Analyse syntaxique par Systèmes Multi-Agents adaptatifs : apport à la construction d'ontologies

MÉMOIRE

Année Universitaire 2003-2004

pour l'obtention du

DEA Représentation de la Connaissance et Formalisation du Raisonnement – Toulouse III

par

Kévin Ottens

Encadrants :	Pierre Glize	Ingénieur de Recherche CNRS - HDR
	Nathalie Aussenac-Gilles	Chargé de Recherches CNRS
	Marie-Pierre Gleizes	Maître de Conférences (Toulouse III)

Table des matières

RÉSUMÉ

Mots-clés : Analyse syntaxique, Architectures multi-agents, Ontologies

Les questions de la construction et de la maintenance d'ontologie à partir de textes doivent être posées conjointement afin de fournir une solution qui favorise les mises à jour simples et fréquentes. Pour cela, nous voulons évaluer l'apport de Systèmes Multi-Agents (SMA) adaptatifs à l'analyse de textes pour le repérage d'éléments d'ontologie. Ce mémoire présente une caractérisation des résultats actuels sur les avantages du Traitement Automatique des Langues (TAL) pour la construction d'ontologies à partir de textes. Il dresse ensuite un état des travaux faisant appel à des SMA pour mener des analyses syntaxiques. Puis, nous développerons les caractéristiques qui paraissent nécessaires à un SMA basé sur la théorie des AMAS (cadre de travail de mon équipe d'accueil) pour faciliter la construction et la maintenance d'ontologies à partir de textes. Enfin, nous tenterons de résoudre un sous-problème d'analyse textuelle pour évaluer la pertinence de l'utilisation d'un SMA basé sur la théorie des AMAS n'exploitant aucune ressource linguistique.

ABSTRACT

Keywords : Syntactic analysis, Multi-agents architectures, Ontologies

The problems of engineering and maintaining ontologies from texts are to be jointly delt with in the perspective of making frequent and easy up-dates. To reach this goal, we want to evaluate the gain brought by adaptive multi-agent systems (MAS) for text analysis and for the identification of ontological components. This thesis presents the current results and advantages of natural language processing systems used for ontology engineering from texts. Then, we list the major works where multi-agent systems carry out syntactical analyses. We detail the characteristics we judge necessary to a MAS based on the AMAS theory (work context of my research team) to easier the build and maintenance of ontologies from texts. We'll finally try to solve a problem of textual analysis to evaluate if the use of a MAS based on the AMAS theory working with no linguistical resource is relevent.

Remerciements

A l'issue de cette année de DEA qui débouche sur le présent mémoire, je tiens à exprimer ma gratitude pour...

... Marie-Pierre Gleizes et Pierre Glize pour leurs qualités scientifiques et leur investissement personnel, même lorsque nos avis divergent.

... Nathalie Aussenac-Gilles qui m'a permis de voir de plus près le monde de la recherche et d'y prendre une part active cette année, et pour sa patience lors des relectures des brouillons de ce mémoire.

... Jean-Luc Soubie pour sa gentillesse, son humour et les discussions fort instructives que j'ai pu avoir avec lui.

... Axel Reymonet que je n'oublierai pas.

... Mario, Ludovic, Lionel et Adrien pour leur confiance en moi.

Et enfin, je conserve une pensée particulière pour mes parents sans lesquels je n'aurais jamais pu effectuer ce travail, et Virginie qui a été mon meilleur soutien dans les moments difficiles.

Merci !

Chapitre 1

Introduction

Pour ce mémoire, nous avons cherché à évaluer l'apport de systèmes multi-agents adaptatifs menant des analyses syntaxiques pour la construction d'ontologies. Ainsi, les travaux présentés dans ce mémoire se situent à la croisée de trois domaines :

- l'analyse syntaxique qui détermine la structure de phrases en langage naturel ;
- les ontologies qui sont des objets informatiques permettant de représenter des connaissances sous la forme de réseaux conceptuels ;
- les systèmes multi-agents qui sont des ensembles d'entités logicielles autonomes en interaction.

1.1 Ontologies, Analyse syntaxique et SMA

Depuis moins d'une dizaine d'années, les recherches sur la construction d'ontologies à partir de textes sont considérées comme une issue possible aux problèmes du coût et de la difficulté de la construction de ces ontologies. Ces facteurs étant les principaux freins à leur utilisation, en particulier dans le cadre du Web sémantique, il devient aujourd'hui stratégique de s'y intéresser pour les prochaines années. Notre recherche se situe dans ce cadre.

Nous souhaitons poser conjointement les questions de la construction et de la maintenance d'ontologie à partir de textes afin de fournir une solution qui favorise les mises à jour simples et fréquentes. Il s'agit déjà d'une hypothèse forte, la plupart des méthodes actuelles se focalisant sur leur construction au détriment de la maintenance.

A l'heure actuelle, la construction d'ontologies à partir de textes utilise intensivement les outils de Traitement Automatique de la Langue (TAL). Ils sont utilisés pour différents types de traitements. Les plus intéressants pour la construction d'ontologies sont les extracteurs de termes et les extracteurs de relations syntaxiques entre ces termes. Toutefois, ces derniers outils reposent sur tout un ensemble de traitements préalables du texte. Chacun de ces traitements correspond à un sous-problème d'analyse (par exemple l'étiquetage morpho-syntaxique) dont les résultats alimentent les étapes suivantes. Les outils développés pour répondre à ces sous-problèmes ayant été conçus séparément (pour des raisons évidentes de simplification), nous avons aujourd'hui un cloisonnement entre les différents niveaux d'analyse. Ils doivent alors être utilisés séquentiellement et fournissent une quantité énorme d'informations à l'analyste.

Pour cela, nous voulons évaluer l'apport de Systèmes Multi-Agents (SMA) adaptatifs à l'analyse de textes pour le repérage d'éléments d'ontologie (candidats termes, classes sémantiques, relations lexicales en vue de définir des concepts et des relations sémantiques). En effet, il s'agit d'un problème difficile dont la résolution pourra tirer partie d'un système ouvert et dynamique, caractéristiques intéressantes des SMA. Nous conservons l'idée qu'une analyse syntaxique est un préalable au repérage d'éléments ontologiques. De ce fait, nous devons définir en quoi la manière de mener une analyse lexicale et syntaxique peut être revue si l'on utilise des agents. En particulier, déterminer jusqu'où l'emploi de connaissances sur la langue peut être réduit pour effectuer ces traitements.

Ce questionnement paraît légitime dans le cadre de l'ingénierie des connaissances où l'extraction de concepts et de relations ontologiques ne nécessite pas une analyse syntaxique exhaustive des textes. En effet, une part réduite des corpus textuels est utilisée, et des entorses à la grammaire sont souvent commises. En ce sens, nous nous inspirons de l'approche en vigueur au GREYC [Vergne, 1999] concernant l'abandon d'une ambition de description exhaustive des langues.

1.2 Un projet ambitieux ?

Comme le prouve l'état de l'art en matière de construction d'ontologies à partir de textes, il s'agit d'un projet ambitieux. C'est pourquoi nous avons fait preuve d'humilité en ne traitant que ce qui paraît raisonnablement envisageable durant une année de DEA. Ce sujet constitue une étude de faisabilité car le problème soulevé n'a été étudié ni dans la littérature, ni par les deux équipes qui ont proposé le sujet et qui collaborent dessus. C'est pourquoi mon stage a consisté à mener un état de l'art des domaines concernés afin de situer le problème ainsi qu'une expérimentation à l'aide de SMA sur une analyse syntaxique réduite qui pourra constituer la base d'une construction d'ontologies. Il faut donc voir les pages qui vont suivre comme une amorce pour un travail de thèse, un premier pas vers une nouvelle façon d'appréhender le travail, actuellement long et fastidieux, de la construction d'ontologies.

Ce stage pose donc les bases de ce projet et en présente les hypothèses. Avant de définir un nouveau système, nous avons essayé de préciser les besoins en matière d'analyse de texte en nous basant sur les approches actuelles utilisant le Traitement Automatique des Langues (TAL) pour la construction d'ontologies. Ensuite, dans cet objectif, nous avons aussi étudié différents systèmes reposant sur une architecture multi-agent et dont une des fonctionnalités est de mener une analyse syntaxique de textes.

Nous envisageons une solution reposant sur des agents réactifs situés au niveau des structures du texte (mot, syntagme...) et basés sur la théorie des Adaptive Multi-Agent Systems (AMAS) [Georgé *et al.*, 2003]. Mettre au point ce genre de système suppose de retenir des hypothèses et de faire une série de choix quant à la nature, aux rôles et aux connaissances des agents, ainsi qu'à leurs échanges et leur manière de gérer des situations dites non coopératives. La fonction globale du système d'agents doit être de fournir des éléments de décision à un analyste chargé d'organiser l'ontologie.

Le système devra effectuer différents types de traitements à plusieurs niveaux : syntaxique puis sémantique et ensuite conceptuel. Pour ce travail de DEA, nous avons pris le parti de nous intéresser, au premier niveau d'analyse, à un étiquetage syntaxique réduit des termes simples trouvés au sein de textes techniques. Les résultats obtenus ont été validés par ce qui peut être dégagé automatiquement à l'aide d'un étiqueteur (bien trop précis pour nos propres besoins). Cette décision a été prise pour une raison simple, il s'agit du traitement préliminaire à tout travail à partir de texte dans les approches classiques. Les résultats obtenus permettront alors de juger des mesures à prendre pour permettre aux agents de contribuer à une organisation sémantique et conceptuelle à partir du niveau syntaxique.

En conclusion, seulement une partie des compétences et des interactions des agents utiles à une analyse syntaxique minimale seront définies. Les recherches futures conduiront à enrichir les agents de compétences complémentaires qui permettront de présenter des regroupements conceptuels pertinents pour construire l'ontologie.

1.3 Organisation du mémoire

Afin d'explorer ces différents aspects, ce mémoire s'articule en cinq chapitres :

- une présentation des problèmes actuels de la construction d'ontologies ;
- une caractérisation des résultats actuels sur les apports du TAL à la construction d'ontologies à partir de textes suivie d'une présentation des différentes techniques de TAL que nous jugeons utiles comme inspiration pour notre futur système ;
- un état des travaux faisant appel à des SMA pour mener des analyses syntaxiques ;
- une présentation de la théorie des AMAS suivie des hypothèses retenues pour un système complet de construction d'ontologies à partir de textes basé sur cette théorie ;
- une évaluation des résultats obtenus pour une partie du problème de l'analyse syntaxique préalable dans le cadre de la construction d'ontologies à partir de textes en se focalisant sur des méthodes de calculs entièrement basées sur des statistiques.

Les études présentées dans les chapitres 2 et 4 de ce mémoire ont donné lieu à une communication [Ottens *et al.*, 2004] lors d'une journée de l'ATALA[1] organisée le 13 mars 2004 sur le thème « Agents et langue ».

[1] `http ://www.atala.org`

Chapitre 2

Problématiques actuelles des ontologies

2.1 Cadre général des Ontologies

2.1.1 Définition

La construction d'ontologies est un domaine qui intéresse à la fois l'ingénierie des connaissances, la linguistique et la philosophie. C'est d'ailleurs cette dernière discipline qui leur a donné leur nom. Initialement, l'Ontologie était un domaine de la philosopie concernant « l'étude de l'être en tant qu'être, c'est-à-dire l'étude des propriétés générales de ce qui existe »[2].

En ingénierie des connaissances, une ontologie devient un objet qu'il convient de construire. Une ontologie est alors une description rigoureuse et structurée du vocabulaire d'un domaine spécialisé (on parlera parfois d'un ensemble structuré de concepts). Cette structuration est à rapprocher des réseaux sémantiques qui permettront de représenter les concepts de l'ontologie (les noeuds du réseau) et les relations qu'ils ont entre eux (les arcs du réseau), la figure 2.1 donne un extrait d'un tel réseau. Lorsque l'on considère l'opérationalisation de l'ontologie, il est nécessaire d'utiliser des langages de description informatique comme le récent standard OWL[3]. Ces langages permettent généralement de manipuler et spécifier des classes, des propriétés de ces classes et des restrictions sur ces propriétés. Il est parfois même possible de définir des axiomes.

Une fois l'ontologie d'un domaine constituée, on peut définir des instances des classes qui devront vérifier les propriétés, les contraintes de leur classe ontologique et satisfaire les axiomes. Lorsque l'on considère le couple ontologie et ensemble d'instances, on peut parler d'une base de connaissances de ce domaine.

Historiquement, la notion d'ontologie est apparue pour satisfaire un besoin d'interopérabilité dans les systèmes informatiques. Il est nécessaire d'avoir à la fois un même langage pour interagir mais aussi des connaissances partagées. Le but des ontologies est donc d'avoir une même description des objets utilisés par plusieurs agents logiciels.

Au sein d'un domaine, des jargons spécifiques (on parle alors de langue de spécialité) finissent toujours par émerger lorsque ce domaine se structure. Cet état de fait entraîne des problèmes de compréhension et freine le partage des connaissances entre personnes de spécialités différentes. Le rôle fondamental des ontologies est d'améliorer la communication, non seulement entre machines seules, mais aussi entre humains et machines ou encore entre humains (principalement par le biais de logiciels). Cette vocation permet une utilisation des ontologies dans un cadre assez large.

[2] cf. l'article de Wikipédia : `http ://fr.wikipedia.org/wiki/Ontologie_(philosophie)`

[3] Recommandation du W3C, cf. `http ://www.w3.org/2001/sw/WebOnt/`

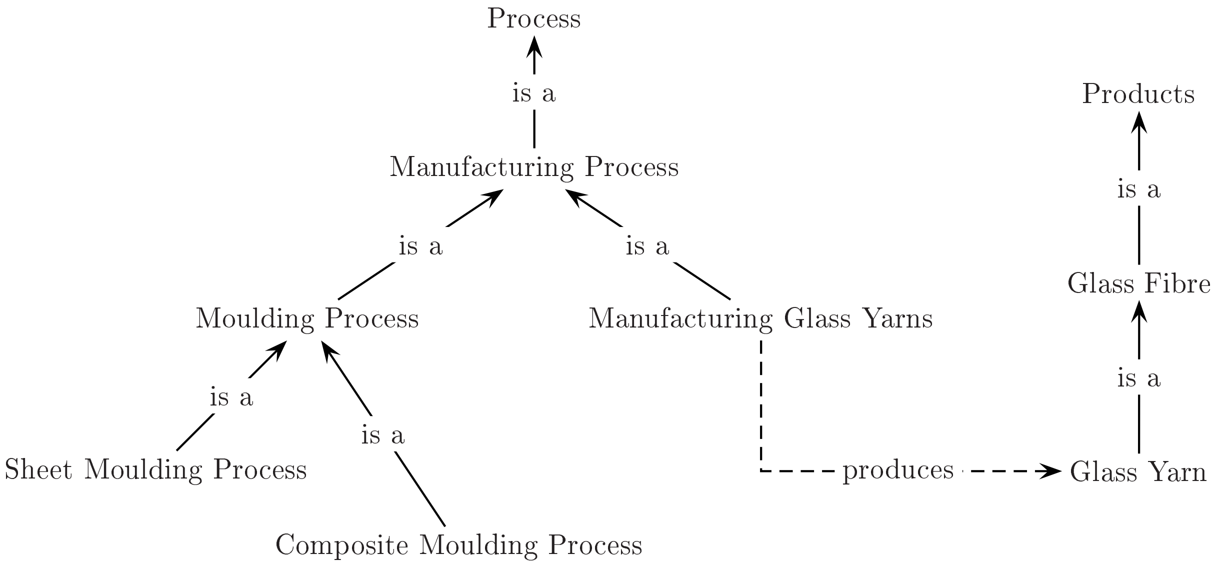

Fig. 2.1 – Extrait d'un réseau sémantique

Au sein d'une entreprise, elles peuvent permettre d'améliorer la compréhension entre professionnels travaillant sur des domaines différents, mais aussi favoriser l'exploitation des informations de l'entreprise pour mieux les gérer et les diffuser. Dans les deux cas, il s'agit d'activités médiatisées par des logiciels. Pour ce dernier point, on pourra mentionner leur application en recherche d'informations, une ontologie permettant entre autre d'indexer des documents ou encore de mieux interpréter des requêtes utilisateur en langage naturel.

Dans le cadre de systèmes informatiques complexes, les ontologies montrent tout leur intérêt puisqu'elles répondent là à leur rôle premier. Pour une conception de système, par exemple, une ontologie peut aider à analyser les besoins et définir les spécifications. De plus, une ontologie a vocation d'être réutilisée dans plusieurs applications différentes. En effet, pour constituer une base de connaissances, rien n'empêche de reprendre une ontologie existante en utilisant de nouvelles instances spécifiques à la nouvelle application. Elles permettent aussi de jouer le rôle de format d'échange pour des systèmes informatiques devant coopérer ou échanger des informations (les webservices, les agents logiciels sur le web, les systèmes multi-agents, et le « pervasive computing » [McGrath *et al.*, 2003] viennent à l'esprit dans ce cas).

