Des mots aux textes. Analyse sémantique pour l'accès à l'information

Thierry Poibeau

Département d'informatique **École doctorale « Galilée »**
Laboratoire d'Informatique de Paris-Nord

Des mots au texte

Analyse sémantique pour l'accès à l'information

MÉMOIRE

présenté le 26 novembre 2008

pour l'obtention de

l'Habilitation à diriger des Recherches
(spécialité informatique)

par

Thierry Poibeau

Composition du jury

Michel Denis, DR, LIMSI-CNRS
Catherine Fuchs, DR, LaTTiCe-CNRS (*rapporteur*)
Benoît Habert, PU, ENS-LSH (*rapporteur*)
Daniel Kayser, PU, Univ. Paris 13
Adeline Nazarenko, PU, Univ. Paris 13
Maria Teresa Pazienza, PU, Univ. Rome Tor Vergata (*rapporteur*)
Pierre Zweigenbaum, DR, LIMSI-CNRS

Remerciements

Je souhaite remercier l'ensemble des membres du jury pour l'intérêt qu'ils ont bien voulu manifester pour mes recherches. Je suis très reconnaissant à Catherine Fuchs, Benoît Habert et Maria Teresa Pazienza d'avoir accepté d'être rapporteurs. Leurs remarques ainsi que celles de Michel Denis et de Pierre Zweigenbaum sont particulièrement précieuses pour la poursuite de ma réflexion. Je remercie aussi Adeline Nazarenko et Daniel Kayser pour leur participation au jury et leurs conseils lors de la préparation de cette Habilitation.

Je tiens par ailleurs à remercier Christophe Fouqueré, Françoise Gayral, Marie-Paule Jacques, François Lévy, Aliyah Morgenstern et Yorick Wilks pour leur relecture attentive et amicale de versions préliminaires de ce manuscrit. Leurs remarques m'ont été précieuses pour l'améliorer sur bien des points. D'autres personnes ont également nourri ma réflexion sur des aspects particuliers, notamment Anne Condamines, Didier Bourigault, Jacqueline Léon, Patrick Saint-Dizier, Isabelle Tellier et Bernard Victorri.

Ce mémoire porte principalement sur les recherches que j'ai pu mener depuis que j'ai rejoint de Laboratoire d'Informatique de Paris-Nord. Je remercie particulièrement Christophe Fouqueré et Adeline Nazarenko pour la grande liberté qu'ils m'ont donnée très tôt au sein du laboratoire et plus particulièrement au sein de l'équipe RCLN (Représentation des Connaissances et Langage Naturel). Je remercie évidemment l'ensemble des membres de l'équipe pour l'environnement à la fois amical et stimulant dont j'ai pu bénéficier toutes ces années. Ces remerciements s'adressent aux membres du LIPN de manière plus générale : c'est un lieu à la fois studieux et chaleureux où il est agréable de travailler. Je remercie notamment Dominique Bouthinon pour ses cafés, et l'ensemble du personnel administratif pour sa sollicitude et son efficacité.

D'autres lieux ont été tout aussi précieux pour l'évolution de ma recherche. Je pense au Centre de Recherche en Ingénierie Multilingue de l'Institut National des Langues et Civilisations Orientales dirigé par Monique Slodzian, où j'enseigne depuis de nombreuses années. Je pense aussi au laboratoire de Célestin Sedogbo à Thales Recherche et Technologie : c'est en particulier grâce à Célestin que j'ai pu faire ma thèse dans un cadre industriel très fécond pour la recherche. Je ne puis énumérer ici tous les collègues que j'ai côtoyés lors de projets, de conférences ou d'ateliers, de travaux d'édition ou de publication, mais je sais que chacune de ces actions à contribué à ma réflexion et à ma recherche.

J'ai pu très tôt encadrer des stages de Masters et, plus récemment, des thèses de doctorat en collaboration avec Adeline Nazarenko ou Daniel Kayser. Mes recherches ont grandement bénéficié des réflexions des doctorants et post-doctorants avec qui j'ai étroi-

tement travaillé ces dernières années, notamment Aurélien Bossard, Amanda Bouffier et Cédric Messiant. Ce mémoire en témoigne largement. J'ai aussi une pensée pour Mani Ezzat, Michel Généreux, Marie-Paule Jacques et Maria Zimina, avec qui j'ai eu l'occasion de travailler sur différents projets.

