à l'oeuvre sont "auto-contrôlés" dans le déroulement de cette construction. Si de tels mécanismes existent dans le langage, une représentation des connaissances de la langue ne doit pas les éviter. On sent bien d'ailleurs que dans certaines phases d'une analyse automatique par exemple, si on pouvait disposer d'un méta-regard sur le déroulement de l'analyse, et si on pouvait fournir des informations

supplémentaires aux moments critiques de celle-ci, il serait possible d'infléchir le cours de l'analyse. Mais il ne faut pas se faire d'illusions hâtives sur une utilisation globale de tels mécanismes. Notre difficulté à isoler les interactions complexes des savoirs à l'oeuvre dans la production d'énoncé accentue la représentation et la construction de tels mécanismes. Il semble possible d'y voir plus clair sur une utilisation de processus de contrôle à un méta-niveau si l'on restreint leur étude à des niveaux de complexité moindre. Notre réflexion portera aussi sur cet aspect fondamental dans la représentation des connaissances pour un dispositif informatique. D'autant plus que les langages de haut niveau que nous utilisons mettent en place des mécanismes qui favorisent les approches réflexives. Celles-ci peuvent éclairer une étude sur la construction de mécanismes de contrôle pour l'analyse et sur la définition même de ces mécanismes : peut-on construire de tels mécanismes capables de restructurer les représentations initiales, de modifier les opérations pré-définies ? C'est en tout cas ce que nous essaierons de présenter dans notre cadre linguistique choisi.

3.4. Cadre choisi : la PàP (Programmation à Prototypes)

Self[10] est le langage qui est utilisé ici pour représenter les mots et leurs comportements. Il a été conçu en 1986 par David Ungar & Randall Smith (Ungar & al. 1987). La première implémentation a été réalisée à Stanford en 1987, la dernière version (Self-4.0) est disponible depuis juillet 1995. Ce langage est désormais développé par Sun Microsystems avec beaucoup de moyens, ce qui semble indiquer l'importance de ce type de représentation dans le développement de la Programmation à Objets. Self est un langage qui permet l'héritage multiple et l'héritage dynamique via la délégation de comportements partagés par un ensemble d'objets. Un prototype ou un objet en Self est une entité composée d'attributs. Ces attributs peuvent porter des données, des méthodes ou pointer sur d'autres objets. Les objets définis dialoguent entre eux via un mécanisme d'envoi de messages.

4. Ce qui a été réalisé

4.1. Travail sur moulin

`Moulin` existe seul pour désigner un moteur, une machine ou un bâtiment où un certain type de moulin est installé ou l'a été. `Moulin` fonctionne aussi comme un opérateur syntaxique qui prend en arguments une préposition et un nom `moulin à blé, moulin à farine...` `Moulin` est donc un opérateur : notre prototype moulin initial peut déclencher la construction d'un prototype `moulin à N2` qui sera construit s'il est possible de lui adjoindre un prototype `Prep` de forme `"à"` et un prototype de catégorie `Nom` en position `N2`. On associe à `moulin` un prédicat dont la valeur prédicative dépend du `N2`. Ce choix permet tout d'abord de maintenir une représentation qui s'accorde avec bon nombre de noms d'instrument. Ensuite, la sous-spécification du prédicat peut être vue soit comme un problème de spécialisation lexicale, soit comme un phénomène de co-compositionalité. Un examen du fonctionnement des emplois métaphoriques comme `moulin à paroles` nous amène ensuite à proposer un modèle de représentation qui se construit par spécialisation. `Moulin à paroles` peut être associé au prédicat `créer` (création d'un excès de paroles) et désigne une entité de type `+humain` (une personne qui telle une machine produit une grande quantité de paroles). Cependant les emplois métaphoriques permettent aussi d'élargir la prise en compte des `N2` de type `Nparole` : `moulin à coups, moulin à systèmes, moulin à thèmes anglais, moulin à images, moulin à ennui...` On a dans ce cas un changement d'état du processus de création de l'argument donné : production massive, importante... Nous choisissons donc une reconfiguration de la représentation initiale et la définition d'un nouveau type de `moulin`. La construction