Paradoxalement, le terme « ontologie » est associé à une grande variété de types d'objets. La plupart des ontologies sont au moins structurées à l'aide de la relation « est-un » qui est la relation de généralisation entre concepts, ainsi on peut définir des hiérarchies de concepts. L'autre relation très utilisée est la relation tout-parties (« est composé de »). Ainsi, on peut considérer les taxonomies comme des variétés d'ontologies n'utilisant que la relation de généralisation. D'ailleurs, tout comme les taxonomies, la hiérarchie d'une ontologie est fortement influencée par son but [Charlet, 2003].

Elles peuvent aussi varier par rapport à d'autre métriques comme le nombre d'axiomes contraignant le sens des concepts (en opposition à une simple hiérarchie de concepts sans autre définition), ou la taille allant de quelques dizaines de concepts à des milliers. Enfin, elles peuvent être générales ou ne couvrir qu'un domaine particulier.

2.1.2 Une construction difficile

Malgré cette grande souplesse d'utilisation, et le caractère réutilisable des ontologies, leur construction est une tâche longue et difficile. Elle a toutefois évolué d'un artisanat omniprésent sans aucune méthode vers la naissance de méthodes fondées sur la linguistique (par exemple TERMINAE [Szulman *et al.*, 2002]) et sur l'Ontologie [Gangemi *et al.*, 2002]. Toutefois, il n'existe pas de processus unifié. Certaines méthodes se concentrent sur l'exploitation de textes à l'aide du TAL, y compris par apprentissage automatique, d'autres se focalisent sur la réutilisation ou sur les principes de structuration et de définition formelle des éléments de l'ontologies.

Lorsqu'elles sont construites à partir de textes, les ontologies comportent beaucoup plus d'informations terminologiques. Par exemple, les différents termes représentant un concept et des informations complémentaires (informations grammaticales, variantes de forme autorisées ou recommandées et même règles de désambiguïsation) lui sont associées. Cet enrichissement provient naturellement de l'examen des textes ayant servi à la construction de l'ontologies. En contre-partie, elles ont une tendance à être moins bien formalisées. Il est très difficile d'avoir un seul point de vue de découpage hiérarchique même avec une ressource textuelle, ce qui peut jouer sur la qualité du produit résultant.

Malgré tout, quelles que soient les approches considérées, certains points restent valables. En particulier, pour un domaine donné, il y a toujours plusieurs ontologies possibles. Dans ce cas, ce sont l'application cible et les objectifs fixés a priori qui doivent servir de guide lors de la modélisation. En général, la construction d'ontologies est un processus itératif. Dans le cadre d'un travail à partir d'un corpus textuel, cela se manifeste par des retours fréquents au texte pour affiner la modélisation. Enfin, on se doit de la valider auprès d'experts du domaine et par son utilisation dans l'application cible.

Dans le cadre d'une méthodologie de construction, les grandes étapes transversales que l'on peut distinguer sont les suivantes :

1. déterminer le domaine et l'objectif de l'ontologie ;
2. considérer les ontologies existantes, dans une optique de réutilisation ;
3. déterminer le vocabulaire de l'ontologie ;
4. définir les classes et leur hiérarchie ;
5. définir les propriétés des classes ;
6. définir les contraintes des propriétés ;
7. enfin, créer les instances pour obtenir une base de connaissances, cette étape pouvant avoir lieu lorsque l'ontologie est en service dans l'application cible.

Toutefois, même si le cadre méthodologique de la construction d'ontologies va se précisant, cette activité laisse une part importante au modélisateur qui orientera toujours plus ou moins le produit résultant. C'est pourquoi, la qualité de l'ontologie ne pourra être pleinement évaluée qu'en phase d'utilisation dans une application donnée. Par la suite, nous nous attacherons plus particulièrement à la construction d'ontologies à partir de textes à l'aide d'outils de TAL. Elle peut être découpée en quatre principales étapes [Aussenac-Gilles et Busnel, 2002] :

1. Constitution d'un corpus à partir d'une analyse des besoins de l'application visée ;
2. Analyse linguistique pour identifier les termes et les relations lexicales en utilisant des outils de traitement de la langue naturelle ;

3. Normalisation sémantique conduisant à des concepts et des relations sémantiques définis dans un langage semi-formel ;
4. Formalisation en logique de description et intégration des concepts au sein d'une ontologie formelle.

Ces étapes ne sont pas complètement indépendantes et l'on passe progressivement de l'une à l'autre.

2.2 TAL pour la construction d'ontologies à partir de textes

Aujourd'hui, la contribution du TAL à la construction d'ontologies prend la forme de plusieurs types de logiciels utilisés pour repérer dans des textes des formes linguistiques révélatrices de connaissances. Notre but n'est pas ici d'en dresser une liste exhaustive, mais plutôt d'en dégager les tendances fortes et les principales caractéristiques. La figure 2.2 présente la place des outils de TAL dans un processus de construction d'ontologies à partir de textes [Aussenac-Gilles et Busnel, 2002].

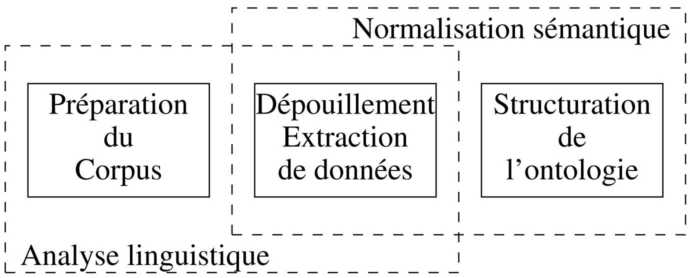

Fig. 2.2 – Positionnement du TAL dans la construction d'ontologies

Les outils de TAL sont utilisés dans toute la phase d'analyse linguistique. Il est alors important de constater que cette phase se recoupe avec la phase de normalisation sémantique lors de laquelle sont dégagés et organisés les concepts du domaine. Le TAL a donc une place importante puisqu'il intervient indirectement sur la structuration de l'ontologie, cette structuration reposant sur des résultats issus d'outils d'analyse linguistique.

2.2.1 Un même point de vue sémantique

Parmi les systèmes les plus utilisés, on trouve des concordanciers, des classifieurs, des extracteurs de termes, de relations ou, plus sophistiqués, des outils de regroupement sémantique ou conceptuel [Bourigault et Aussenac-Gilles, 2003]. Soit ces systèmes s'appuient sur des techniques purement linguistiques, soit ils combinent critères statistiques et analyses linguistiques.

La plupart des aides à l'identification de termes pour le français, comme LEXTER ou SYNTEX [Bourigault et Aussenac-Gilles, 2003], de classes sémantiques, comme Lexiclass [Assadi, 1998] ou de relations lexicales comme Prométhée [Morin, 1999] ou Caméléon [Séguéla, 2001] utilisent des connaissances sur la langue. Ce type de connaissances est aussi utilisé par les systèmes statistiques qui effectuent des décomptes sur des formes lemmatisées ou tenant compte des classes grammaticales ou même sémantiques dont relèvent les termes [Daille, 1994]. Les connaissances linguistiques utilisées correspondent à des ressources (dictionnaires sémantiques, lexiques spécialisés, listes de morphèmes ou de mots vides, etc.) ou à des règles d'analyse grammaticale, syntaxique ou lexicale. Ainsi, beaucoup de systèmes s'appliquent sur des textes étiquetés par des analyseurs grammaticaux, ce qui permet ensuite de mieux caractériser l'information linguistique.

Seuls quelques systèmes comme l'extracteur de termes ANA [Enguehard et Pantera, 1995] s'affranchissent de ce type de connaissances. A partir des régularités du corpus, ANA apprend des formes permettant d'identifier des termes. Il présente alors le double avantage de s'adapter à différentes langues et de pouvoir traiter des textes grammaticalement incorrects.

Les systèmes de TAL utilisés pour construire des ressources ontologiques reposent sur une hypothèse commune. On peut l'énoncer ainsi : en s'appuyant sur des informations au niveau syntaxique, il sera possible de suggérer à un analyste des regroupements ou des structurations sémantiques ou conceptuels. Ces principes communs renvoient aux travaux de Z. Harris, s'appliquant à des langages de spécialité dans des domaines bien délimités. Selon Harris, « Le sens donné à des entités et le sens des relations grammaticales entre celles-ci est lié à la restriction des combinaisons possibles de ces entités entre elles » [Harris, 1968] dans [Assadi, 1998]. Autrement dit, dans un domaine particulier, la distribution des mots et les configurations syntaxiques dans lesquelles ils se trouvent ne sont pas illimitées et sont révélatrices du sens de ces mots. C'est dans cette même optique que nous envisageons de définir un système multi-agent capable d'analyser des textes et de s'appuyer sur des régularités syntaxiques associées à l'usage des termes pour suggérer des éléments au niveau sémantique.

2.2.2 Des limites bien identifiées

Plusieurs analyses comme celles dressées au niveau national [Aussenac-Gilles et Condamines, 2003] ou international [Aussenac-Gilles et Maedche, 2002] font ressortir les perspectives nécessaires pour une meilleure diffusion et utilisation de ces résultats. Un premier point faible est le manque de compatibilité et d'intégration entre ces différents systèmes : il est difficile de combiner un extracteur de termes et un extracteur de relations par exemple. Ou encore, il serait souhaitable de pouvoir facilement récupérer leurs résultats sous des formats standards dans des environnements de modélisation de terminologies ou d'ontologies. Ainsi, des tentatives encore insuffisantes ont été réalisées entre SYNTEX, Caméléon et Terminae [Szulman *et al.*, 2002].

Une autre limite des logiciels actuels est leur caractère fermé, alors que l'on souhaiterait parfois n'utiliser qu'une partie des fonctionnalités, ou travailler sur des résultats intermédiaires ou encore adapter certaines fonctions à des besoins spécifiques.

Enfin, une dernière évolution souhaitée, conditionnée par les deux précédentes (meilleure modularité et meilleure intégration), est de pouvoir définir des chaînes de traitement adaptées à différents types d'applications. En effet, les besoins varient selon le type de ressource à construire et l'application dans laquelle elle va être utilisée [Aussenac-Gilles et Condamines, 2003]. Or il est encore complexe de mesurer les conséquences des choix méthodologiques (outils et approches d'analyse de la langue, type de structure de donnée) sur la qualité des résultats et l'apport de l'ontologie dans l'application ciblée. Des recherches sont en cours pour caractériser la manière de traiter le problème par type d'application. Ce n'est que grâce à l'accumulation de telles études que l'on pourra spécifier des méthodes adaptées à chaque type de modèle ontologique.

2.2.3 La maintenance : un problème mal géré

Un des arguments en faveur de la construction d'ontologies à partir de textes est d'améliorer la maintenance des modèles obtenus. En effet, lorsque le modèle conserve un lien vers les textes utilisés, ceux-ci fournissent des informations qui justifient la manière dont les connaissances ont été modélisées. Ils servent de documentation du modèle et, a priori, en facilitent la maintenance.

D'autres éléments importants à conserver sont les critères et principes d'organisation des concepts ainsi que la justification de la sémantique des relations définies.

La maintenance n'en reste pas moins délicate [Aussenac-Gilles et Condamines, 2003] : il est souvent difficile de retrouver la logique de l'analyste qui a choisi d'organiser les concepts d'une certaine façon, ou encore de savoir comment une évolution du domaine ou des besoins liés à l'application doivent se répercuter sur l'ontologie. Si on se réfère aux textes comme indicateurs de l'évolution des connaissances ou de la terminologie du domaine, le problème revient à savoir comment prendre en compte de nouveaux textes, comment les traiter (séparément ou avec les premiers textes), puis comment intégrer les connaissances extraites de ces textes dans la structure existante. Ce problème est d'autant plus compliqué que la structure est formelle car la formalisation suppose une représentation normalisée des connaissances qui éloigne du texte.

Aujourd'hui, les ontologies sont vues comme des structures figées, relativement stables, dont la construction requiert un processus long et coûteux. Implicitement, on s'attendrait donc à ce qu'elles nécessitent une maintenance réduite. Or pratiquement, en particulier dans les cas de domaines spécialisés innovants et évoluant rapidement, certaines ontologies doivent régulièrement être revues, surtout au niveau terminologique. Dans ce contexte, le scénario de construction de l'ontologie et celui de maintenance gagnent à être envisagés de manière cohérente. Ainsi, on peut se demander si les mêmes outils sont utilisables pour ces deux phases, et le prévoir dès la construction de l'ontologie. Finalement, les outils de construction devraient être sauvegardés avec l'ontologie de manière à la reconstruire à partir de nouveaux textes et selon un nouveau point de vue [Bachimont, 2001].

Dans ce même objectif de cohérence, une vision alternative commence à émerger en s'appuyant sur des logiciels de TAL et d'apprentissage. Cette tendance vise à construire le plus automatiquement possible des structures plus proches de réseaux terminologiques et donc moins complexes à élaborer. Ces structures, moins formelles, sont souvent suffisantes pour des applications comme la recherche d'information, l'indexation de documents ou la gestion des connaissances. Dans ce cas, construire ou maintenir ces structures revient à relancer le même processus sur une nouvelle version d'un corpus de référence.

2.3 Aspects concernés par ce mémoire

Le système à réaliser devra aussi bien mener des activités d'analyse syntaxique, d'extraction de termes, d'identifications de relations que des regroupements conceptuels pour mener à bien une construction d'ontologies à partir de textes. Toutefois dans le cadre de mon DEA, ce n'est pas un objectif accessible. C'est pourquoi il me faut me concentrer sur la toute première étape d'une chaîne de travail complète, à savoir l'étiquetage.

Malgré tout, il est important d'examiner la manière dont l'étiquetage est exploité ensuite afin d'essayer de prendre de bonnes décisions pour ce premier pas.

Une des contraintes voulue pour cette étude est l'indépendance envers une quelconque ressource linguistique extérieure (dictionnaire, lexique ou règles grammaticales) ce qui pousse naturellement à focaliser sur des techniques purement statistiques. Ce choix comporte le risque de déboucher sur un échec car il n'y a pas, à notre connaissance, de précédent et accomplir des analyses syntaxiques sans ressource d'ordre linguistique parait être un problème difficile. Mais, ce résultat négatif permettrait au moins d'écarter définitivement cette option. C'est pourquoi les réflexions qui suivront dans ce mémoire aborderont non seulement la question de l'indépendance envers les ressources, mais aussi celle du choix des techniques permettant au moins de limiter les ressources nécessaires.

Chapitre 3

Analyse syntaxique

3.1 Qu'est-ce que l'analyse syntaxique ?

Nous pouvons définir succintement l'analyse syntaxique comme l'art de segmenter un texte en composants, d'identifier ces composants et de les relier [Sabah, 1989]. Derrière cette définition simple se cachent un grand nombre de questions pointues et de problèmes à considérer. Ils pousseront à faire des choix qui donneront à l'analyseur résultant des forces mais aussi des faiblesses. La notion d'analyseur syntaxique couvre donc une grande diversité d'approches que de réponses à ces choix.

Par exemple, on peut légitimement se demander : comment traiter la ponctuation ? Faut-il d'ailleurs en tenir compte ? Comment segmenter correctement ? Un exemple de l'impact de ces choix est l'analyseur GREYC'98, examiné en détail dans [Lebarbé, 2002], qui commet des erreurs dues au fait que la virgule n'est pas prise en compte. Ne pas en tenir compte revient à devenir aveugle aux appositions et autres énumérations ce qui restreint d'autant la capacité de l'analyseur à effectuer de bonnes conclusions dans ces cas là.

Un autre exemple de difficulté est la segmentation en mots qui pose de nombreux problèmes à elle seule. Le cas de l'élision d'une voyelle est un problème typique de la discipline lors de cette segmentation. En effet, il se pose très vite la question de l'endroit où l'on doit couper lorsqu'une apostrophe est présente. En français, par exemple, l'observation d'une forme telle que « l'archive » pousserait à couper juste après l'apostrophe. Ce constat n'est malheureusement pas réutilisable en langue anglaise, qui présente un comportement moins régulier que le français sur ce point avec des formes comme « don't » et « Franck's » où l'on couperait respectivement avant la lettre *n* et avant l'apostrophe.