J'ai principalement rédigé ce mémoire lors de séjours de recherche à Cambridge. J'y ai bénéficié d'une atmosphère à la fois chaleureuse et stimulante, sans compter les chocolats et les crackers qu'Anna m'a fournis sans relâche... Ma réflexion lui doit beaucoup. Kiitos kaikesta !

Résumé

Pourquoi est-il si difficile de comprendre une langue de manière automatique, même si on ne vise qu'une compréhension limitée, factuelle et orientée vers des faits connus ? La langue, telle qu'elle s'offre à nous, semble trop malléable pour être directement appréhendable par ordinateur. C'est pourtant à ce problème que je me suis intéressé : comment identifier du semblable dans des productions langagières si variées, comment repérer des fragments de signification au milieu d'un océan de textes ?

Ce mémoire se compose de quatre chapitres. Je reviens tout d'abord sur certains développements récents de la linguistique informatique, pour montrer que la disponibilité de gros corpus a entraîné une forte opérationnalisation du domaine. Cette évolution n'est pas neutre théoriquement : je pense que l'apport des corpus et des techniques d'acquisition dynamique de connaissances (notamment par les techniques d'apprentissage) rend tout à fait plausible l'idée d'une sémantique fondée sur l'usage.

Les trois chapitres suivants portent chacun sur un niveau d'analyse différent (niveau lexical pour l'annotation sémantique, niveau prédicatif pour l'extraction de relations, niveau textuel pour la modélisation de documents spécialisés). Je défends l'idée d'un continuum entre ces niveaux, du fait notamment que tous partagent des similarités fondamentales, ce qui peut se manifester parfois de manière très visible et influer sur les techniques utilisées.

Je reviens, dans la conclusion, sur les similitudes observées entre ces différents paliers : la question de la relation entre mots et concepts, les bords flous des catégories envisagées, leur grande variabilité sur le plan linguistique. Je m'interroge sur le lien entre traitement automatique des langues (TAL) et linguistique, avant de proposer quelques perspectives permettant de poursuivre ce travail par d'autres chemins.

Abstract

Why is it so difficult to automatically understand a language — even when what is targeted is only a limited kind of understanding, based on known facts? A key reason is the great variability in language which is too challenging for a computer. This is the problem I try to tackle: how to identify similar meanings among different expressions? How to identify fragments of meaning in a sea of texts?

This thesis consists of four chapters. I first consider recent developments in computational linguistics: I show that the availability of large corpora has resulted in more functional Natural Language Processing (NLP). This evolution carries the potential of a major impact on theory: corpora and automatic acquisition of knowledge from corpora (especially using machine learning techniques) makes it possible to get semantics based on language use.

Each of the next three chapters deals with a different level of analysis (lexical semantics for semantic annotation, predicative semantics for relation extraction, and text semantics for technical document modelling). I suggest the idea of a continuum between these levels, since they all share fundamental similarities that affect the techniques used.

I emphasize, in the conclusion, the similarities between these three different levels: the complex problem of the relations between words and concepts, the fuzziness of linguistic categories, the great variability of language. I conclude with a discussion on the relationship between NLP and linguistics, before proposing future research through alternative routes.

Table des matières

In some ways, [Artificial Intelligence] is akin to medieval alchemy. We are at the stage of pouring together different combinations of substances and seeing what happens, not yet having developed satisfactory theories.

T. Winograd "On some Contested Suppositions of Generative Linguistics about the Scientific Study of Languages". *Cognition*, vol. 5. 1977.

Introduction

Pourquoi est-il si difficile de comprendre une langue de manière automatique, même si on ne vise qu'une compréhension limitée, factuelle et orientée vers des faits connus ? La langue, telle qu'elle s'offre à nous, semble trop malléable pour être directement appréhendable par ordinateur. C'est pourtant à ce problème que je me suis intéressé : comment identifier du semblable dans des productions langagières si variées, comment repérer des fragments de signification au milieu d'un océan de textes ?

Ce mémoire décrit mes activités de recherche, comme il est d'usage pour une Habilitation à Diriger des Recherches, mais j'espère faire aussi ressortir un peu de mon intérêt pour les langues et leurs « mystères insondables ». Pour reprendre les mots de P. Boldini [2006], « dans ce genre d'exercice de nature épistémologique, il me paraît important d'indiquer d'où l'on parle, c'est-à-dire à partir de quelle position dans les champs scientifique et académique et à partir de quelle pratique ». Qu'il me soit donc permis de commencer par quelques touches personnelles.