10 *"Self is more than a language. It is the combination of a language, a set of reusable objects, a programming environment, a user interface, and a virtual machine. /./ Self is based on prototypes and has no classes or meta-classes. Simplicity is ensured through uniformity: all data are objects, all expressions are message sends, and computation is performed solely by sending messages"* (Agesen 1996).

sémantique des composés en `moulin à N2` propose une interprétation privilégiée : un moulin est avant tout un nom de la classe instrument, la sémantique "de base" pour `moulin` porte la valeur prédicative `Faire-Changer-D'état`. C'est le "poids" de `moulin` (du `N1` dans le composé en `N1 à N2`) qui déclenche la construction sémantique du composé visé; celle-ci peut à son tour être modifiée en tenant compte des informations co-textuelles rencontrées : en l'occurrence, le "poids" sémantique du `N2` peut sous-spécifier cette sémantique initiale soit la respécifier complètement. Cette solution permet de préserver une évolution potentielle pour l'interprétation des configurations `en moulin à N2`. Si l'on rencontre une séquence nouvelle, il reste possible à partir des interprétations existantes, de construire une nouvelle interprétation. Ces ajustements peuvent être réalisés "à la main" : mise à jour dynamique des attributs des prototypes visés. Si l'on dispose de méta-connaissances qui permettent de mettre en lumière les informations pertinentes qui éclairent une nouvelle interprétation, ces ajustements peuvent être réalisés via l'utilisation de méthodes idoines : mise en place de comportements généraux ou de comportements spécifiques au niveau des prototypes rencontrés.

4.2. Un dispositif pour la représentation et le classement automatiques des mots et de leurs comportements

Nous montrons tout d'abord comment la PàP permet la génération automatique des prototypes lexicaux et des arbres syntaxiques associés sur la base de savoirs extraits d'un corpus arboré. Les savoirs extraits sont des arbres d'analyse fournis par un outil d'extraction terminologique (Lexter[11], Bourigault 1993), ces arbres étant ensuite simplifiés dans le but de déterminer les arbres élémentaires de dépendance qu'il est possible d'associer aux entrées lexicales (Cyclade[12], Habert & Nazarenko 1996). La PàP permet ensuite d'amorcer un début de classement des unités lexicales représentées sur la base de leurs contraintes syntaxiques en construisant des hiérarchies locales de comportements[13] partagés. On classe les prototypes lexicaux définis sur la base des (prototypes d') arbres qui leurs sont associés. Les savoirs attachés aux prototypes lexicaux se composent en effet d'une liste d'arbres élémentaires (avec des contraintes possibles sur les noeuds de l'arbre) et pour chacun de ces arbres élémentaires d'une éventuelle liste d'arbres d'analyses (avec là aussi des contraintes possibles). On recherche donc les arbres élémentaires communs à un ensemble de prototypes lexicaux. Si on trouve des arbres élémentaires communs à un sous-ensemble de prototypes, on construit un pôle de comportements partagés qui va porter les arbres élémentaires communs. On établit ensuite un lien de délégation entre ce pôle et les prototypes concernés. On met donc en place, automatiquement, un réseau de pôles de comportements partagés en définissant des hiérarchies locales sur des sous-ensembles de prototypes. L'évaluation de connaissances définies préalablement comme prototypiques permet d'envisager une classification hiérarchisée des savoirs représentés, même si celle-ci reste évolutive. Notre objectif est de tendre vers la détermination de classes sémantiques, de manière inductive, dans une approche symbolique[14], celle des sous-langages. Il convient de souligner que le classement opéré s'appuie principalement sur des contraintes syntaxiques. Ce premier classement ne dit rien de plus sur les agrégats de comportements partagés construits. La syntaxe est utilisée pour dégrossir le classement. A l'inverse des approches harrissiennes[15] et statistiques[16], notre approche ne

[11] Lexter est un outil d'extraction terminologique conçu et réalisé dans un environnement industriel: la Direction des Etudes et Recherches de EDF, pour aider à la mise au point de terminologie. Il prend en entrée des textes arbitrairement longs et produit des arbres d'analyses de séquences nominales en décomposant ces séquences en Tête et Expansion, de manière récursive

[12] Le programme Cyclade (Habert & al. 1995) a pour tâche de simplifier les arbres d'analyse fournis par une analyse, ici par Lexter, et de mettre en évidence les arbres qui présentent des relations de dépendance entre mots pleins (vs. mots vides : préposition, déterminant, conjonction...)