De plus, il existe toute une myriade de techniques d'analyse à base de grammaires. Nous pouvons y distinguer deux principales approches concurrentes. Premièrement, les grammaires syntagmatiques prenant leurs racines dans les travaux de Chomsky : cette famille regroupe les grammaires Lexical Functional Grammar (LFG), Generalised Phrase Structure Grammar (GPSG), Tree Adjoining Grammar (TAG), Head-driven Phrase Structure Grammar (HPSG) aujourd'hui très utilisées. Deuxièmement, les grammaires de dépendances issues des travaux de Tesnière : elles se distinguent des précédentes par leur mise en avant des relations entre mots par opposition à la structure en constituants de la phrase. Cette focalisation sur les relations entre mots provient d'une interprétation des travaux de Tesnière dont l'extrait le plus souvent cité est le suivant [Tesnière, 1959] :

« Tout mot qui fait partie d'une phrase cesse par lui-même d'être isolé comme dans le dictionnaire. Entre lui et ses voisins, l'esprit aperçoit des connexions, dont l'ensemble forme la charpente de la phrase. [...] Les connexions structurales établissent entre les mots des rapports de dépendance. »

On constate un retour de ces dernières approches depuis quelques années pour des analyses superficielles. On en trouve une présentation synthétique dans [Kahane, 2000]. Un des points importants en faveur des grammaires de dépendance est la quantité de ressources (lexiques ou règles) nécessaires pour mener à bien une analyse, qui peut être très réduite en comparaison des grammaires syntagmatiques. Il suffit pour s'en convaincre d'examiner les travaux menés dans l'équipe SYNTAXE du GREYC (voir en particulier le système HACTAR présenté au chapitre 4).

A la lumière des considérations précédentes, nous constatons que l'analyse syntaxique fait donc partie de ces domaines définissables aisément mais qui, malgré un énoncé naturel, comporte un nombre étonnant de ramifications et de difficultés qui lui sont propres.

3.2 Exemples d'objets de l'analyse syntaxique

Dans cette section, nous passerons en revue quelques objets manipulés par des outils d'analyse syntaxique. Nous n'avons pas la volonté d'être exhaustif, le domaine étant très vaste. C'est pourquoi seront présentés seulement trois objets qui nous paraissent intéressants pour leurs propriétés et raisons d'exister. On pourra s'étonner de ne pas trouver l'objet « mot » dans cet exposé. Il s'agit d'une omission tout à fait volontaire afin de ne pas donner une vision trop « grammairienne » de l'analyse syntaxique.

Les objets présentés ici et définis par la suite sont donc :

1. le chunk (aussi appelé syntagme minimal) qui est tout comme le mot une unité linguistique, ce qui garantit les fondements de traitements basés sur cette notion (en effet, une unité linguistique dispose de certaines propriétés comme le marquage prosodique, la délimitation par des bornes, ou encore la possibilité de typer l'unité) ;
2. le « segment » qui n'est pas une unité linguistique, mais comporte quand même quelques propriétés d'ordre linguistique (en particulier son existence du point de vue de la prosodie [Lebarbé, 2002]) ;
3. le « réseau terminologique » (au sens de [Bourigault, 1994]) qui est essentiellement un objet d'opérationnalisation et d'exploration de corpus, il n'a aucune existence syntaxique dans le sens où on ne peut le reconstituer simplement en observant la syntaxe phrastique ; sa mise en évidence suppose une analyse du corpus entier (contrairement aux deux précédents).

Enfin, le dernier intérêt de cette présentation est le contraste entre les deux premiers objets qui sont des supports de l'analyse syntaxique, tandis que le dernier en est un résultat.

3.2.1 Le chunk

Dans [Lebarbé, 2002], l'auteur considère la notion de « chunk » découverte par Abney comme un élément fondamental et acquis de l'analyse syntaxique. Il est aussi appelé syntagme minimal, et on en trouve une définition plus précise dans [Giguet, 1998] :

« Le syntagme minimal est une unité constituée d'une série de mots tous contigus les uns aux autres et regroupés autour d'une tête lexicale, le plus souvent un nom ou un verbe, plus rarement un adverbe, un adjectif ou un pronom. Toutes les relations syntaxiques entre mots internes au syntagme ont la particularité d'être calculables de manière positionnelle et non ambigüe. »

Dans [Lebarbé, 2002], le chunk est utilisé pour le calcul syntaxique, il n'est pas là pour supplanter la structuration en mot. Il s'agit juste d'un grain d'observation et d'analyse car il est plus simple d'utilisation pour le calcul des relations syntaxiques au sein de la phrase. La figure 3.1 illustre bien ce point, il suffit de constater que les relations au sein du chunk sont principalement des relations de dépendance vers la tête lexicale du chunk (en gras), et que les relations de dépendance entre chunks sont entre têtes lexicales.

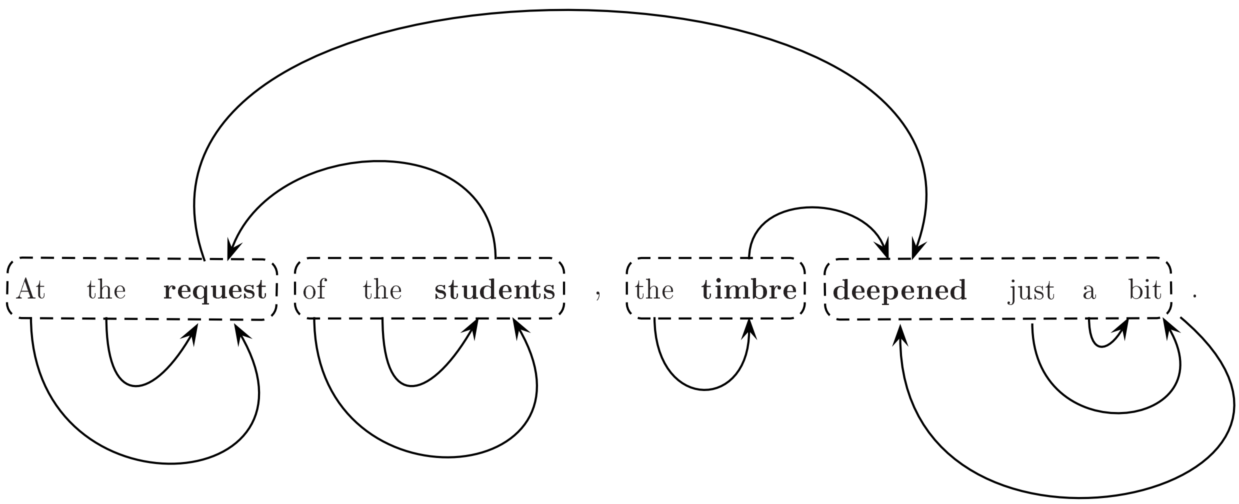

FIG. 3.1 – Exemple de délimitation de « chunks »

L'algorithme de découpage en chunks tient compte de marques morpho-syntaxiques pour fonctionner. Il est donc fortement dépendant de la langue considérée. Toutefois, il existe des recherches dont l'objet est la découverte automatique de ces marques morpho-syntaxiques par l'exploration d'un corpus d'apprentissage (voir à ce propos [Déjean, 1998] et [Houben, 2002]).

3.2.2 Le « segment »

Dans [Lebarbé, 2002], l'auteur s'est attaché à dégager une unité d'analyse intermédiaire entre le chunk et la phrase. Ainsi, il peut constituer une hiérarchie d'unités au sein d'un texte lui permettant d'effectuer des traitements selon différentes granularités. Ce travail a été fait en vue d'une amélioration des résultats de l'analyseur GREYC'98 qui fonctionne à base de chunks.

L'algorithme de segmentation est assez simple et repose entièrement sur la détection de bornes clairement définies. Ces bornes marquent une rupture dans la linéarité de la phrase et leur fonction est d'entretenir la logique phrastique [Lebarbé, 2002]. Il s'agit des conjonctions, pronoms relatifs et ponctuations non finales. Encore une fois, la représentation textuelle de ces bornes est changeante en fonction des langues et l'on peut se poser la question de l'existence d'une méthode de découverte de ces bornes à l'aide d'un corpus d'apprentissage.

Lebarbé mentionne différentes propriétés du segment, les plus intéressantes concernent les relations de dépendance entre les chunks associés, qui le composant peuvent servir de base à la construction de relations sémantiques. Tout d'abord, il simplifie la mise en relation des chunks internes au segment car il limite et simplifie la fenêtre de recherche des dépendants [Lebarbé, 2002]. En outre, il permet de mieux appréhender les relations de dépendance distantes au sein de la phrase.

3.2.3 Le « réseau terminologique »

Ce dernier objet présenté est tout à fait différent des précédents. En effet, il n'est pas une aide à l'analyse, mais le résultat d'une analyse syntaxique. Il s'agit du réseau terminologique déjà présent dans le logiciel d'extraction terminologique LEXTER [Bourigault, 1994] pour les syntagmes nominaux, puis étendu aux autres types de syntagmes (verbaux, adjectivaux...) dans SYNTEX, autre logiciel d'analyse terminologique.

On obtient ce réseau en exploitant les relations de composition en tête et en expansion des syntagmes d'un corpus. En effet, le but de cette analyse étant la mise en place d'un réseau de termes, on tire partie de la structure de ces termes. Or *« un terme complexe est constitué d'une tête (ou base), qui est le terme (représentant la classe) générique et d'une expansion (ou déterminant), qui est un complément adjectival ou nominal mentionnant un caractère spécifique »* [Bourigault, 1994]. Cette décomposition est évidemment récursive puisque la tête ou l'expansion d'un terme peut être à son tour un terme.

Dès lors que les décompositions des termes entre tête et expansion sont effectuées comme dans le tableau 3.1, il devient tout à fait possible d'extraire un réseau terminologique comme celui de la figure 3.2 [Bourigault, 1994].Ce réseau est structuré d'une manière pertinente pour un terminologue qui peut s'appuyer sur ce résultat pour définir des fiches terminologiques par exemple. Il peut aussi être utilisé de manière opérationnelle pour des applications telles que des moteurs de recherche.

Syntagme	**Tête**	**Expansion**
batterie froide	batterie	froide
batterie de refroidissement	batterie	de refroidissement
batterie chaude terminale	batterie chaude	terminale
batterie chaude	batterie	chaude
circuit de refroidissement	circuit	de refroidissement

Tab. 3.1 – Exemple de liens tête-expansion

3.3 La loi de Zipf

Dans cette section, nous nous attacherons à présenter une loi statistique particulière : la loi de Zipf. Il est légitime de s'étonner du fait qu'il n'y ait qu'une unique loi statistique examinée ici. La principale raison est qu'il s'agit de la seule loi exploitable et valable sans ressources extérieures. L'utilisation du modèle de Markov caché ou encore d'une méthode comme celle utilisée dans TreeTagger [Schmid, 1994], nécessite de disposer d'un dictionnaire d'associations entre des mots et les étiquettes auxquelles ils peuvent être associés ou encore de ressources constituées à partir d'un corpus d'apprentissage annoté à la main comme pour TreeTagger [Schmid, 1994]. Comme le but de notre travail de recherche était d'effectuer un état de l'art dans le domaine de l'analyse syntaxique, mais surtout de mener une étude de faisabilité d'analyses n'utilisant pas de ressources extérieures, ces méthodes ont été examinées puis écartées et non exploitées car elles ne correspondaient pas aux contraintes qui m'ont été fixées. C'est pourquoi notre attention s'est tournée vers la loi de Zipf.

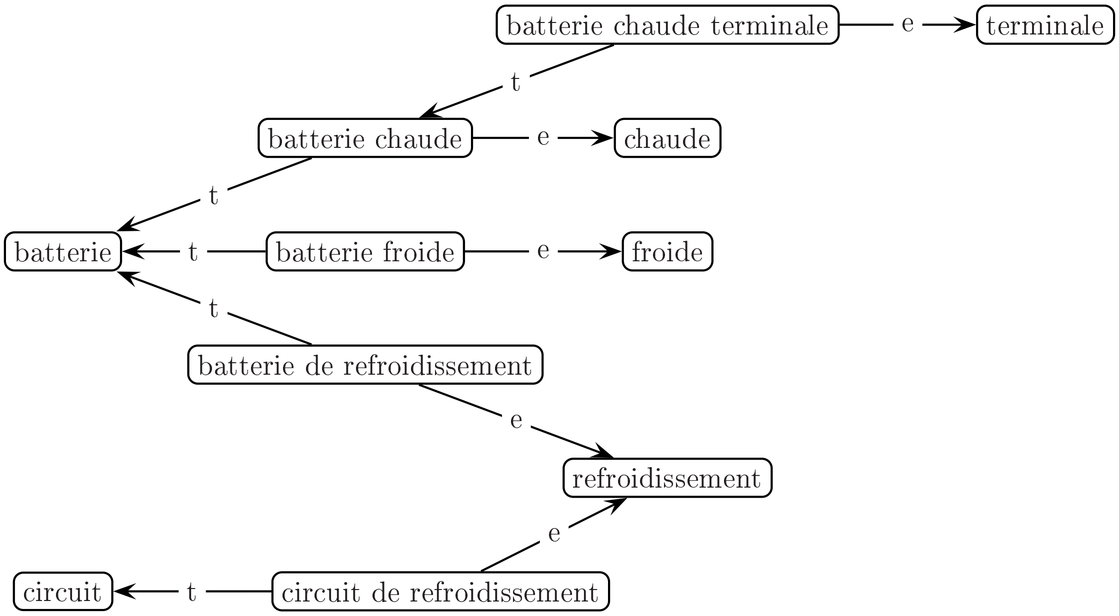

Fig. 3.2 – Exemple de réseau terminologique

3.3.1 Présentation

Initialement, il s'agit d'une loi statistique qui s'applique à des domaines aussi variés que la population des villes, la distribution des données dans une base de données, les prénoms, ou encore les occurrences des mots dans un texte. C'est ce dernier domaine qui sera particulièrement traité ici, mais pour des considérations plus complètes on pourra se reporter à [Poosala, 1997].

En étudiant l'ensemble des mots différents d'un texte quelconque et en les classant par fréquences décroissantes, on constate que ce classement vérifie l'égalité suivante :

$$f \times r = C \tag{3.1}$$

où f est la fréquence du mot, r son rang et C une constante. Il faut toutefois garder à l'esprit qu'il s'agit d'une approximation.

A partir de la précédente relation il est possible de tracer une hyperbole dans laquelle trois zones se distinguent [Houben, 2002]. La zone des mots grammaticaux et de la ponctuation est située dans le bras de l'hyperbole suivant l'axe des fréquences. La plupart des mots thématiques sont regroupés dans le « virage » de l'hyperbole. La zone des autres mots lexicaux se situe dans le bras de l'hyperbole suivant l'axe des rangs. Toutefois, les frontières de ces zones ne sont pas nettes et la distribution des mots à l'intérieur des recoupements n'est absolument pas connue.

De plus, il est important de noter que les mots grammaticaux regroupent environ 50% des mots du texte. A titre d'exemple, la figure 3.3 a été tracée à partir d'un corpus important de recettes de cuisines en français.

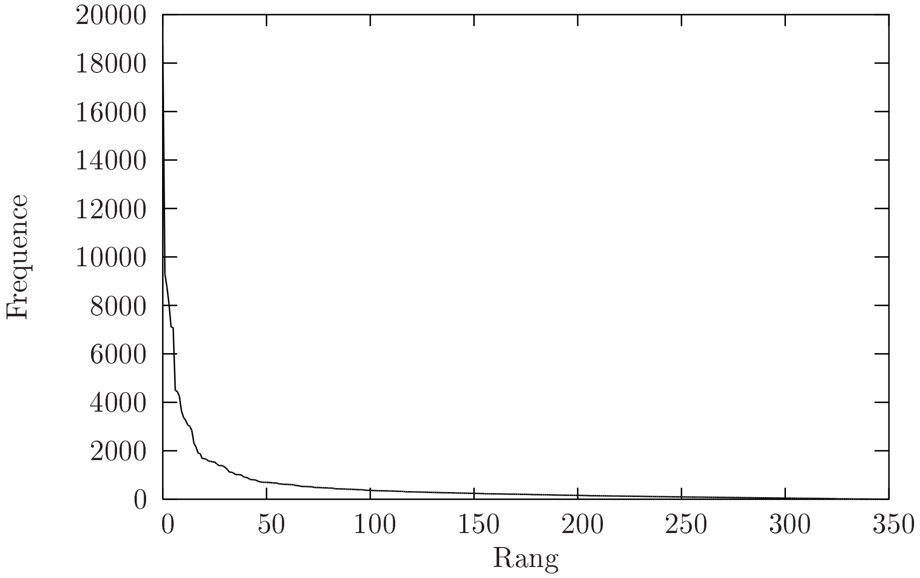

FIG. 3.3 – Fréquence/Rang dans un corpus de recettes de cuisine

Un des principaux intérêts de cette loi lorsqu'elle est appliquée telle que présentée ci-dessus est son indépendance envers les locuteurs, les types de textes et les langues.

3.3.2 Une brève justification

On trouve plusieurs hypothèses sur la théorie sous-jacente à la loi de Zipf. La plus intéressante par rapport à cette étude est celle de Mandelbrot [Poosala, 1997]. En effet, il a supposé que le but du langage est de transmettre le plus possible d'informations par symbole (dans le sens de la théorie de l'information selon Shannon) avec le moins d'effort. Il a alors obtenu la relation suivante :

$$f(r) = K(r+c)^{-\theta} \tag{3.2}$$

où $f(r)$ est la fréquence du mot et r son rang.