1 Quelques éléments d'un parcours

Le domaine du Traitement Automatique des Langues (TAL) semble avoir toujours été tiraillé entre linguistique et informatique. Mon poste actuel est à cet égard révélateur : Chargé de Recherche au CNRS, je suis rattaché à la section linguistique (ou plus exactement à la section 34 « Langues, langage, discours ») mais, au sein de l'Université Paris 13, je travaille au Laboratoire d'Informatique de Paris-Nord[1], lui-même rattaché au département « Sciences et Technologies de l'Information et de l'Ingénierie » (ST2I).

D'une manière plus générale, le développement de données textuelles disponibles sur support informatique semble ces dernières années avoir fait pencher la balance au profit de l'informatique. Le contenu des conférences de traitement des langues est un bon indicateur de cet état de fait, les linguistes y étant aujourd'hui largement absents. C'est pourtant par les langues que je suis venu à ce domaine.

Ce qui est fascinant dans les langues, c'est leur incroyable complexité, leur diversité mais aussi leur ressemblance. L'étude des langues permet d'en entrevoir l'« architecture » : leur grammaire, leur lexique, mais aussi leur histoire et leur comparaison. On retrouve un

[1] Au sein de l'équipe Représentation des Connaissances et Langage Naturel (RCLN), dirigée par Adeline Nazarenko.

peu de cet univers dans le TAL. Après mon DEA de linguistique (à l'Université Paris 7), j'ai eu la chance d'être recruté au centre de recherche de Thales (Thales Recherche et Technologie). J'y ai travaillé de 1998 à 2003, d'abord en thèse, puis comme ingénieur, sur des questions d'extraction d'information. J'ai ensuite été recruté par le CNRS sur le poste que j'occupe actuellement.

La nature de la recherche dans une entreprise privée comme Thales peut sembler très éloignée des questions fondamentales sur la langue. C'est sans doute en partie vrai : durant les années que j'ai passées à Thales, à travers les projets et les collaborations, ce sont surtout les aspects applicatifs du domaine que j'ai pu explorer. Il s'agissait de développer des « démonstrateurs », c'est-à-dire des maquettes si possible fonctionnelles et opérationnelles. Ces aspects pratiques ne sont à mon avis ni à négliger, ni à dédaigner : la langue est un outil pratique, qui sert essentiellement à communiquer, et il serait inopportun d'en faire une abstraction philosophique. Les applications obligent à formaliser, à décrire explicitement les ressources nécessaires, mais aussi à « faire avec ce que l'on a ». Autrement dit, les applications sont aussi intéressantes en ce qu'elles montrent les limites de notre connaissance sur la langue et la complexité de celle-ci, dans la mesure où on est encore loin de savoir réaliser des systèmes de compréhension automatique[2].

Ce mémoire tente de faire une synthèse de mes travaux et de mes réflexions, à partir de ce rapide arrière-plan personnel. Les quatre années que j'ai passées au CNRS m'ont permis de poursuivre mes recherches dans un cadre moins directement appliqué que celui offert par une entreprise privée. Ce cadre ne change pourtant pas fondamentalement les questions que je continue de me poser sur la langue.

2 La compréhension automatique de textes

Les applications sur lesquelles je me suis penché ont trait à la compréhension automatique de textes. Il faut toutefois souligner les limites de ce que l'on entend par là : la compréhension correspond en fait à la reconnaissance, au typage et à la mise en relation de certains éléments pertinents par rapport à une tâche. Ce type d'applications est donc très limité et nécessite une étude de besoins préalable. Il ne s'agit pas de mettre au point des systèmes de compréhension généraux et valables hors domaine, même si certaines applications peuvent parfois en donner l'impression (comme les systèmes de questions-réponses).

Précisons un peu les différents types d'applications abordés :

- **l'annotation sémantique** permet de mettre en évidence, à même le texte (en général par un jeu de couleurs approprié), des éléments d'information pertinents (mots clés, noms propres, *etc.*).
- **l'extraction d'information** vise à extraire des informations structurées pour remplir une base de données (concernant par exemple des rachats d'entreprises, des réseaux d'interactions géniques, *etc.*).

[2]Pour une autre expérience en milieu industriel, voir [Bourigault, 2007].