[13] On classe donc les objets représentant des mots sur la base des comportements réels associés aux mots ainsi modélisés. Les comportements informatiques qui ne décrivent les mots dans leur activité langagière ne sont pas pertinents pour ce classement.

[14] (Resnik 1993) présente une approche de construction de classes sémantiques de manière inductive dans une optique statistique.

[15] Dans l'approche suivie par N.Sager, les classes de mots et les formules syntaxiques d'une grammaire de sous-langage sont en correspondance étroite avec les classes d'objets du monde réel et à leurs relations, ce qui conduit à sous-estimer le travail d'interprétation des résultats construits.

[16] Dans les approches statistiques (Church & Hanks 1990, Hindle 1990), les classes de mots sémantiquement pertinentes sont obtenues statistiquement et induites du corpus à partir des seules mesures de similarités entre mots : *"En réalité, les listes de mots obtenues ne constituent pas de véritables classes de mots cohérentes et homogènes. Il est souvent difficile de leur attribuer une étiquette globale, et ce d'autant plus que les mesures de similarité constituent des résumés statistiques bruts dont les critères de construction sont effacés"* (Habert & Nazarenko 1996).

conduit pas à la détermination de classes sémantiques satisfaisantes mais elle constitue une méthode d'amorçage pour l'élaboration de l'ontologie du domaine, nous suivons sur ce point la démarche suivie par (Habert & Nazarenko 1996) : la construction de l'ontologie du domaine étudié nécessite un part d'interprétation, *"il y a, entre le flou notionnel inhérent aux langues naturelles (Kayser 1992), y compris aux langues de spécialités (Dachelet 1994), et la stabilité conceptuelle qui est requise dans les ontologies construites, un seuil qui ne peut être franchi automatiquement"* (Habert & Nazarenko 1996).

4.3. Un méta-niveau pour l'analyse

Le travail de représentation mis en place ne construit que des ébauches. Il s'agit ensuite d'ajuster les représentations initialement définies en tenant compte des informations que l'analyse dévoile. Pour la construction du sens, il convient de décrire au mieux les mécanismes qui opèrent pour réaliser cette mise en actions de mécanismes complexes et de s'approcher au plus près de ces mécanismes. On ne les touchera de près que lorsque nous serons capables de travailler sur de nombreuses connaissances et de produire de nouvelles connaissances à partir de savoirs initiaux. Le traitement automatique de la construction du sens est en effet confronté au problème de travailler sur de nombreuses connaissances et sur la nécessité d'en reconstruire de nouvelles. Cet impératif n'est pas encore accessible[17]. 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. Ce travail sur des processus réflexifs pour un traitement automatique passe impérativement par une phase de développement de processus de traçage des activités d'un programme de traitement automatique. Dans notre travail, les processus réflexifs disponibles avec le cadre de représentation choisi ont permis de développer des outils de suivi ou de mise au point de l'analyse. Ces outils présentent, de manière raisonnée, les différents états pertinents des traitements réalisés et des résultats construits à chaque étape.

[17] *"Atteindre le même niveau de performance que les êtres humains dans leur traitement des langues naturelles est un problème extraordinairement difficile que beaucoup croient utopique. Je doute que nous puissions y ariver sans résoudre au préalable les deux problèmes clés de l'IA : utiliser efficacement une grande quantité de connaissances déclaratives et découvrir automatiquement de nouvelles connaissances. Et je ne pense pas que les langues naturelles soient un bon support pour étudier ces deux problèmes, parce que ce domaine est beaucoup trop complexe; il vaut mieux faire progresser ces deux domaines de recherche en étudiant des domaines qui n'exigent pas une si grande quantité de connaissances. Travailler sur l'utilsation de métaconnaissances est un détour qui aura des conséquences pour les systèmes manipulant les langues naturelles, mais nous ne devons pas en attendre des résultats spectaculaires à court terme. Vu l'importance du traitement des langues naturelles, ce domaine doit être attaqué de tous les angles possibles. Aussi est-il indispensable de poursuivre simulatément des travaux dès maintenant utilisables dans le domaine des langues naturelles, mais il est douteux que cela permettra d'atteindre une qualité de traitement comparable à la nôtre même à moyen terme"* (Pitrat 1992).