Il a été démontré par Booth que la relation précédente et la loi de Zipf sont équivalentes. On trouve donc ici une bonne justification théorique de la loi de Zipf dans le cadre d'une étude sur textes.

3.3.3 Utiliser cette loi pour l'extraction de termes

Dans [Vergne, 2003], l'auteur s'attache à décrire une méthode d'extraction de termes candidats qui exploite certaines propriétés de la loi de Zipf. Cette méthode est un bon exemple de ce que l'on peut réaliser par le biais de cette loi sans autre ressource linguistique que le corpus traité lui-même (on parlera alors de méthode « endogène »).

Dans la grande majorité des cas, s'intéresser uniquement à la fréquence des mots n'est pas un critère suffisant pour distinguer les mots lexicaux des mots grammaticaux. En effet, il est souvent difficile de trancher entre les mots grammaticaux rares des mots lexicaux fréquents. Pour cela, il faut s'intéresser aux fréquences des mots à certaines positions ce qui n'est pas exploité par la loi de Zipf. Or, une observation de Zipf lui-même est que les mots grammaticaux sont fréquents et courts tandis que les mots lexicaux sont plus rares et longs (par la loi d'économie d'effort dans un code). Dans la suite, on parlera des mots vides et des mots pleins pour désigner respectivement les mots grammaticaux et les mots lexicaux.

L'idée est donc de travailler sur les différences entre la fréquence **et** la longueur des mots. Le texte est alors une succession de mots vides et de mots pleins (en fait de mots non vides, car tout mot non vide est considéré comme plein). Voici un exemple issu de [Vergne, 2003] :

Manifestazioni per la pace in tutto il mondo

L'outil produit alors la sortie suivante :

Manifestazioni *per la* ***pace*** *in* ***tutto*** *il* ***mondo***

où les mots en gras sont des mots non vides. Le problème devient alors une détection de différences selon les critères suivants :

- différence de longueur en nombre de lettres : ***Manifestazioni*** *per* (14 lettres puis 3 lettre)
- différence de fréquence dans le corpus : *il* ***mondo*** (19 occurrences puis 3 occurrences)

Fort de ces hypothèses, il est possible d'extraire des termes candidats en détectant la séquence `plein - vide - plein` (i.e. `long - court - long` et `rare - fréquent - rare`), et la séquence `plein - vide - vide - plein`. Cette recherche est faite au sein du « virgulot », un segment de corpus délimité par des ponctuations [Vergne, 2003].

Cette méthode semble présenter de bons résultats et fonctionne sur corpus multilingue sans prétraitement ni ressource autre que le corpus lui-même (contrairement à d'autres méthodes comme celle appliquée par SYNTEX qui s'appuie sur une analyse syntaxique). En revanche, il faut bien noter qu'elle reste très spécialisée à l'extraction de termes mais n'est pas capable (en l'état) de distinguer des relations entre ces termes. En outre, elle a tendance à faire ressortir les groupes nominaux au détriment des groupes verbaux (qui ont tendance à être plus rares et variés).

3.4 Analyse syntaxique et Ingénierie des Connaissances

L'ingénierie des connaissances entretient des relations étroites avec l'analyse syntaxique dès qu'il s'agit d'extraire de la connaissance depuis des corpus textuels. En effet, la masse de données à traiter requiert une automatisation et une aide au repérage de structures pertinentes, pour la construction d'ontologies et de terminologies, dans le corpus :

- les noms et syntagmes nominaux pour dégager des concepts ;
- les relations syntaxiques et certains adjectifs pour extraire des propriétés ;
- certains verbes et des relations de causalités pour déterminer des axiomes.

L'extraction d'axiomes étant encore peu étudiée dans le domaine, nous distinguons trois tâches principales qui tirent parti de l'analyse syntaxique ou qui sont entièrement couvertes par des outils d'analyse syntaxique :

1. l'extraction de termes ;
2. l'identification de classes sémantiques ;
3. la mise en relation de ces classes.

3.4.1 Extraction de termes

L'extraction terminologique est une tâche déterminante pour l'ingénierie des connaissances lorsque l'on travaille à partir d'un corpus textuel. En effet, elle permet de repérer des syntagmes particuliers du corpus : les termes du domaine. Initialement, on ne s'intéressait lors de ce travail qu'aux syntagmes nominaux, mais à présent les outils permettent de repérer d'autres types de syntagmes (verbaux, adjectivaux...) également utiles pour les ontologies.

Cette extraction est faite en utilisant des outils d'analyse syntaxique et des techniques statistiques en exploitant leurs résultats. Il s'agit d'une tâche pointue et difficile car une extraction de termes peut générer beaucoup de bruit. Il faut donc faire attention à se recentrer réellement sur les termes du domaine. Pour cela, des critères de visualisation et de sélections des termes sont nécessaires, l'interface homme-machine de ce type d'outils est parfois une aide (voir par exemple LEXTER et SYNTEX qui fournissent des critères statistiques mais aussi directement un réseau terminologique). Dans ce type de travaux, il n'est pas rare de faire appel à des experts du domaine pour valider les extractions effectuées.

Un autre point qui peut énormément aider est le choix du corpus textuel exploité. En effet, dans [Aussenac-Gilles et Busnel, 2002], il s'agit d'une des difficultés rencontrées. Il a été nécessaire de scinder le corpus initial en trois sous-corpus car au milieu du domaine cible (l'industrie de la fibre de verre) se trouvaient représentées dans les textes choisis des considérations juridiques (dépôt de brevet) ou économiques (veille). Les termes extraits sont un moyen de valider et caractériser le corpus et inversement ces termes sont d'autant plus pertienents que le corpus est bien choisi.

Une fois toutes les précautions d'usage prises, ces termes peuvent devenir des noms de concepts (ou d'actions, ou de relations, généralement suivant la nature du syntagme) s'ils sont jugés pertinents pour le travail en cours.

3.4.2 Identification de classes sémantiques

Cette tâche permet d'aller plus loin que l'extraction de termes. Depuis un ensemble de termes extraits d'un corpus, on cherche à dégager des classes sémantiques. Mais identifier ces classes ne repose pas que sur une extraction de termes, elle requiert aussi de faire appel à des analyses plus élaborées comme le repérage des fonctions grammaticales et des dépendances syntaxiques. En effet, ces analyses syntaxiques préparent à une analyse distributionnelle « Harrissienne ».

Le sens des termes est en relation étroite avec leur distribution, c'est-à-dire la somme des environnements dans lesquels ils apparaissent : « Deux morphèmes qui ont des significations différentes diffèrent quelques part dans leur distribution » [Harris, 1968]. Ainsi, il est possible de classer les sens d'un terme en fonction des constructions où il est présent. Des distributions comparables entre deux termes sont le signe de parties de sens communes. C'est entre autre par ce type d'analyse que l'on franchit, modestement, la frontière entre le niveau syntaxique et le niveau sémantique.

On peut effectuer des regroupements de termes par l'étude de cooccurrences. Par ces regroupements il est possible de décrouvrir des relations de synonymie, d'hyponimie ou de méronymie. Toutefois, cette interprétation est toujours le fait d'un humain, les cooccurrences permettent seulement de voir qu'il existe une relation entre ces termes, pas de la caractériser.

Il existe des outils menant des analyses distributionnelles tout en exploitant des décompositions basées sur des patrons (Nom Préposition Nom, Nom Adjectif, Syntagme Adjectif...). Ces décompositions permettent de raffiner en partie les rôles joués par les termes composants ou encore de savoir quels seront les termes les plus intéressants pour la structuration en concepts, etc. On pourra se reporter à [Assadi, 1998] pour une présentation d'un outil de ce type utilisant entre autres les résultats issus de LEXTER.

3.4.3 Mise en relation

Enfin, une dernière tâche présentée ici, reposant sur des analyses syntaxiques préalables, est la mise en relation des concepts. Aux concepts déjà définis sont rattachés des termes dans le corpus. Or, les relations syntaxiques reliant ces termes sont souvent révélatrices des relations sémantiques entre les concepts correspondants. De plus l'analyse distributionnelle a permis de regrouper des termes sans pour autant caractériser précisément les relations qui les unissent. C'est pourquoi il nous faut les repérer en utilisant par exemple des patrons syntaxiques appelés marqueurs.

Ce type de travaux est abondamment présenté dans [Séguéla, 2001] et la méthode Caméléon. L'auteur y détaille aussi une méthode d'acquisition de marqueurs à partir d'un corpus technique proposée par Marti Hearst :

1. Sélectionner une relation lexicale R intéressante ;
2. Etablir une liste de termes entre lesquels R est identifiée (en utilisant par exemple des marqueurs génériques, non dépendants d'un domaine, mis au point à la main) ;
3. Trouver des endroits dans le corpus où les termes sont cooccurrents et enregistrer leurs environnements ;
4. Trouver les régularités dans ces environnements et faire l'hypothèse que ces régularités sont à la base de nouveaux marqueurs pour la relation R ;
5. Une fois un nouveau marqueur étudié et validé, l'utiliser pour trouver d'autres couples de termes en relations et revenir au point 2.

On voit donc toute l'importance de connaissances génériques préalables, les marqueurs spécifiques devant être découverts grâce aux marqueurs génériques initiaux. Enfin, rien n'est précisé sur la manière d'exprimer les marqueurs ni pour trouver les régularités. C'est pourquoi il existe de nombreuses approches provenant de cette proposition de Hearst, comme utiliser :

- des expressions régulières ;
- des patrons morpho-syntaxiques et/ou lexico-syntaxiques ;
- la structure des syntagmes ;
- etc.

Par exemple, la méthode Caméléon [Séguéla, 2001] utilisait de simples expressions régulières à ses débuts, mais les dernières versions de l'outil utilisent un concordancier. On dispose par ces programmes d'une grande souplesse dans l'expression des patrons de recherche.

3.5 Conclusions

Dans ce chapitre, nous avons fait un panorama de techniques d'analyse syntaxique pour finir sur des considérations liées à leur utilisation dans le cadre de l'ingénierie des connaissances. Il est toutefois légitime de se demander si cette présentation est suffisament large. En effet, il s'agit d'un domaine extrêmement vaste et il paraît bien difficile d'en faire un état de l'art exhaustif en un seul chapitre.

Malgré tout, au moins un point est frappant. Il n'existe pratiquement pas de lois statistiques exploitables pour faire de l'analyse sans ressources. Seule la loi de Zipf semble être une piste pour effectuer un étiquetage partiel d'un texte sans utilisation de ressources. Toutefois, dans le cas d'un échec d'un étiquetage sans aucune ressource et basé uniquement sur la loi de Zipf et certains constats dérivés, les approches à grammaire de dépendances semblent intéressantes car elle peuvent fonctionner avec des ressources réduites (ce qui semble un compromis acceptable), et sont plus facilement utilisables comme support pour les compétences utilisées dans un système multi-agent (contrairement aux grammaires syntagmatiques pensées pour des approches montantes ou descendantes éminemment séquentielles).

Enfin, la dernière section concernant l'utilisation de l'analyse syntaxique en ingénierie des connaissances renforce bien le sentiment que la route sera extrêmement longue pour construire un système multi-agent permettant d'aider à la construction d'ontologies à partir de textes. En effet, il existe de nombreux types d'analyse syntaxique à mener pour couvrir les besoins de la construction d'ontologies.

Chapitre 4

Architectures distribuées pour l'analyse syntaxique

4.1 Une longue histoire

Les architectures distribuées pour l'analyse syntaxique ont une histoire à peu près aussi longue que l'histoire des architectures distribuées en intelligence artificielle (IAD). En effet, un des premiers système d'IAD est aussi un système de TAL : HEARSAY II (1975).

Dans ce chapitre, nous allons présenter les principaux systèmes distribués pour l'analyse syntaxique. Nous nous attacherons à faire ressortir les caractéristiques saillantes de ces systèmes de TAL. Il nous a paru important d'insister sur leurs caractéristiques architecturales plutôt que sur les techniques et connaissances purement linguistiques présentes dans les agents de ces systèmes. En effet, ces techniques sont pour la plupart des techniques éprouvées du TAL pour mener des analyses syntaxiques. En revanche, nous nous sommes concentrés sur les collaborations nécessaires entre les différents types d'agents au sein de ces systèmes. Toutefois la dernière partie de ce chapitre présentera un comparatif qui abordera les ressources linguistiques nécessaires à ces systèmes.

Nous avons choisi de découper ce chapitre en trois parties correspondant aux trois grandes périodes du domaine :

1. les systèmes multi-expert ;
2. les systèmes tendant vers la décentralisation et présentant un découpage fonctionnel ;
3. les systèmes découpés selon une approche objet.

4.1.1 Premier âge : Multi-expert

HEARSAY II

HEARSAY II est un système de compréhension de la parole qui, à partir d'un signal sonore, reconstitue des requêtes d'interrogation de bases de données [Erman et Lesser, 1975], [Lesser et Erman, 1977]. Il s'agit du premier système de TAL à utiliser une architecture distribuée. Il dispose d'un vocabulaire de seulement 1000 mots, qu'il faudrait étendre pour valider l'architecture.

Ce système est constitué de douze sources de connaissances, d'un module de contrôle centralisé et d'une structure de données commune, appelée le tableau noir, dans laquelle la solution est construite. Les sources de connaissances travaillent à différents niveaux d'interprétation : le segment, la syllabe, le mot, la séquence de mots et la phrase. Elles ont un format condition/action et lorsque la partie condition est vérifiée, le module de connaissances infère des hypothèses et éventuellement modifie des informations dans le tableau. Le module de contrôle est composé d'un agenda, d'un moniteur et d'un ordonnanceur. Le moniteur repère les parties condition des sources de connaissances intéressées par les modifications effectuées au sein du tableau noir et remplit l'agenda avec ces sources de connaissances. L'ordonnanceur applique des heuristiques pour sélectionner la source de connaissances à activer.

CARAMEL

CARAMEL [Sauvage-Wintergerst, 1992] est un système multi-expert pour le traitement automatique de la langue naturelle fondé sur le modèle du tableau noir. Ce système est constitué d'agents cognitifs à granularité forte qui interagissent par l'intermédiaire d'un superviseur. Ce superviseur SIROP met en œuvre de manière centralisée le contrôle du système. Il détermine les modules à activer en fonction de l'état courant de la solution. Les différents modules de connaissances correspondent aux différentes expertises nécessaires telles que l'analyse de la phrase qui utilise des connaissances syntaxiques et sémantiques, la résolution de pronoms, la résolution d'ellipses, le dialogue, la génération de réponses pour l'utilisateur, etc. Ces modules sont constitués d'une base de connaissances écrites sous la forme de règles, d'un moteur d'inférences, d'une mémoire locale et d'une boîte aux lettres. Ses applications sont la compréhension de récits et la conception d'un standard automatique. Dans ce système, le parallélisme et la collaboration entre les modules de connaissances sont utilisés pour accélérer et améliorer la résolution.

4.1.2 Deuxième âge : Décentralisation et découpage fonctionnel

TALISMAN

Le but de TALISMAN est l'analyse morpho-syntaxique du français écrit. Il est décrit dans [Stefanini, 1993] et abondamment analysé puis amélioré dans [Warren, 1998]. Ce système permet de construire l'arbre syntaxique complet d'une phrase. Sur le plan architectural, il s'agit d'un système multi-agent gouverné par des lois. Il existe deux types de lois : locales aux agents et globales (valables dans tout le système). Il est découpé selon un axe fonctionnel, chaque agent correspondant à une activité de l'analyse de la phrase.

Les agents peuvent être regroupés en deux familles :

- Les agents « de niveaux » ont été définis en fonction de la structure de données sur laquelle porte leur analyse. Ils ont donc comme activité la reconnaissance de différents types d'unités linguistiques. Il s'agit des agents « Prétraitements », « Morphologie », « Syntaxe » et « Segmentation ».
- Les agents « transversaux » peuvent manipuler différents types d'unités. Ils doivent traiter de divers phénomènes linguistiques complexes. Leur intervention n'est pas toujours nécessaire, et ils sont souvent amenés à collaborer avec d'autres agents pour assurer la satisfaction de leurs buts. Il s'agit des agents « Transformation », « Ellipse », « Coordination », « Négation ».

Le fait d'avoir un système dans lequel différents agents sont amenés à collaborer pour mener à bien leurs activités d'analyse est extrêment intéressant pour fiabiliser les résultats. En effet, il n'est pas rare dans les systèmes séquentiels classiques de s'apercevoir que certains résultats dégagés par l'étape précédente deviennent erronés à la lumière des analyses de l'étape courante.

TALISMAN illustre parfaitement ce point puisque les agents sont amenés à collaborer pour lever des ambiguïtés.