- **les systèmes de questions-réponses** visent à répondre de manière précise à des questions posées en langage naturel, en cherchant des réponses au sein d'un ensemble de textes.
- **le résumé automatique de textes** vise à produire un texte cible reprenant de façon concise les principales informations contenues dans un texte source. La cible est généralement une simple sélection de phrases pertinentes du texte source.

Le domaine de la compréhension de textes a fortement évolué depuis les débuts de la recherche en TAL, il y a plus de 50 ans. Paradoxalement, cette évolution est en partie fondée sur une série de revers, révélant progressivement la complexité de la tâche à accomplir.

La compréhension de textes est une tâche mentionnée dès les débuts de l'informatique : le test de Turing (1950), qui vise à simuler un dialogue humain à travers une interaction avec la machine, implique une compréhension minimale de la parole humaine[3]. Par la suite, la compréhension est devenue un passage obligé après l'échec des premières tentatives de traduction automatique dans les années 1950–1960 [Dreyfus, 1992; Hutchins, 2000; Léon, 2002b][4]. Les systèmes génériques, visant à comprendre des textes tout venant, ont aussi rapidement fait faillite devant l'ampleur des connaissances à modéliser. Les recherches se sont alors focalisées sur des domaines de plus en plus limités. Enfin, les années 1980-1990 ont vu un fort déclin des systèmes de compréhension de textes proprement dits (au sens où il s'agit de donner une représentation d'un texte de manière globale) au profit des systèmes visant une compréhension très partielle, mettant en évidence quelques éléments essentiels et les relations qu'ils entretiennent entre eux.

Même dans le cas d'une compréhension limitée à quelques éléments clés, l'infinie variabilité de la langue rend la tâche si difficile que les taux d'erreur restent importants. Par exemple, le meilleur système de questions-réponses à TREC-QA 2006 fournit la bonne réponse (en première position dans la liste de réponses possibles) dans moins de 50 % des cas. Cela ne veut pas dire que les applications soient complètement inenvisageables et encore moins qu'il faille arrêter la recherche dans ces domaines : sur des tâches bien ciblées, les performances s'améliorent régulièrement. La généralisation relativement récente du recours à l'apprentissage permet une meilleure adaptabilité des systèmes. Enfin, la prise en compte de contraintes ergonomiques rend plus supportable les manques et les erreurs : par exemple, l'annotation sémantique permet une prise de connaissance rapide d'informations pertinentes, même si tout ce qui est souligné n'est pas pertinent. L'apparition d'applications nouvelles, aussi bien pour le grand public que pour les professionnels (ceux-ci ayant souvent besoin d'informations plus précises et plus complexes), est un signe encourageant pour le domaine.

[3]En fait, les systèmes de dialogue comme Eliza [Weizenbaum, 1966] montreront rapidement qu'une stratégie à base de mots clés et de patrons de réponse peut être efficace, sans inclure de mécanisme de compréhension à proprement parler.

[4]L'arrêt brutal des recherches en traduction automatique après la publication du rapport ALPAC de 1966 ne pas doit masquer le foisonnement d'idées et d'approches au cours de la période. Certaines de ces idées seront d'ailleurs reprises avec succès par la suite (voir chapitre 1, section 1.2.2).

Il n'empêche qu'au-delà de ces aspects applicatifs, il semble important de continuer à s'interroger sur les raisons de résultats limités en dépit de plus de cinquante ans de recherche.

3 La quête du sens

La compréhension de textes vise à rendre compte du sens des textes. Or, que veut dire comprendre ? Peut-on déterminer le sens d'un texte ? Y a-t-il un sens à parler de sens ?

Une première difficulté vient de l'imprécision des termes employés et des résultats visés. On peut rappeler ici la polémique autour de la notion de sens, à la fin des années 1980 entre D. Kayser d'une part, G. Kleiber et M. Riegel d'autre part. Dans un article initial, Kayser [1987] critique la notion de sens, il n'y voit qu'une « abstraction » qui « n'est pas fondée scientifiquement ». Kleiber et Riegel [1989] constatent, quant à eux, que « la tendance actuelle, (...) serait plutôt à l'interprétation des phrases selon une géométrie variable qui articule leur sens littéral (compositionnel) avec des données contextuelles et illocutionnaires (ce qui est dit littéralement, qui le dit, à qui, comment, quand, où, pourquoi, *etc.*), sans oublier les connaissances générales et particulières que les interlocuteurs se prêtent réciproquement. » Cette description me semble correspondre assez fidèlement au paysage actuel en sémantique (de ce point de vue, les choses n'ont pas trop changé en 20 ans).

Dans ce mémoire, nous faisons nôtres les interrogations de D. Kayser. En effet, « il est connu, que sauf entraînement spécial, un interlocuteur ne sait pas donner une représentation du sens d'un énoncé ». En revanche, ce qui est à la portée d'un interlocuteur, c'est fournir une paraphrase ou faire des inférences à partir d'un énoncé[5]. Comme le dit Ricœur [2006], comprendre, c'est « dire la même chose autrement » ou, d'une manière plus générale, « comprendre, c'est traduire » [Steiner, 1998].

Tout semble par ailleurs démontrer que la compréhension est inséparable du contexte (notons d'ailleurs que les positions ne s'opposent pas sur ce point). Comme il a été dit, comprendre implique de représenter « les données contextuelles et illocutionnaires (...), sans oublier les connaissances générales et particulières que les interlocuteurs se prêtent réciproquement ». On admettra aisément que cela est difficilement envisageable, au moins à large échelle (c'est pourquoi des auteurs comme Sampson [2001] rejettent la sémantique à l'extérieur des sciences, par opposition aux autres branches de la linguistique). Même pour l'analyse de textes écrits (où le contexte de production n'a pas l'importance qu'a le contexte illocutoire pour la compréhension du dialogue), la modélisation de connaissances générales nécessaires à la compréhension reste un travail de titan qui n'a été effectué qu'à une échelle très réduite, pour des applications très ciblées. Une première simplification s'impose donc.

[5] L'existence d'un sens « littéral », accessible sans que l'on ait connaissance du contexte, a elle aussi été fort contestée ([Kayser, 1989; Kleiber et Riegel, 1991]). Voir [Lyons, 1995] pour un point de vue d'ensemble sur la question, [Recanati, 2007] ou [Rastier, 2001] pour deux points de vue engagés.

Les applications informatiques ici décrites reposent sur des modèles très simplifiés du monde. Que ce soit pour des tâches d'annotation, d'extraction ou de questions-réponses, on identifie des séquences textuelles renvoyant à des objets du monde, puis des relations entre ces objets. Le modèle ainsi obtenu est souvent représenté sous forme d'*ontologie*, où les objets du monde sont appelés *concepts* et sont liés par des *relations*. Le modèle esquisse un lien référentiel avec le monde réel tandis que l'application vise à « instancier » ce modèle avec des connaissances repérées au sein des textes.

4 Du corpus au modèle

Les applications reposent, on l'a vu, en partie sur des modèles formels permettant de fournir une représentation simplifiée du monde. Cette représentation est nécessaire afin de pouvoir nommer des séquences textuelles (par exemple des entités) et les relations identifiées entre ces séquences au sein du corpus. Le modèle permet donc de ramener un ensemble de formes à un type (ou à un champ dans une base de données) et d'affranchir en partie l'analyse du problème de la référence ; il permet de figer un cadre de manière pré-déterminée.

D'un autre côté, cette analyse exige des ressources et des connaissances qui sont souvent spécifiques à la tâche et ne peuvent provenir complètement de sources de connaissance extérieures. Il est donc nécessaire de définir des méthodes d'acquisition de connaissances, d'exploration et de « moissonnage » des textes. On sait qu'un corpus, assemblé selon des critères rigoureux, peut former un tout cohérent et, par suite, renfermer des informations précieuses pour l'interprétation [Sinclair, 2007]. Ces régularités (au niveau du mot, du syntagme ou de « périodes de texte » plus larges) peuvent souvent être repérées par des moyens automatiques — par apprentissage notamment — et contribuent à la compréhension [Habert *et al.*, 1997; Nazarenko, 2004].

On a alors affaire à deux pôles, à la fois complémentaires et contradictoires. D'un côté le cadre réducteur de l'application qui fige le sens de manière artificielle et *a priori*. De l'autre, l'acquisition, qui permet d'identifier de manière dynamique régularités et nuances de sens, mais qui ne garantit pas que ces régularités correspondent exactement au modèle visé. Dès lors, il semble intéressant de voir les conséquences de ces deux points de vue. Jusqu'où sont-ils cohérents ? contradictoires ? En quoi les limites des modèles développés nous dévoilent-elles la nature et la complexité de la langue ? Que nous révèlent les applications quant à l'inadéquation de l'approche ontologisante (c'est-à-dire à base de modèles figeant le sens *a priori*) ? Ce sont quelques unes des questions auxquelles ce mémoire essaie de répondre.