MASPAR

MASPAR est un système multi-agent pour l'analyse syntaxique de textes arabes présenté dans [Aloulou *et al.*, 2003]. Il adopte une approche similaire à celle de TALISMAN en dégageant des agents responsables d'une activité d'analyse du texte. Ainsi, les agents « Segmentation », « Lexical », « Morphologie », « Syntaxe », «Anaphore », et « Ellipse » y sont implémentés. La ressemblance avec TALISMAN est donc frappante puisque ils sont répartis en trois agents « de niveaux » et deux agents « transversaux ». Toutefois, les agents de MASPAR utilisent des connaissances très différentes pour mener des analyses de l'arabe.

4.1.3 Troisième âge : Découpage objet

HACTAR

Ce dernier système est le plus récent que nous ayons étudié. Il est présenté dans [Lebarbé, 2002]. Le modèle d'agent utilisé est un modèle hybride réactif/cognitif. Ce système est original par rapport à ses prédécesseurs par son découpage « objet ». En effet, chaque agent ne correspond pas purement à une activité d'analyse syntaxique mais à une structure de l'analyse effectuée : mot, chunk, segment et phrase. Chacune de ces structures a des moyens d'introspection et peut interagir avec ses voisins.

Lorsqu'un groupe d'agents d'un même niveau hiérarchique dispose de suffisamment d'informations, il peut décider de fusionner en un agent de niveau supérieur, l'ancien groupe d'agents étant perdu, ce qui se justifie pour l'objectif visé d'analyse de texte. Selon nous, ce passage abrupt d'un niveau hiérarchique à un autre présente l'inconvénient de limiter les possibilités d'interaction entre les différents niveaux. Ainsi cette fusion est irréversible et ne permet pas aux agents de réviser les hypothèses conduisant à leur disparition. Une fois l'analyse achevée, il ne reste plus qu'un agent par phrase. Le système a alors identifié tous les constituants de la phrase et leurs relations grammaticales.

HACTAR est certainement le système le plus proche de notre façon d'appréhender le problème par son découpage innovant focalisé sur les structures analysées plutôt que sur les activités d'analyse. Toutefois, il reste selon nous encore du chemin à parcourir pour obtenir des systèmes adaptés aux problématiques de la création d'ontologies exploitant une telle décomposition.

4.2 Comparaisons des systèmes

Après l'historique des systèmes d'IAD pour le TAL précédent, il est tentant de le compléter par des comparatifs selon différents points de vue. Nous avons donc effectué un comparatif en terme d'architecture et un comparatif en terme de ressources linguistiques utilisées.

4.2.1 Comparaison en architecture

Les critères architecturaux retenus pour comparer les systèmes étudiés sont récapitulés dans le tableau 4.1. Nous situons ci-dessous les différents systèmes présentés dans la partie précédente relativement à chacun de ces critères.

Système	Contrôle	Communication	Constituants
HEARSAY II	Centralisé	Tableau noir	Experts (découpage fonctionnel)
CARAMEL	Centralisé (plusieurs niveaux)	Tableaux noirs	Experts (découpage fonctionnel)
TALISMAN	Système gouverné par des lois	Envois de messages	Agents cognitifs
MASPAR	?[a]	Envois de messages	Agents cognitifs (découpage fonctionnel)
HACTAR	Décentralisé	Actions sur l'environnement	Agents mixtes (découpage objet)

[a]Nous n'avons pas trouvé d'information claire sur ce point. Toutefois, vue la forte similitude avec TALISMAN, nous pouvons supposer qu'il s'agit d'un système gouverné par des lois.

TAB. 4.1 – Comparaison des systèmes, point de vue architecture

Contrôle

Le seul système véritablement décentralisé est HACTAR. A l'opposé, HEARSAY II est complètement centralisé par le biais d'un gestionnaire de tâches et un moniteur sur le tableau noir. Entre les deux se trouvent CARAMEL et TALISMAN qui ont un contrôle « mixte ». En effet, le premier dispose d'un superviseur et de sous-superviseurs et le second est un système contrôlé par le biais de lois locales et globales.

Communication

La grande majorité des systèmes présentés utilisent une communication indirecte (c'est-à-dire par modification de l'environnement). Seul TALISMAN emploi l'envoi de messages directs, conjointement à un protocole d'apprentissage.

Constituants

Deux familles de systèmes sont présentées ici, les systèmes multi-experts (HEARSAY II et CARAMEL) et les systèmes multi-agents. La terminologie peut changer d'un système à un autre. Par exemple HEARSAY II voit ses experts qualifiés de « sources de connaissances » tandis que CARAMEL dispose de « processus ». De plus, les systèmes multi-agents peuvent être conçus avec un découpage fonctionnel (TALISMAN et MASPAR) ou objet comme HACTAR. Enfin, les agents sont classés cognitifs ou réactifs. Ici nous n'avons présenté que des SMA utilisant des agents cognitifs ou mixtes (réactifs et cognitifs). Cette caractéristique du système HACTAR provient du modèle d'agent APA (Anticipation par Perception Augmentée) utilisé où chaque agent est lui-même un système multi-agent.

4.2.2 Comparaison en ressources linguistiques utilisées

Les ressources linguistiques considérées pour comparer les systèmes étudiés sont présentées dans le tableau 4.2.

Système	Lexique	Morphologie	Syntaxe
HEARSAY II	Vocabulaire de 1000 mots	?	Grammaire hors-contexte
CARAMEL	Lexique de 15000 mots	Analyse morphologique flexionnelle	Grammaire de cas et Grammaire systémique
TALISMAN	Lexique de 50000 mots	Analyse morphologique flexionnelle	Grammaire hors-contexte de constituants
MASPAR	Lexique	Règles de dérivation	Head-Driven Phrase Structure Grammar
HACTAR	Lexique restreint (mots grammaticaux, environ 400)	Règles morphologiques	Grammaire de dépendance

TAB. 4.2 – Comparaison des systèmes, point de vue ressources linguistiques utilisées

Lexique

Tous les systèmes présentés utilisent un lexique plus ou moins important. Nous n'avons pas d'information sur la taille exacte du lexique utilisé par HEARSAY II. Mais il dispose d'une faible couverture de la langue (1000 mots) qui nous fait dire que la ressource employée doit disposer d'un nombre d'entrées du même ordre de grandeur. CARAMEL et TALISMAN utilisent des lexiques imposants. Enfin on remarquera le lexique extrêmement restreint de HACTAR contenant uniquement 400 mots, mais lui permettant toutefois de disposer d'une large couverture de la langue dans ses analyses.

Syntaxe

Les grammaires mises en oeuvre pour l'analyse syntaxique sont variées (grammaire hors-contexte, HPSG, grammaire de dépendances) ce qui donne aux différents systèmes des capacités différentes. Ainsi, la grammaire de cas et la grammaire systémique utilisées par CARAMEL lui permettent de mener simultanément l'analyse syntaxique et l'analyse sémantique.

Enfin, un constat interessant provient de la corrélation entre grammaire de dépendances (au plan syntaxique) et découpage objet (au plan architectural). En effet, cela renforce le sentiment mentionné dans la section 3.5 que les grammaires de dépendances sont plus adaptées pour mener des fonctions d'analyse syntaxique à l'aide d'un système multi-agent (et non pas à partir d'un agent encapsulant le traitement comme dans les systèmes à découpage fonctionnel).

4.3 Conclusions

De l'examen de ces systèmes de TAL exploitant des SMA [Ottens *et al.*, 2004], nous tirons les conclusions suivantes.

En premier lieu, la grande majorité de ces systèmes s'appuie sur une architecture multi-agent calquée sur les fonctions d'analyse TAL disponibles dans des systèmes plus conventionnels. Cette organisation transparaît clairement par la présence d'un nombre réduit d'agents cognitifs aux traitements très sophistiqués. De plus, les agents utilisent une quantité de ressources extérieures importante. Seul le système HACTAR semble s'éloigner de cette tradition TAL en repensant totalement la structure des agents qui est calquée sur la structure des éléments de phrase analysés ; toutefois, il conserve un fort couplage avec des ressources linguistiques.

En second lieu, ils ne répondent pas aux besoins de la construction et de la maintenance d'ontologies. Mener conjointement construction et maintenance suppose une plus grande dynamique des analyses de textes permettant d'enrichir l'ontologie sans que l'utilisateur ne soit obligé de manipuler des résultats intermédiaires aux niveaux les plus fins de l'analyse linguistique. Or, les systèmes que nous venons de présenter effectuent un travail phrase à phrase exhaustif, sans aller jusqu'à une vision d'ensemble du texte nécessaire pour nos travaux. De plus, ils ne présentent pas d'aspect dynamique, une fois l'analyse effectuée, les conclusions ne sont pas réutilisables pour améliorer les analyses futures nécessaires à la maintenance de l'ontologie. On doit donc utiliser ces SMA comme des systèmes de TAL « classiques » pour lesquels l'analyste examine les résultats de l'analyse pour les transformer manuellement en éléments de son ontologie.

Enfin, nous notons que ces systèmes conçus pour mener des analyses linguistiques de surface ne fournissent pas de résultat de plus haut niveau. Par exemple, ils ne s'attachent pas à effectuer des analyses plus poussées comme l'analyse distributionnelle « à la Harris ». Il s'agit clairement d'un manque à combler : parvenir à un réseau d'agent qui facilite la construction d'ontologies. Toutefois, l'ajout de niveaux d'analyse doit s'accompagner de précautions tenant compte des problèmes cités précédemment afin de réellement apporter une valeur ajoutée par rapport aux techniques de TAL utilisées aujourd'hui dans le domaine des ontologies.

Cependant, on peut en retenir un certain nombre de points positifs. Les architectures distribuées ont d'ores et déjà permis d'améliorer les résultats obtenus dans leurs analyses (en particulier une meilleure robustesse des outils et l'élimination de certaines erreurs des outils classiques), et cela, grâce à une meilleure collaboration entre les différents niveaux d'analyse. Ainsi les « frontières » sont moins nettes que dans un système purement séquentiel et modulaire où chaque niveau est isolé des autres. Nous avons aussi constaté que l'on retrouvait, dans ces systèmes, les qualités inhérentes aux SMA et en particulier une extrême modularité. Nous prendrons en exemple TALISMAN II [Warren, 1998] qui obtient de meilleurs résultats que son ancêtre principalement en améliorant la communication entre certains agents.

Tout ceci nous mène à penser que la pertinence des architectures distribuées en TAL n'est plus à prouver. Ce sont les qualités présentées ci-dessus que nous voudrions préserver tout en tenant compte des manques actuels par une approche ambitieuse, basée sur la théorie des AMAS, et adaptée aux travaux sur les ontologies.

Chapitre 5

Vers un système multi-agent de construction d'ontologies

5.1 La théorie des « Adaptive Multi-Agent Systems » (AMAS)

Le but de cette théorie est de permettre à un système utilisant des critères de réorganisation locale au niveau de ses entités d'atteindre l'adéquation fonctionnelle. Par « adéquation fonctionnelle », nous nous référons au comportement global du système qui doit être adapté à la tâche courante. La théorie des AMAS se base sur le théorème suivant démontré dans [Camps *et al.*, 1998] :

Théorème 1 *Pour tout système fonctionnellement adéquat, il existe au moins un système à milieu intérieur coopératif qui réalise une fonction équivalente dans le même environnement.*

Un système possède un « milieu intérieur coopératif » lorsqu'il n'existe plus en son sein d'incompréhension, d'ambiguïté, d'incompétence, de conflit, de concurrence ou d'inutilité entre les entités composant ce système. La démonstration de ce théorème repose sur un axiome et quatre lemmes. Notre but n'est pas de refaire cette démonstration en détails, mais de présenter succintement cet axiome, ces quatres lemmes et leurs conséquences.

Axiome 1 *Un système fonctionnellement adéquat n'a aucune activité antinomique ou indifférente sur son environnement.*

Une activité antinomique sur l'environnement est une activité allant à l'encontre des intérêts de l'environnement.

Lemme 1 *Tout système coopératif est fonctionnellement adéquat.*

Car par définition un système coopératif n'a pas d'activités antinomiques ou indifférentes. Donc l'ensemble des systèmes coopératif est inclus dans l'ensemble des systèmes fonctionnellement adéquats.

Lemme 2 *Pour tout système S fonctionnellement adéquat, il existe au moins un système coopératif S^* qui soit fonctionnellement adéquat dans le même environnement.*

Ce lemme lève l'incertitude restante vis-à-vis de l'existence d'un système coopératif équivalent pour chaque système fonctionnellement adéquat dans le même environnement. Sans ce lemme, l'intérêt porté aux systèmes coopératifs serait amoindri puisque pour certains systèmes fonctionnellement adéquats, il ne serait plus possible de trouver un système coopératif équivalent.

Lemme 3 *Tout système à milieu intérieur coopératif est un système coopératif.*

Donc l'ensemble des systèmes à milieu intérieur coopératif est inclus dans l'ensemble des systèmes coopératifs. Il reste donc comme précédemment à s'assurer que chaque système coopératif dispose d'un système à milieu coopératif équivalent.

Lemme 4 *Pour tout système coopératif, il existe au moins un système à milieu intérieur coopératif avec une fonction équivalente dans le même environnement.*

La dernière incertitude étant levée, il est alors possible de définir une application surjective de l'ensemble des systèmes fonctionnellement adéquats vers l'ensemble des systèmes à milieu intérieur coopératifs. Ces derniers systèmes constituants un sous-ensemble de l'ensemble des systèmes fonctionnellement adéquats.

Ainsi le théorème central de la théorie des AMAS permet de se focaliser uniquement sur une classe de systèmes pour obtenir des systèmes fonctionnellement adéquat dans un environnement donné (voir figure 5.1).

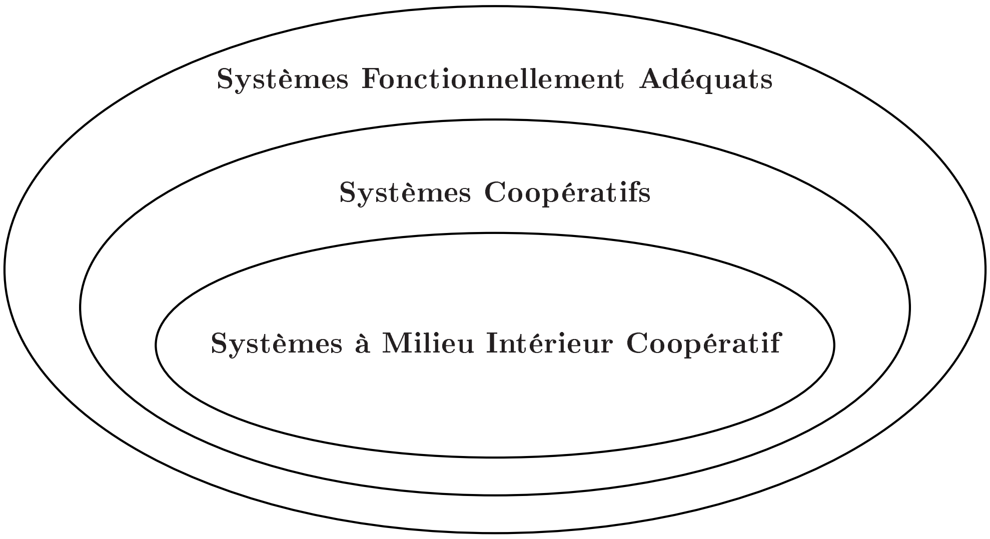

FIG. 5.1 – Relation entre les systèmes fonctionnellement adéquats et les systèmes à milieu intérieur coopératif

Les systèmes à milieu intérieur coopératif est la notion centrale de la théorie des AMAS. Cette théorie s'intéresse donc en particulier aux relations entre les entités constituantes des systèmes. Il devient alors clair qu'elle est aisément applicable aux systèmes multi-agents. La figure 5.2 présente comment doit s'adapter d'un tel système [Georgé *et al.*, 2003].

Le concepteur s'intéresse uniquement aux agents P_i qui produisent chacun leur fonction f_{P_i} et leur donne le moyen de décider de changer les liens les unissant. Ce sont ces liens qui règlent comment se combinent les fonctions f_{P_i} qui engendrent la fonction globale f_S. Ainsi en fonction des interactions du système avec son environnement, l'organisation des agents se modifie pour faire face aux perturbations de l'environnement. Le système s'adapte à tout moment pour produire une nouvelle fonction f_S.