Chemin faisant, on verra en quoi la démarche se distingue des théories qui tiennent le texte pour un objet autonome, susceptible d'être analysé par lui-même, indépendamment de toute connaissance extérieure. Je pense que le lecteur verra aussi en quoi la démarche n'adhère pas au courant dit de « sémantique formelle », bien que l'utilisation de modèles

semi-formels[6] puisse y faire penser. Nous aurons par exemple l'occasion de « croiser » la notion de sens littéral, de tenter de la cerner et de la critiquer : la participation à des campagnes d'évaluation a, entre autres, cet avantage qu'on est amené à se colleter avec d'autres cadres théoriques, ce qui permet de les examiner « de l'intérieur ».

La nature des questions soulevées amène aussi à se pencher sur des théories et des réflexions au-delà du domaine du TAL. C'est particulièrement le cas avec Wittgenstein, qui tient une place privilégiée dans ce mémoire car sa philosophie me semble offrir des pistes de réflexion particulièrement riches et éclairantes par rapport aux questions posées. J'ai essayé de ne pas faire trop de rapprochements indus. Le constat — souvent fait après coup, parfois au hasard de mes lectures — que je n'étais pas le seul à avoir abordé cette voie (cf. [Wilks, 2005]) m'a conforté dans l'idée qu'elle n'était peut-être pas complètement erronée.

5 Contenu et plan du mémoire

Ce mémoire vise à rendre compte de mes activités de recherche en traitement des langues depuis une dizaine d'années (j'insisterai surtout sur les travaux effectués depuis que je suis entré au CNRS). J'essaie aussi, autant qu'il est possible, de prendre du recul par rapport à des recherches souvent appliquées. Il ne s'agit pas d'opposer la théorie à la pratique, mais plutôt d'observer comment une pratique révèle ses propres limites et laisse voir, en filigrane, une partie du fonctionnement de la langue.

Je détaille des réalisations qui visent à répondre à des besoins concrets : le plus souvent, il s'agit de démonstrateurs développés en collaborations au sein de projets de recherche. De ce point de vue, ce mémoire se rattache au domaine de l'ingénierie linguistique. Il est donc en partie composé de descriptions de systèmes, évalués en fonction de critères précis.

Au-delà, il me semble intéressant de prendre du recul par rapport à ces applications pour les examiner avec un autre regard. L'hypothèse que je formule pour cette investigation est la suivante : on a ici affaire à des réalisations visant à répondre à des besoins avérés et reposant sur des modèles (semi-)formels. Or, ces modèles permettent difficilement de « passer à l'échelle », au-delà des applications visées : il semble dès lors légitime de s'interroger sur leur adéquation au problème. Cette question me semble valoir pour l'ensemble du domaine de l'ingénierie linguistique, tant celui-ci peine à proposer des outils vraiment robustes face à la diversité des corpus[7].

C'est cette double démarche scientifique que j'essaie de décrire ici : d'une part, la production d'outils et d'autre part, une réflexion épistémologique sur la valeur de ces outils. J'étudie ce deuxième point à partir de réflexions théoriques sur la langue que je

[6] On parle ici de modèles semi-formels dans la mesure où l'on dispose d'outils de représentation des connaissances rigoureux. En revanche, on ne cherche pas à valider automatiquement ces connaissances.

[7] L'apprentissage semble donner un début de réponse à ce problème, en permettant d'adapter les outils par un ré-entraînement sur un corpus typique, mais les corpus annotés sont rares et coûteux à constituer. On est par ailleurs encore loin de disposer de systèmes réflexifs pouvant adapter dynamiquement la stratégie d'analyse en fonction du corpus à analyser.

me donne *a priori*, sans que ces réflexions soient directement guidées par ma pratique du traitement des langues. Malgré ce décalage — ou plutôt à cause de lui —, il me semble intéressant de continuer dans cette voie, pour au moins deux raisons. D'abord, cette démarche peut être poursuivie parallèlement à la recherche de solutions pratiques, répondant aux besoins concrets exprimés par des utilisateurs, qui ne peuvent attendre le résultat de recherches théoriques aux débouchés aléatoires. Ensuite, les réflexions sur la langue (autant les miennes que celles d'autres chercheurs) ne permettent pas à l'heure actuelle de dégager une théorie ou une approche formelle précise qui guideraient de manière directe l'ingénierie du domaine, comme on le verra dans le premier chapitre de ce mémoire. A l'inverse, une réflexion sur les résultats pratiques obtenus peuvent valablement alimenter des réflexions sur la langue[8].