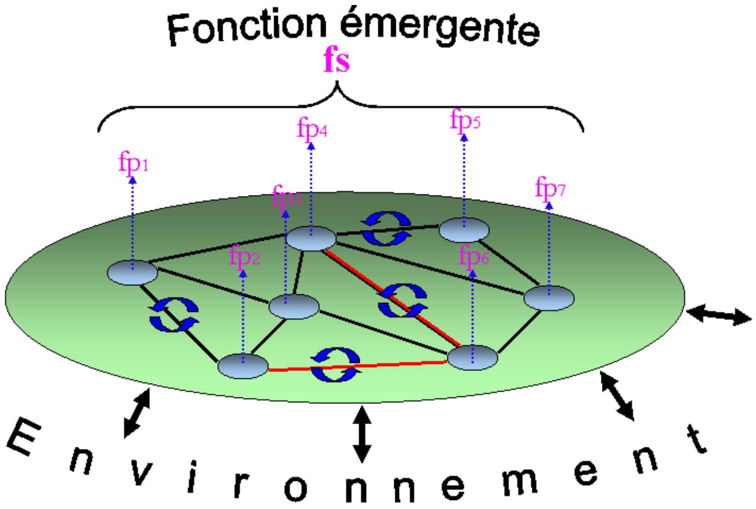

FIG. 5.2 – Adaptation par auto-organisation

5.2 Une théorie appliquée aux systèmes multi-agents

Malgré son nom, il est tout à fait envisageable d'appliquer la théorie des AMAS qui est basée sur la coopération pour d'autres paradigmes que celui du multi-agent. D'ailleurs, nous avons entrepris de ne mentionner les systèmes multi-agents que vers la fin de la section précédente. Il s'avère pourtant que c'est ce paradigme qui est aujourd'hui le mieux adapté à une telle théorie.

Pour pourvoir l'appliquer, il existe une spécification générale de l'architecture d'un agent AMAS. Un tel agent dispose en général des éléments suivants pour mener son travail à bien :

1. des connaissances d'un domaine particulier permettant à l'agent de réaliser sa fonction partielle ;
2. une représentation de lui-même, des autres et de l'environnement utilisant la notion de confiance sur ses croyances ;
3. une attitude sociale coopérative, c'est-à-dire qu'il dispose de critères locaux lui permettant de savoir s'il est ou non coopératif et d'agir en conséquence ;
4. un langage d'interaction (par messages ou par l'environnement) ;
5. des capacités de raisonnement sur ses connaissances.

L'élément le plus important du point de vue de la coopération est ici l'attitude sociale de l'agent. Si l'on prend un point de vue d'assez haut-niveau, la détection des situations non-coopératives (SNC) est exprimable en utilisant l'équation suivante [Georgé *et al.*, 2003] :

$$\textbf{NON COOPERATION} = \neg C_1 \vee \neg C_2 \vee \neg C_3 \tag{5.1}$$

- $\neg C_1$ **Incompréhension** : Un signal est incompris ou comporte des ambiguïtés. Dans ce cas, l'agent va tenter de transmettre l'information à d'autres agents potentiellement plus compétents ou de se faire aider par autrui afin de lever les ambiguïtés.
- $\neg C_2$ **Défaut de raisonnement** : Une information reçue est déjà connue ou n'a aucune conséquence logique. L'agent ne peut en tirer profit et va donc chercher d'autres agents pouvant en bénéficier.
- $\neg C_3$ **Inutilité de l'action** : D'après ses croyances courantes, l'agent considère que la transformation qu'il peut opérer n'est pas bénéfique à autrui (conflit ou concurrence).

Un système plongé dans un environnement dynamique va forcément devoir traiter des situations imprévues du point de vue local de l'agent. Cet imprévu sera pris en compte comme une SNC au niveau de l'agent qui va chercher à modifier ses relations avec les autres. Les fonctions partielles du système se réorganisent, changeant la fonction globale, pour réagir à la situation imprévue et la prendre en charge. Ainsi, un agent doit non seulement réaliser sa fonction partielle, mais aussi agir sur l'organisation du système pour résoudre les SNC. Ces décisions sur la coopération sont locales aux agents et ne doivent pas être dictées par la fonction globale (qui évolue au cours du temps et en fonction des imprévus).

5.3 Quelques problèmes traités avec des agents AMAS

Une telle théorie nécessite d'être expérimentée, d'autant plus que l'adéquation fonctionnelle est un concept sujectif qui ne peut être validé que par un observateur dans un contexte donné. C'est pourquoi, nous présentons trois systèmes pour lesquels la théorie des AMAS a été appliquée avec succès [Georgé *et al.*, 2003]. Nous n'avons pas ici pour but de détailler avec précision chacune des solutions retenues, mais simplement de présenter quelques problèmes résolus aux caractéristiques distinctes.

Tout d'abord, le système FORSIC est un système de recherche d'informations dans un cadre d'e-éducation. A partir des formateurs inscrits, le système pour répond aux requêtes d'un utilisateur cherchant des cours dans un domaine donné. L'intérêt d'une approche AMAS dans ce cas est la possibilité de tenir compte de la dynamique des besoins des utilisateurs et des services offerts. Les descriptions de formations et les requêtes utilisateurs sont saisies en texte libre. On exploite alors les co-occurrences, qui sont souvent un critère statistique suffisant pour la recherche d'informations, et des réajustements par petits incréments ou décréments en fonction des situations non coopératives détectées.

Un deuxième exemple est le système STAFF de prévision des crues. Il s'agit là d'un problème très complexe dépendant de nombreux paramètres : topographie, superficie, pluviométrie... De plus, ces caractéristiques ont tendance à évoluer avec le temps. Le système a donc été développé en tirant parti de la coopération, il est entraîné en utilisant des historiques de crues des stations sur lesquels il est installé. Il effectue un réajustement permanent relativement aux mesures d'entrée, aux résultats des autres agents et à l'erreur de prévision commise. Cette réorganisation se fait par petites variations sur les coefficients d'une somme pondérée utilisée par chaque agent pour conclure.

Enfin, ABROSE est un système de courtage électronique de services sur internet. Il travaille sur les accointances des agents que ces derniers exploitent pour faire suivre les requêtes utilisateur à des agents jugés plus compétents lorsqu'ils ne peuvent pas satisfaire complètement les besoins. L'utilisateur a la possibilité de critiquer les résultats du système en les notant « bon », « mauvais » ou « peu importe ». Cette critique est alors utilisée par les agents pour modifier leurs accointances si nécessaire.

Afin de rester dans cette obligation de confrontation de la théorie des AMAS à des domaines variés, il parait naturel de tenter de l'adapter à des domaines complexes comme la création d'ontologies à partir de textes ou le traitement automatique des langues.

5.4 Notre proposition

L'objectif général est de construire et de maintenir des ontologies en contexte dynamique en s'appuyant sur des textes. Pour cela, un outil doit construire de manière automatique un réseau conceptuel à partir de textes. Tout d'abord, il doit interagir avec l'analyste pour que ce dernier valide les résultats du système et lui permette éventuellement de se corriger. Dans cette phase, il fera des propositions de réseau de type conceptuel à l'analyste à partir du résultat de son analyse de textes. Puis, l'ontologie sera construite à partir des résultats de l'outil. En phase de maintenance, le système fera des propositions de modifications du réseau conceptuel existant en fonction de l'information présente dans les anciens textes et les nouveaux textes fournis. Du point de vue des agents, il n'y aura pas de réelle différence entre la phase de construction et la phase de maintenance, construire le réseau conceptuel initial revient à proposer des modifications sur un réseau conceptuel vide.

Les hypothèses de travail de départ, qui pourront évoluer en cours de projet, sont de trois ordres. Tout d'abord, la première hypothèse consiste à avoir une utilisation minimale de ressources externes et à rester le plus indépendant possible de la langue naturelle. La deuxième hypothèse est de permettre à l'analyste d'accéder via une interface à un réseau conceptuel proposé par le système et lui permettre de valider les résultats ou de forcer certains aspects du réseau en question. La dernière hypothèse réside en l'utilisation de systèmes multi-agents comme outil de construction du réseau à partir de textes. Les systèmes multi-agents considérés sont basés sur la théorie des AMAS.

Chaque agent sera doté de règles de comportement locales (linguistiques, lexicales) lui permettant de développer et de modifier des liens avec d'autres agents au fur et à mesure de l'analyse du texte ou de l'acquisition de nouveaux textes. La validation des liens que développe un agent est locale au niveau de l'agent. Sur ce processus de création, plusieurs questions restent ouvertes : nature et importance de connaissances linguistiques que doivent posséder les agents, niveau de granularité des agents (mot, segment...), nature des connaissances entre agents...

Ensuite, des parties du réseau vont être sélectionnées et présentées à l'analyste pour validation au travers d'un outil interface. La sélection tiendra compte de la nature et de l'importance des liens entre agents. La validation consiste à confirmer ou à critiquer les caractéristiques syntaxiques, grammaticales, ou sémantiques que l'agent aura identifiées sur lui-même ou sur les agents avec lesquels il communique. Ces interactions entre l'analyste et le système permettent au système multi-agent de se modifier pour s'adapter à ce que l'analyste souhaite. La validation donnée par l'analyste se fait ici à un niveau macro et aura une conséquence au niveau micro des agents. Les connaissances au sein des agents évoluent pour constituer un meilleur réseau. Ces connaissances seront ensuite utilisées quand le réseau sera augmenté en phase de maintenance lorsque de nouveaux textes seront à traiter. En utilisant les propriétés des systèmes multi-agents adaptatifs, le réseau se constitue puis se stabilise.

Utiliser la technique multi-agent pour concevoir une ontologie est un point original car comme nous l'avons vu, peu de travaux utilisent les agents pour l'analyse syntaxique et il n'en existe aucun à notre connaissance pour la construction d'ontologies. En général, les ontologies sont manipulées par les agents pour leur permettre de se comprendre dans leurs interactions [Gandon, 2002]. A l'inverse, c'est la technique multi-agent qui est ici au service du traitement automatique de la langue pour la construction d'ontologies.

L'objectif de construction d'ontologies étant bien trop vaste dans le cadre d'un DEA, notre travail cette année s'est limité à une des toutes premières étapes de ce type de tâche : l'étiquetage. De plus, il s'agit d'un étiquetage simplifié mais qui cherche à ne tirer partie d'aucune ressource linguistique (lexique ou règles). Ce sont ces travaux que nous allons maintenant examiner plus en détails.

Chapitre 6

Réalisation & Évaluation

6.1 Comportement général du système

Notre système est un système multi-agent composé d'agents réactifs. Leur structure interne est donc très simple. Ils ne sont pas situés (ne sont pas plongés dans un environnement disposant d'un repère normé) et ne considèrent que leur environnement social c'est-à-dire les autres agents. Toutefois, ce milieu est initialement constitué d'une manière particulière, dépendante du corpus textuel. Au démarrage du système, le corpus textuel segmenté (l'algorithme de segmentation utilisé est simple et considère par exemple « don't » comme un seul mot) est parcouru, et pour chaque nouveau mot rencontré, un agent est créé pour représenter ce mot (nous disposons à la fin de ce processus d'un seul agent par forme d'un mot présente dans le texte). Chacun de ces agents se voit attribuer comme réseau d'accointances l'ensemble des mots présents dans tous les virgulots[4] où il apparaît aussi. Le tout reste partitionné par virgulot, et l'ordre phrastique est conservé. Pour un agent, accéder à une de ces connaissances revient à se demander quel est le n-ième agent à droite ou à gauche de sa m-ième occurrence (cf. figure 6.1).

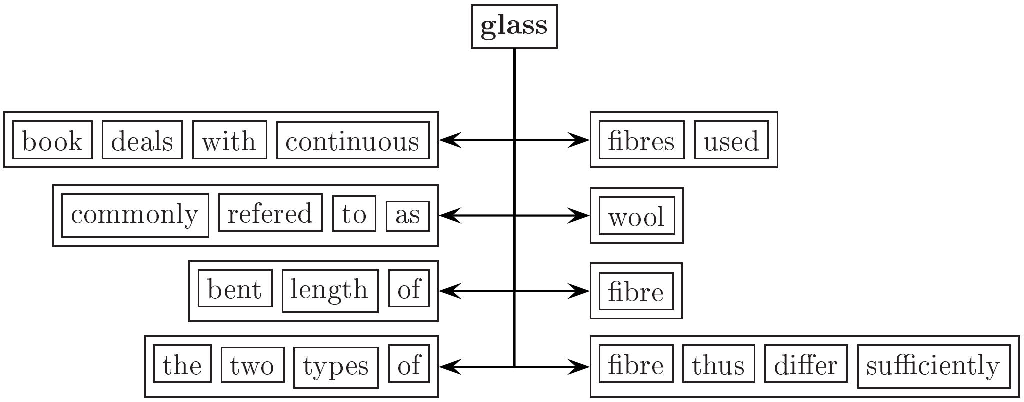

FIG. 6.1 – Exemple de réseau d'accointances créé pour l'agent « glass » par un extrait du corpus sur l'industrie de la fibre de verre contenant quatre occurrences du mot en question

Le fonction globale du système étant d'effectuer un étiquetage simplifié (on ne distingue que

[4] tout segment de corpus situé entre deux ponctuations

trois catégories : fonctionnel, nom et verbe) de textes en anglais, les agents disposent d'un jeu de confiances qu'ils doivent attribuer à trois rôles différents, en plus de leurs accointances. Ils peuvent modifier ces confiances selon plusieurs critères qui seront détaillés par la suite. Le rôle en lequel l'agent a le plus confiance à un instant donné est appelé rôle dominant. La somme des confiances des rôles est toujours de 100% pour chaque agent. Il n'apporte donc aucune information au système s'il a une égale confiance dans chacun de ses rôles. Le but local d'un agent est d'avoir des confiances compatibles avec celles de ses voisins, et il doit donc les recalculer chaque fois que cela est nécessaire.

D'ores et déjà cette « nécessité » de recalcul n'est pas une notion triviale et doit être sujette à caution : quand est-il réellement opportun de recalculer son rôle ? Dans notre système la demande de recalcul est émise par un agent à ses accointances lorsqu'il a changé ses propres confiances et que les agents qu'il connaît doivent utiliser cette information. Deux heuristiques ont été testées pour traiter ceci :

1. la requête est émise lorsque le rôle dominant a changé ;
2. la requête est émise lorsque le rôle dominant a changé **ou** que la confiance de ce rôle a subi une variation de $+/-10\%$.

Trois rôles ont été définis : « nom », « verbe » et « fonctionnel ». Ce choix a été guidé à la fois par la notion de chunk qui est constitué d'un mot lexical (ici pris au sens restreint : nominal ou verbal) entouré de mots fonctionnels (au sens large : déterminants, pronoms, adverbes, adjectifs...) et par notre but à long terme de construction d'ontologies pour lesquelles les structures nominales et verbales sont les plus importantes à identifier au sein du corpus.

Ainsi, chaque agent cherche à déterminer sa confiance en chacun de ses rôles en fonction des confiances de ses accointances. Le système s'arrête lorsqu'il a atteint un état stable et pour chaque mot du corpus un ou plusieurs rôles dominants sont choisis. Maintenant que le comportement général du système est défini, détaillons les compétences que les agents mettent en oeuvre pour ajuster leurs confiances.

6.2 Compétences des agents

Pour ce travail, nous nous sommes concentrés uniquement sur la langue anglaise, le corpus de test étant écrit dans cette langue. Il s'agit là d'une difficulté supplémentaire, l'anglais étant une langue disposant d'une stylistique plus variable que le français. Il est toutefois clair qu'une évolution de notre système devrait au moins prendre en compte le français et l'anglais.

6.2.1 Analyse contextuelle

La compétence minimum des agents (non désactivable) est la capacité d'analyse contextuelle de leur environnement gauche. Cette capacité est nécessaire, car l'environnement gauche d'un mot prend une part importante dans la détermination de la catégorie grammaticale de ce mot. Intuitivement, si un locuteur prononce le mot « the » qui est un article, on s'attend à ce que le mot suivant soit un nom. En prenant un point de vue centré sur l'agent « mot » on peut traduire cette compétence par : « mes précédents me déterminent, et je détermine mes suivants ».

Un marquage manuel a été effectué sur des extraits de corpus en anglais (cf. l'exemple figure 6.2), choisis pour être suffisamment généraux sur les styles observés.

Ce marquage manuel a bien entendu été effectué en utilisant uniquement les catégories précédemment définies à savoir : fonctionnel, nom et verbe. Sur ce marquage nous avons ensuite

---

I. A BRIEF HISTORY
The possibility of drawing heat-softened glass into fine fibres was known to glassmakers in antiquity and is older than the technique of glass blowing. Many Egyptian vessels were made by winding coarse glass fibres onto a mandrel of clay, suitably shaped, heating the assembly until the glass fibres flowed into one another and, after cooling, removing the clay core.

Légende : Fonctionnel, Nom, Verbe

FIG. 6.2 – Exemple de marquage manuel d'un extrait de corpus

passé un analyseur calculant des statistiques basées sur le contexte gauche de chaque mot (jusqu'à une distance de quatre mots). Ainsi nous avons dressé quatre tableaux (cf. 6.1) donnant les probabilités d'appartenance à une catégorie d'un mot précédent à une distance donnée sachant la catégorie du mot courant.