J'ai conscience qu'il s'agit là d'une démarche scientifique peu classique et sans doute critiquable dans la mesure où elle met en évidence deux paradigmes qui se font écho sans se répondre toujours. Dans le mémoire, j'ai essayé de préciser, par des renvois réguliers, les points de contact entre théorie et pratique, principes et applications. Ces points de contact sont bien sûr à affiner et à affermir ; cette recherche constitue la perspective de mes travaux.

Ce document n'est pas le résultat d'un travail solitaire : il s'agit au contraire d'un travail d'équipe mené d'abord avec mes collègues de Thales puis du Laboratoire d'Informatique de Paris-Nord, sans oublier ceux de l'INaLCO[9]. Il faudrait également mentionner les collaborations nationales et internationales dont j'ai pu bénéficier, ayant eu la chance de travailler au sein d'organisations fortement impliquées dans des réseaux de recherche actifs. Enfin, j'ai depuis quelques années maintenant la chance de diriger à mon tour des stages de Master et des thèses. Ces rencontres ont été autant d'opportunités d'enrichir mon parcours de recherche en bénéficiant de compétences et de points de vue variés, suscitant toujours la réflexion.

Ce mémoire se compose de quatre chapitres. Je reviens tout d'abord, dans le **chapitre 1**, sur certains développements récents de la linguistique informatique, pour montrer que la disponibilité de gros corpus a entraîné une forte opérationnalisation du domaine. Cette évolution n'est pas neutre théoriquement : je pense que l'apport des corpus et des techniques d'acquisition dynamique de connaissances (notamment par les techniques d'apprentissage) rend tout à fait plausible l'idée d'une sémantique fondée sur l'usage.

[8] Dans la bouche de Culioli : « *Quelquefois, on se place dans une direction, qui a un certain nombre d'intérêts, et on peut avoir la tentation de se dire : maintenant j'ai trouvé un filon, je continue. Mais on peut aussi avoir le désir d'aller regarder ailleurs. C'est comme la clé de Barbe Bleue : qu'est-ce qu'il y a derrière, est-ce qu'il y a autre chose ?* » [Culioli et Normand, 2005]

[9] Je suis associé au Département TIM (Textes, Informatique, Multilinguisme) de l'Institut National des Langues et Civilisations Orientales depuis 2002.

Les trois chapitres suivants portent chacun sur un niveau d'analyse différent. Je défends l'idée d'un *continuum* entre ces niveaux, du fait notamment que tous partagent des similarités fondamentales (Rastier *et al.* [1994] parlent de micro-, méso- et macrosémantique). On reconnaît notamment l'influence du global sur le local, ce qui peut se manifester parfois de manière très visible et influer sur les techniques utilisées (notamment quand la valeur d'une séquence se détermine à partir de caractéristiques calculées sur tout le corpus).

Le **chapitre 2** traite du niveau lexical (*microsémantique*), essentiellement à travers l'analyse des « entités nommées » : ce type de séquences comprend notamment les noms propres, qui sont des éléments essentiels pour une prise de connaissance rapide du contenu des documents. Ces séquences, et plus particulièrement les noms propres, ont été largement étudiées dans le cadre des approches logiques ; nous travaillons dans un cadre en partie hérité de cette tradition : nous verrons que celui-ci offre bien des avantages applicatifs mais que les noms propres, comme le vocabulaire courant, sont soumis aux mêmes phénomènes de variation et de brouillage de sens, du fait des tropes notamment.

Le **chapitre 3** traite essentiellement des relations prédicatives (*mésosémantique*) : le repérage de ces séquences est essentiel pour les systèmes d'extraction d'information et de questions-réponses. Ces applications reposent en effet sur la mise en correspondance d'entités autour d'un prédicat. Nous détaillons différentes techniques permettant d'acquérir automatiquement ces structures à partir de corpus (classes sémantiques, cadres de sous-catégorisation et restrictions de sélection). Nous montrons que ces catégories sont floues et que les analyses à partir de corpus remettent en cause certaines classifications de la grammaire traditionnelle.