Précédent	F	N	V
F	46,43%	66,92%	67,35%
N	22,62%	22,56%	20,41%
V	15,48%	4,51%	0,00%
P	15,48%	6,02%	12,24%

Distance 1

Précédent	F	N	V
F	43,25%	57,14%	42,86%
N	17,46%	19,55%	30,61%
V	10,32%	10,53%	4,08%
P	28,97%	12,78%	22,45%

Distance 2

Précédent	F	N	V
F	39,29%	42,86%	44,90%
N	16,67%	19,55%	12,24%
V	7,94%	12,03%	4,08%
P	36,11%	25,56%	38,78%

Distance 3

Précédent	F	N	V
F	35,71%	38,35%	24,49%
N	13,10%	18,80%	14,29%
V	6,75%	7,52%	12,24%
P	44,44%	35,34%	48,98%

Distance 4

Légende : Fonctionnel (F), Nom (N), Verbe (V), Ponctuation (P)

TAB. 6.1 – Répartition des probabilités

Ces probabilités sont ensuite utilisées par l'agent pour calculer les confiances en chacun de ses rôles possibles. L'algorithme utilisé (cf. algorithme 6.1) permet d'obtenir une confiance pour chacun des rôles sur l'ensemble des contextes où le mot est présent. Les règles supplémentaires mentionnées à la ligne 1 de l'algorithme 6.1 sont optionnelles et présentées en détail dans les sections 6.2.3 et 6.2.4. Elles tiennent compte du contexte courant C et modifient le tableau *role_local*.

L'algorithme général traduit le fait que l'agent tente de maximiser les chances de ses voisins dans leurs propres rôles dominants. C'est pourquoi nous utilisons des tables de probabilités d'un rôle du voisin sachant le rôle de l'agent courant tout en examinant successivement l'ensemble des rôles possibles pour cet agent.

6.2.2 Détermination des confiances initiales

Une autre capacité des agents est une introspection simple permettant de déterminer les confiances initiales des agents. Cette compétence exploite une information globale fournie à chaque agent une fois que le réseau est constitué. Il s'agit de la fréquence du mot au sein du

Algorithme 6.1: Calcul des confiances d'un agent en fonction de ses contextes

Données : ensemble des contextes
Résultat : tableau *role* des confiances en chaque rôle
$n \leftarrow 0$
pour chaque *contexte* C **faire**
$n \leftarrow n + 1$
pour *d allant de 1 à 4* **faire**
pour chaque *rôle* R *dans l'ensemble* $\{F, N, V\}$ **faire**
$A \leftarrow$ unité à la distance d à gauche de l'agent dans le contexte C
si A *est une ponctuation* **alors**
$proba[R][d] \leftarrow table[d][R][P]$
sinon
$m \leftarrow 0$
$acc \leftarrow 0$
$role_A \leftarrow$ tableau des confiances de A
pour chaque *rôle dominant* R' *de* A **faire**
$m \leftarrow m + 1$
$acc \leftarrow acc + table[d][R][R'] \times role_A[R']$
fin
$proba[R][d] \leftarrow \frac{acc}{m}$
fin
fin
fin
$role\_local[F] \leftarrow \frac{proba[F][1]+\frac{2}{3}proba[F][2]+\frac{1}{2}proba[F][3]+\frac{1}{4}proba[F][4]}{1+\frac{2}{3}+\frac{1}{2}+\frac{1}{4}}$
$role\_local[N] \leftarrow \frac{proba[N][1]+\frac{2}{3}proba[N][2]+\frac{1}{2}proba[N][3]+\frac{1}{4}proba[N][4]}{1+\frac{2}{3}+\frac{1}{2}+\frac{1}{4}}$
$role\_local[V] \leftarrow \frac{proba[V][1]+\frac{2}{3}proba[V][2]+\frac{1}{2}proba[V][3]+\frac{1}{4}proba[V][4]}{1+\frac{2}{3}+\frac{1}{2}+\frac{1}{4}}$
appel de règles supplémentaires
normaliser $role\_local$
$role[F] \leftarrow role[F] + role\_local[F]$
$role[N] \leftarrow role[N] + role\_local[N]$
$role[V] \leftarrow role[V] + role\_local[V]$
fin
$role[F] \leftarrow \frac{role[F]}{n}$
$role[N] \leftarrow \frac{role[N]}{n}$
$role[V] \leftarrow \frac{role[V]}{n}$

corpus et son classement par rapport à l'ensemble des mots du corpus. Bien évidemment nous avons cherché à tirer parti de la loi de Zipf. Deux lois de répartition des confiances des agents en fonction de leur rang ont été implémentées afin d'être évaluées. Leurs comportements sont sensiblement différents.

Premièrement, l'algorithme 6.2 fait décroître la confiance pour le rôle fonctionnel en fonction du rang en partant d'une valeur forte de 70% de confiance. Symétriquement, les confiances des rôles de nom et verbe sont augmentées chacune pour la moitié de la quantité. Cette variation se fait sur les dix premiers pour cent des agents, puis l'attribution de confiances devient stationnaire favorisant très légèrement les rôles de nom et verbe.

Algorithme 6.2: Initialisation basée sur Zipf 1

Données : rang R de l'agent, nombre total T d'agents
Résultat : tableau *role*
si $R < T \times 10\%$ **alors**
  $val \leftarrow 70\% \times \frac{(R-T)^2}{T^2}$
sinon
  $val \leftarrow 25\%$
fin
$role[F] \leftarrow val$
$role[N] \leftarrow 50\% - \frac{val}{2}$
$role[V] \leftarrow 50\% - \frac{val}{2}$

Deuxièmement, l'algorithme 6.3 fait décroître la confiance pour le rôle fonctionnel, et augmenter la confiance dans les rôles nom et verbe comme l'algorithme précédent. En revanche, il démarre plus haut avec 80% de confiance. Cependant, cette confiance devient stationnaire lorsque la loi de décroissance utilisée passe sous la barrière des 33.33% de confiance. On garantit toutefois que cela arrivera avant les cinq premiers pour cent des agents.

Algorithme 6.3: Initialisation basée sur Zipf 2

Données : rang R de l'agent, nombre total T d'agents
Résultat : tableau *role*
$val \leftarrow 80\% \times \frac{5\% \times T - R}{5\% \times T}$
si $val < \frac{1}{3}$ **alors**
  $role[F] \leftarrow \frac{1}{3}$
  $role[N] \leftarrow \frac{1}{3}$
  $role[V] \leftarrow \frac{1}{3}$
sinon
  $role[F] \leftarrow val$
  $role[N] \leftarrow 50\% - \frac{val}{2}$
  $role[V] \leftarrow 50\% - \frac{val}{2}$
fin

6.2.3 Régulation locale de population

Le marquage manuel et le comptage effectués ont aussi permis de calculer au niveau global les proportions des différents types de mots dans un texte. Cette répartition est comme suit :

- Fonctionnels : 58,06%
- Noms : 30,65%
- Verbes : 11,29%

Nous avons alors décidé de tester un critère local pour maintenir ces proportions. La taille maximale d'un contexte étant de neuf mots, une approximation peut être de simplement compter le nombre de mots par catégorie au sein d'un contexte. Comme ce critère est local, nous n'avons pas cherché à complètement forcer les proportions mais à s'assurer que chacune des catégories ne dépassait pas localement sa quantité moyenne.

Ainsi dans un contexte d'environ dix mots, l'agent pénalisera (*i.e.* réduira la confiance) les rôles qui peuvent lui faire dépasser le seuil maximum du contexte (à savoir six fonctionnels, trois noms et un verbe). L'algorithme 6.4 a été utilisé à cette fin. Cette règle lorsqu'elle est activée est appelée à la ligne 1 de l'algorithme 6.1.

Algorithme 6.4: Regulation de population

Données : contexte courant C, tableau *role_local*
Résultat : tableau *role_local* modifié
$total[F] \leftarrow$ nombre de mots de C dont le rôle dominant est fonctionnel
$total[N] \leftarrow$ nombre de mots de C dont le rôle dominant est nom
$total[V] \leftarrow$ nombre de mots de C dont le rôle dominant est verbe

si $total[F] > 6$ **alors** pénaliser $role\_local[F]$
si $total[N] > 3$ **alors** pénaliser $role\_local[N]$
si $total[V] > 1$ **alors** pénaliser $role\_local[V]$

6.2.4 Analyse de frontières

Cette dernière compétence tire parti de la loi de Zipf augmentée de la considération sur la longueur des mots (cf. 3.3). Dans ce cas, un mot dont au moins un des voisins est moins fréquent et plus long que lui est un bon indicateur que ce mot est un fonctionnel (en fait une sous-classe des fonctionnels). De manière duale, un mot dont au moins un des voisins est plus fréquent et plus court que lui réduit ses chances d'être fonctionnel.

Cette notion est implémentée dans l'algorithme 6.5. Il s'agit d'un algorithme dont le comportement peut-être modifié suivant le sens attribué à l'opérateur $<_\alpha$, pour les expérimentations deux variantes ont été testées. La première utilise $<_\alpha$ comme un opérateur de comparaison numérique classique (définition absolue). La seconde définit $<_\alpha$ de manière relative, c'est-à-dire que $x <_\alpha y$ est vrai si et seulement y est supérieur à x de 10%. Cette règle lorsqu'elle est activée est appelée à la ligne 1 de l'algorithme 6.1 selon une de ses deux variantes.

6.3 Le code

Notre implémentation est basée sur un ancien système développé au sein de mon équipe d'accueil. Le but de cet ancien système étant très différent (recherche d'informations) seuls le grain des agents (un agent par mot) et l'implémentation général du système multi-agent ont été conservés. En particulier, ce dernier ne tire pas parti du parallélisme car beaucoup d'agents sont

Algorithme 6.5: Analyse de frontières - cadre général

Données : contexte courant C, tableau $role\_local$
Résultat : tableau $role\_local$ modifié
$l \leftarrow$ longueur du mot courant
$f \leftarrow$ fréquence du mot courant

$A \leftarrow$ première unité à gauche de l'agent dans le contexte C

$l_A \leftarrow$ longueur de A
$f_A \leftarrow$ fréquence de A

si $f_A >_\alpha f$ *et* $l_A <_\alpha l$ **alors**
    pénaliser $role\_local[F]$
sinon
    si $f_A <_\alpha f$ *et* $l_A >_\alpha l$ **alors**
        récompenser $role\_local[F]$
    fin
fin

$A' \leftarrow$ première unité à droite de l'agent dans le contexte C

$l_{A'} \leftarrow$ longueur de A'
$f_{A'} \leftarrow$ fréquence de A'

si $f_{A'} >_\alpha f$ *et* $l_{A'} <_\alpha l$ **alors**
    pénaliser $role\_local[F]$
sinon
    si $f_{A'} <_\alpha f$ *et* $l_{A'} >_\alpha l$ **alors**
        récompenser $role\_local[F]$
    fin
fin

créés (de l'ordre de la dizaine de milliers sur notre corpus de test), une *thread* par agent serait donc une consommation de ressources trop importante vis-à-vis du gain obtenu. L'implémentation utilise donc un parallélisme simulé, chaque agent est représenté par un objet du système disposant d'une méthode appelée chaque fois qu'il doit s'exécuter. Les agents en attente d'exécution (ici ceux ayant reçu une requête de recalcul) sont placés dans une liste (sans doublon) et, à chaque fin d'exécution d'un agent, un nouvel agent est tiré au hasard de cette liste afin d'être exécuté à son tour.

Enfin, les compétences ont été implémentées selon ce qui a été présenté dans les sections précédentes. Les sections suivantes s'attacheront à présenter comment nous avons mené les expérimentations sur ces compétences et dresse un bilan des résultats obtenus.

En résumé, un agent de notre système est implémenté comme un objet disposant d'une méthode publique déclenchant son exécution ainsi que de méthodes privées pour chacune de ses compétences. Ces méthodes modifient les valeurs de trois variables membres représentant la confiance de l'agent pour chacun des rôles fonctionnel, nom et verbe.

6.4 Démarche expérimentale

Il est bien sûr nécessaire de chercher à évaluer les compétences que nous avons dégagées, introduisent-elles beaucoup d'erreur ? quel type d'erreurs favorisent-elles ? etc. C'est pourquoi un certain nombre d'expérimentations ont été conduites en activant certains types de compétences (cf. tableau 6.2). Et nous avons ainsi cherché à dégager un ensemble de compétences donnant un bon compromis au niveau des résultats.

Cette série de douze expérimentations doit être vue comme deux séries de six expérimentations. La deuxième série utilisant des compétences plus raffinées (utilisation de l'analyse de frontières relative et la deuxième loi d'initialisation) et un recalcul plus fin de chaque agent. Hormis ces raffinements, les deux séries sont tout à fait similaires. La seconde constitue donc une évolution de la première série d'expérimentation. Enfin, au sein de chaque série nous avons cherché à évaluer comment fonctionnait les agents uniquement avec l'analyse de contexte puis nous avons testé la loi d'initialisation seule, ainsi que les comportements respectifs et les interactions entre l'analyse de frontières et la régulation de population. Enfin chaque série se termine avec une expérimentation mettant en jeu l'ensemble des compétences disponibles.

Un corpus anglais sur l'industrie de la fibre de verre a été utilisé comme corpus de test. Une fois les résultats sur ce corpus obtenus pour les douze combinaisons de compétences des agents, nous les avons examinés selon un point de vue qualitatif mais aussi quantitatif.

6.5 Résultats obtenus

6.5.1 Qualitatif

L'examen qualitatif des résultats est basé sur un petit ensemble de mots choisis pour leurs propriétés. Pour chaque expérimentation, nous avons comparé la catégorie dominante donnée par le système à la catégorie attendue. Le tableau 6.3 donne une synthèse de cet examen. Les mots correctement identifiés lors d'une expérimentation se voient gratifiés d'une croix dans la colonne correspondante.

Les mots choisis peuvent être regroupés en cinq groupes :

1. les noms du domaine très fréquents : *glass* et *fibre* ;

N°	Heuristique recalcul	Confiance initiale	Régulation population	Analyse frontières
1	Rôle dominant	Aucune	Aucune	Aucune
2	Rôle dominant	Loi 1	Aucune	Aucune
3	Rôle dominant	Aucune	Activée	Aucune
4	Rôle dominant	Aucune	Aucune	Absolue
5	Rôle dominant	Aucune	Activée	Absolue
6	Rôle dominant	Loi 1	Activée	Absolue
7	$+/-10\%$	Aucune	Aucune	Aucune
8	$+/-10\%$	Loi 2	Aucune	Aucune
9	$+/-10\%$	Aucune	Activée	Aucune
10	$+/-10\%$	Aucune	Aucune	Relative
11	$+/-10\%$	Aucune	Activée	Relative
12	$+/-10\%$	Loi 2	Activée	Relative

Tab. 6.2 – Contenu des expérimentations

2. des noms du domaine bien moins fréquents : *yarns*, *yarn*, *device* et *sand* ; *yarns* et *yarn* ont été choisis pour vérifier s'il n'y avait pas de décorellation entre un singulier et son pluriel, et *sand* parce qu'il est le moins fréquent de tous (il n'apparaît que six fois dans le corpus) ;
3. des fonctionnels : *or*, *continuous* et *secondary* ;
4. des verbes courts : *is*, *used* et *chopped* ; *is* étant très fréquent ;
5. des verbes plus longs : *manufactured*, *published* et *purified* ; qui ont un autre point commun, celui d'être assez peu fréquents.

Mot	Attendu	1	2	3	4	5	6	7	8	9	10	11	12
glass	Nom	X			X	X		X			X	X	
fibre	Nom	X			X	X		X			X	X	
yarns	Nom	X			X	X		X	X		X	X	X
yarn	Nom	X			X	X		X	X		X	X	X
device	Nom		X		X	X	X				X	X	X
sand	Nom	X	X		X	X	X	X	X		X	X	X
or	Fonctionnel	X	X	X	X	X	X	X	X	X	X	X	X
continuous	Fonctionnel		X	X			X			X			
secondary	Fonctionnel	X		X				X		X			
is	Verbe												
used	Verbe												
chopped	Verbe												
manufactured	Verbe												
published	Verbe												
purified	Verbe				X	X					X	X	

Tab. 6.3 – Résultats qualitatifs : tableau de synthèse

Le constat qui saute immédiatement aux yeux est à quel point le système est aveugle aux verbes. Seul *purified* est détecté comme verbe lorsque l'analyse de frontières est activée. Nous remarquons aussi les bons résultats du système sur les noms lorsque l'analyse de frontières est activée en conjonction ou non avec la régulation de population.

La régulation de population employée seule donne de bons résultats sur les fonctionnels même s'ils sont longs. En revanche, l'analyse de frontières dégrade cette performance sur les fonctionnels longs lorsqu'elle est ajoutée. D'une manière générale, l'analyse de frontière influence plus fortement les résultats que la régulation de population.

6.5.2 Quantitatif

L'examen quantitatif des résultats exploite les catégories données par le système dans son ensemble. Nous les comparons à la catégorie grammaticale de ce mot telle que la retroune un étiqueteur. Nous avons utilisé TreeTagger comme étiqueteur de référence en le considérant comme parfait (bien que sa précision soit autour de 96%). La sortie produite sur notre corpus a ensuite été filtré en mettant en correspondance les catégories de TreeTagger avec notre ensemble fonctionnel, nom et verbe. Ensuite nous avons effectué un tirage aléatoire pour l'attribution des catégories. Enfin, les résultats de nos expérimentations et de ce tirage aléatoire ont été confrontés à ceux de TreeTagger. Nous avons mesuré l'écart de précision entre nos expérimentations et le tirage aléatoire. Les mesures obtenues sont récapitulées dans le tableau 6.4.

N°	Réussite	Echec	Ecart avec aléatoire
1	35,25%	64,75%	5,39%
2	38,76%	61,24%	8,89%
3	27,61%	72,39%	-2,25%
4	39,40%	60,60%	9,54%
5	38,26%	61,74%	8,39%
6	39,95%	60,05%	10,08%
7	35,24%	64,76%	5,38%
8	34,54%	65,46%	4,67%
9	27,78%	72,22%	2,08%
10	39,22%	60,78%	9,35%
11	38,15%	61,85%	8,29%
12	38,80%	61,20%	8,94%
Tirage aléatoire	29,86%	70,14%	0,00%

TAB. 6.4 – Résultats quantitatifs : tableau de synthèse

On voit alors que le gain de notre système par rapport à un tirage aléatoire est tout à fait minime. Le meilleur gain étant aux alentours de 10%. De plus si l'on considère les expérimentations 4, 5, 10 et 11 qui donnent de meilleurs résultats sur les noms du domaine (ce qui nous intéresse plus particulièrement pour la construction d'ontologies), le gain est encore plus faible et tourne autour de 9%.

6.6 Bilan

Les bilan de ces expérimentations est plutôt négatif. En effet, notre système est en total échec sur le repérage des verbes et même pour les autres catégories nous n'avons pu dégager une combinaison de compétences stabilisant nos résultats. Enfin à une échelle plus macroscopique nous avons constaté que notre système est tout juste plus précis qu'un tirage aléatoire.

Ces constats nous poussent aussi à nous poser la question de la qualité des statistiques utilisées. On peut légitimement se demander s'il n'aurait pas été utile de considérer encore plus d'échantillons, voire utiliser une procédure automatique de constitution de ces statistiques à partir de textes étiqueté à l'aide de TreeTagger. Toutefois cela reviendrait à constituer des ressources en utilisant un corpus d'apprentissage important pour réutilisation ultérieure, ce que l'on voulait éviter initialement.

Chapitre 7

Conclusions & Perspectives

7.1 Résultats d'expérience négatif

Ce travail de DEA a permis d'effectuer une étude de faisabilité d'une approche statistique et endogène d'un étiquetage simplifié de corpus textuel basée sur un système multi-agent. Ce problème s'inscrit dans une volonté plus générale de construction d'ontologies à partir de textes utilisant le même type de système.

Malheureusement, nos expérimentations ont confirmé le risque pressenti à la section 3.5. Nos résultats sont à peine meilleurs qu'un tirage aléatoire si l'on prend TreeTagger comme étiqueteur de référence. N'oublions pas qu'il s'agit juste d'un premier pas qui devra se compléter d'autres niveaux d'analyse. A priori, ajouter des compétences syntaxiques ou un niveau sémantique supplémentaires, par exemple, peut améliorer les résultats. Mais ces niveaux devront pouvoir exploiter ces mêmes résultats pour leurs analyses et je pense que nous partons d'un niveau de qualité trop bas pour pouvoir construire quoi que ce soit par dessus.

Cela provient peut-être de la qualité des statistiques utilisées comme nous l'avons déjà soulevé, ou encore du fait que certaines formes peuvent appartenir à plusieurs catégories et ainsi propager des erreurs importantes s'il s'agit d'un mot fréquent. Mais le réel problème se situe certainement ailleurs. Nous ne disposons pas de critères statistiques en nombre suffisant pour être exploitables. De plus, améliorer les statistiques existantes passerait nécessairement par un système d'apprentissage sur des corpus textuels préparés. Il s'agit là d'une constitution de ressources que nous voulions éviter, mais les statistiques existantes constituent déjà une forme de ressources. A utiliser des ressources, par apprentissage, ou directement implémentées dans le système, autant s'inspirer directement d'approches éprouvées (même s'il sera nécessaire de les adapter et repenser pour un paradigme multi-agent) avec des connaissances linguistiques de qualité.

Enfin, il est juste de chercher à tirer des enseignements sur le côté purement multi-agent du problème, en particulier en considérant la théorie des AMAS. En effet, le fait que ce système soit complètement coopératif ou même émergent est à mon avis discutable. Certes, il en tire parti puisque les compétences des agents ont été pensées dans un but coopératif. Leur capacité d'analyse contextuelle telle qu'elle a été spécifiée va dans ce sens, puisqu'un agent cherche à être compatible avec les conclusions de ceux qui le déterminent plutôt que de mettre une forte confiance dans un rôle qui forcerait à remettre en cause trop d'agents de son voisinage. Toutefois cette compétence pousse nos agents à être altruistes ce qui les éloigne un peu d'une situation idéale de coopération.

On pourra considérer qu'il n'est pas suffisamment coopératif et que certaines situations non coopératives ne sont pas résolues. Toutefois, il est extrêmement difficile d'évaluer ce qu'est une situation non coopérative entre agents représentant des mots. Dans certains systèmes une métaphore anthropomorphique peut aider à les dégager. Ce type de réflexion semble ici impossible à transposer.

J'insisterais sur le fait qu'une analyse syntaxique se positionne sur un type de problème discontinu. De petits réajustements de la part d'un agent peuvent remettre en cause toutes les conclusions faites par les agents situés dans son voisinage social. La difficulté du problème devient extrêmement ardue à traiter à l'aide de la théorie des AMAS qui fonctionne bien sur les problèmes continus (cf. section 5.3) les agents cherchant en permanence à se réajuster pour converger vers une solution acceptable. Dans un tel cadre, il est plus évident de mesurer l'apport de chaque agent et un retour de l'environnement forçant les agents responsables d'une mauvaise solution de reconsidérer leurs calculs est envisageable. Mais ici il est impossible de savoir quels agents sont responsables sans disposer de la solution du problème.

Finalement, je pense qu'il est extrêmement ardu de se positionner sur un tel problème sans utiliser de ressources. De plus, l'enjeu d'un tel système me semble être à un niveau d'analyse plus haut.

7.2 Vers où se diriger maintenant ?

Comme nous venons de le voir, il paraît sage de recentrer nos objectifs sur une analyse syntaxique plus classique et de s'intéresser en parallèle à des niveaux plus élevés d'analyse linguistique et conceptuels pour nous focaliser sur l'objectif de construction d'ontologies. Il paraît très difficile de remettre en cause tout un pan du traitement automatique des langues simplement en le transposant au paradigme multi-agent.

Les solutions ayant à mon avis le plus de potentiel seront celles basées sur les grammaires de dépendances plus facilement transposables dans un système multi-agent. Il peut être intéressant de considérer un système tel que HACTAR comme base. Un tel socle paraît envisageable pour développer le nécessaire concernant la construction d'ontologies en tirant parti de la théorie des AMAS.

Une piste possible est d'intégrer la notion de confiance aux grammaires de dépendances. La plupart des systèmes basés sur ces grammaires prennent une décision par défaut si aucune règle ne s'applique. Dans notre cas, marquer cette indécision peut-être informatif pour le reste du système. En effet, il se peut que cette indécision provienne d'un autre agent non coopératif. Un retour lui permettrait peut-être de revenir sur une décision erronée.

Enfin, quelle que soit l'approche employée une évolution importante à faire au niveau syntaxique du futur système est de changer le grain des agents. Notre expérience a fait ressortir qu'il ne faut pas simplement disposer d'un agent par mot, mais d'un agent par couple mot/catégorie, voire même lemme/catégorie afin de traiter correctement les problèmes d'homonymie complètement éludés. On pourra se poser le même genre de questions à d'autres niveaux pour traiter de la polysémie.

Un sujet de thèse dans le prolongement de ce travail serait complètement centré sur la construction d'ontologies à partir de textes en utilisant un système multi-agent. Le travail à réaliser se déroulerait alors en deux temps. En premier lieu, il faudrait adapter une grammaire de dépendances au paradigme multi-agent en utilisant le moins possible de ressources extérieures, mais sans aller jusqu'à fonctionner sans ressources. Puis, très vite il serait requis de se focaliser sur des travaux de plus haut niveau pour effectuer des regroupements de termes et des extractions de relations, mais aussi permettre au système de tenir compte de retours de la part d'un utilisateur cherchant à construire son ontologie.

Bibliographie

[Aloulou *et al.*, 2003] Chafik Aloulou, Lamia Hadrich Belguith, Ahmed Hadj Kacem, et Abdelmajid Ben Hamadou. Apport des approches agent pour l'analyse syntaxique de l'arabe. In *JFSMA*, pages 47–60, 2003.

[Assadi, 1998] Houssem Assadi. *Construction d'ontologies régionales à partir de textes techniques.* Thèse de Doctorat, Université Paris 6, 1998.

[Aussenac-Gilles et Busnel, 2002] Nathalie Aussenac-Gilles et Agnès Busnel. Une méthode de construction d'ontologies pour l'industrie de la fabrication de la fibre de verre. Rapport Technique IRIT/2002-11-R., IRIT, Avril 2002.

[Aussenac-Gilles et Condamines, 2003] Nathalie Aussenac-Gilles et Anne Condamines. Terminologies et corpus. Rapport Technique IRIT/2003-23-R., IRIT, Octobre 2003. Rapport final de l'Action Spécifique ASSTICCOT.

[Aussenac-Gilles et Maedche, 2002] Nathalie Aussenac-Gilles et Alexander Maedche. the machine learning and natural language processing for ontology engineering workshop. In *ECAI*, Lyon (F), 2002.

[Bachimont, 2001] Bruno Bachimont. Modélisation linguistique et modélisation logique des ontologies : l'apport de l'ontologie formelle. In *Conférence Ingénierie des Connaissances IC 2001*, pages 349–368, Grenoble (F), 2001.

[Bourigault et Aussenac-Gilles, 2003] Didier Bourigault et Nathalie Aussenac-Gilles. Construction d'ontologies à partir de textes. In *Actes de la 10ème conférence sur le Traitement Automatique des Langues Naturelles TALN2003*, pages 27–47, Batz-sur-mer (F), juin 2003.

[Bourigault, 1994] Didier Bourigault. *LEXTER, un Logiciel d'Extraction de TERminologie. Application à l'acquisition des connaissances à partir de textes.* Thèse de Doctorat, Ecole des Hautes Etudes en Sciences Sociales, 1994.

[Camps *et al.*, 1998] Valérie Camps, Marie-Pierre Gleizes, et Pierre Glize. Une théorie des phénomènes globaux fondée sur des interactions locales. In *Actes des Sixième Journées Francophones Intelligence Artificielle Distribuée et Systèmes Multi-Agents.* Hermès, 1998.

[Charlet, 2003] Jean Charlet. *Développements, résultats et perspectives pour la gestion des connaissances médicales.* Université Pierre et Marie Curie, février 2003. Habilitation à diriger des recherches.

[Daille, 1994] Béatrice Daille. *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques.* Thèse de Doctorat, Université Paris 7, 1994.

[Déjean, 1998] Hervé Déjean. *Concepts et algorithmes pour la découverte des structures formelles des langues.* Thèse de Doctorat, Université de Caen, 1998.

[Enguehard et Pantera, 1995] Chantal Enguehard et Laurent Pantera. Automatic natural acquisition of a terminology. *Journal of Quantitative Linguistics*, pages 27–32, 1995.

[Erman et Lesser, 1975] Lee D. Erman et Victor R. Lesser. A multi-level organization for problem solving using many, diverse, cooperating sources of knowledge,. In *IJCAI 1975*, volume 2, pages 483–490, 1975.

[Gandon, 2002] Fabien Gandon. *Intelligence artificielle distribuée et gestion des connaissances : ontolgies et systèmes multi-agents pour un web sémantique organisationnel*. Thèse de Doctorat, INRIA et Université de Nice - Sophia Antipolis, 2002.

[Gangemi *et al.*, 2002] Aldo Gangemi, Nicola Guarino, Claudio Masolo, Alessandro Oltramari, et Luc Schneider. Sweetening ontologies with dolce. In *EKAW, Knowledge Engineering and Knowledge Management*. Springer, octobre 2002.

[Georgé *et al.*, 2003] Jean-Pierre Georgé, Marie-Pierre Gleizes, et Pierre Glize. Conception de systèmes adaptatifs à fonctionnalité émergente : la théorie des amas. *Revue d'Intelligence Artificielle*, 2003.

[Giguet, 1998] Emmanuel Giguet. *Méthode pour l'analyse automatique de structures formelles sur documents multilingues*. Thèse de Doctorat, Université de Caen, 1998.

[Harris, 1968] Zellig S. Harris. *Mathematical Structures of Language*. John Wiley and Sons, New-York, 1968.

[Houben, 2002] Frédérick Houben. Production semi-automatique de ressources à partir de corpus brut pour un analyseur syntaxique multilingue minimal. Mémoire de DEA, Université de Caen, 2002.

[Kahane, 2000] Sylvain Kahane. Les grammaires de dépendance. In *Traitement Automatique des Langues*, volume 41. Hermes, 2000.

[Lebarbé, 2002] Thomas Lebarbé. *Hiérarchie Inclusive des Unités Linguistiques en Analyse Syntaxique Coopérative*. Thèse de Doctorat, Université de Caen, 2002.

[Lesser et Erman, 1977] Victor R. Lesser et Lee D. Erman. A retrospective view of the hearsay ii architecture. In *IJCAI 1977*, pages 790–800, 1977.

[McGrath *et al.*, 2003] Robert E. McGrath, Anand Ranganathan, Roy H. Campbell, et Denis Mickunas. Use of ontologies in pervasive computing environments. Rapport Technique UIUCDCS-R-2003-2332 UILU-ENG-2003-1719, University of Illinois, Department of Computer Science, Avril 2003.

[Morin, 1999] Emmanuel Morin. *Acquisition de patrons lexico-syntaxiques caractéristiques d'une relation sémantique*. Hermès, Paris, 1999.

[Ottens *et al.*, 2004] Kévin Ottens, Nathalie Aussenac-Gilles, Marie-Pierre Gleizes, et Pierre Glize. Systèmes multi-agents pour l'extraction d'ontologies à partir de textes : revue de questions. *Journée ATALA AGENTAL : Agents & Langues*, mars 2004.

[Poosala, 1997] Viswanath Poosala. Zipf's law, 1997.

[Sabah, 1989] Gérard Sabah. *L'intelligence artificielle et le langage*, volume 2. Processus de compréhension, chapitre 2. Les principes de l'analyse syntaxique des phrases, pages 37–71. Hermès, Paris, 1989.

[Sauvage-Wintergerst, 1992] Caroline Sauvage-Wintergerst. *Parallélisme et traitement automatique des langues : application à l'analyse des énoncés elliptiques*. Thèse de Doctorat, Université Paris Paris-Sud Centre d'Orsay, mars 1992.

[Schmid, 1994] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. 1994. IMS-CL, Université de Stuttgart.

[Stefanini, 1993] Marie-Hélène Stefanini. Vers une architecture multi-agents pour le traitement automatique des langues naturelles. In *Actes des Premières Journées Francophones Intelligence Artificielle Distribuée et Systèmes Multi-Agents*, pages 153–165, 1993.

[Szulman *et al.*, 2002] Sylvie Szulman, Brigitte Biébow, et Nathalie Aussenac-Gilles. Structuration de terminologies à l'aide d'outils d'analyse de textes avec terminae. In *TAL*, volume 43, pages 103–128. Hermès, Paris, 2002.

[Séguéla, 2001] Patrick Séguéla. *Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques.* Thèse de Doctorat, Université Paul-Sabatier - Toulouse III, mars 2001.

[Tesnière, 1959] Lucien Tesnière. *Eléments de syntaxe structurale.* Klincksieck, Paris, 1959.

[Vergne, 1999] Jacques Vergne. *Étude et modélisation de la syntaxe des langues à l'aide de l'ordinateur, analyse syntaxique automatique non combinatoire*, chapitre 6. Synthèse des concepts principaux. Université de Caen, 1999. Habilitation à diriger des recherches.

[Vergne, 2003] Jacques Vergne. Un outil d'extraction terminologique endogène et multilingue. *TALN 2003*, 2003.

[Warren, 1998] Karine Warren. *Gestion de conflits dans une architecture multi-agents d'analyse automatique de textes.* Thèse de Doctorat, Université Stendhal - Grenoble III, 1998.