Le **chapitre 4** traite du contenu et de la structure de textes complexes, essentiellement techniques (*macrosémantique*). Nous montrons que le texte forme un tout cohérent, marqué par un ensemble de séquences (ou périodes) liées entre elles ; cette architecture textuelle est normée et elle est significative pour la compréhension globale. Nous essayons d'étendre ce travail à des ensembles de textes cohérents, afin d'aboutir à une typologie. Nous montrons là aussi le flou et la difficulté à définir des typologies cohérentes et, surtout, fondées linguistiquement.

Nous revenons, dans **la conclusion**, sur les similitudes observés entre ces différents paliers : les difficultés d'une classification adéquate, les bords flous des catégories envisagées, leur grande variabilité sur le plan linguistique. Nous proposons enfin quelques perspectives permettant de poursuivre ce travail par d'autres chemins.

Chapitre 1

Une linguistique fondée sur l'usage

La disponibilité de corpus volumineux dans les années 1990 a en grande partie renouvelé le TAL. La nécessité d'accéder efficacement à une information pertinente requiert des analyses rapides et efficaces à l'échelle du document. Les techniques d'informatique documentaire ont donc connu un renouveau dans les années 1990, ainsi que le courant appelé « linguistique de corpus » [Cori et Léon, 2002]. Il s'agit en apparence d'une linguistique plus appliquée, moins théorique que les courants en vogue récemment encore [Abeillé, 1993] ; cette linguistique est aussi marquée par un souci d'opérationnalisation, dans la mesure où elle doit aussi faire face rapidement à des besoins précis. Encore faut-il se demander si ces applications restent en phase avec l'objet principal de la linguistique : donnent-elles à penser par rapport à la nature même de la langue ou du sens ?

1.1 Une linguistique sans théorie ?

Le renouveau de l'approche en traitement des langues est marqué par plusieurs faits caractéristiques :

- des visées applicatives assumées. L'arrivée conjointe des micro-ordinateurs, des réseaux informatiques puis du web a fait exploser les besoins en matière d'informatique documentaire. Alors que cette dernière restait confinée jusque là aux bibliothèques et aux grandes institutions, les années 1980 puis 1990 ont poussé au développement de techniques de recherche rapides, efficaces et distribuées.
- des avancées mesurables. Les années 1990 voient fleurir les campagnes d'évaluation récurrentes, où les systèmes sont mesurés par des métriques précises permettant de voir l'évolution des techniques en jeu. Des conférences comme TREC (Text Retrieval Evaluation Conference, 1990- [Voorhees et Buckland, 2006]), MUC (Message Understanding Conference, 1987-1998 [Grishman et Sundheim, 1996]), ACE (Automatic Content Extraction, 2000- [Doddington *et al.*, 2004]) et TREC-QA (Question Answering, 2000- [Voorhees, 2004]) ont permis de développer des systèmes de compréhension restreints, focalisés sur des éléments très précis au sein du texte.

– des approches traitant le texte en surface ; ces approches peuvent être symboliques (le plus souvent en ayant recours aux automates à nombre fini d'états [Roche et Schabes, 1997]), ou statistiques (permettant un accès rapide et relativement efficace au document à partir de mots clés ou de co-occurrences de mots clés [van Rijsbergen, 2004]).

La nécessité de traitements efficaces a relégué au second plan les formalismes linguistiques qui se voulaient, dans le même temps, des théories linguistiques (HPSG, LFG... [Abeillé, 1993]). Le lien entre traitement informatique et cognition est devenu très ténu quand il n'a pas purement et simplement disparu (voir [Sabah, 2006] pour un constat nuancé).

Le terme « ingénierie des langues » est parfois utilisé pour caractériser ce courant de recherche [Pierrel, 2000]. Celui-ci s'appuie largement sur des corpus de grande taille pour inférer des informations diverses. Les techniques se sont affinées en mélangeant plus ou moins heureusement statistique et linguistique : les calculs peuvent par exemple opérer sur un texte partiellement normalisé, suite à une lemmatisation, une analyse morpho-syntaxique ou sémantique partielle. Il s'agit en général d'échapper à la variation de surface pour parvenir à une représentation partiellement normalisée du texte [Nazarenko, 2004].

La mise en avant du corpus repose sur l'idée que la linguistique est une science empirique, qui a pour objet d'étude le langage à travers ses manifestations [McEnery et Wilson, 2001; Sampson, 2001; Sinclair, 2004]. Ce point de départ peut sembler évident, au point de pousser un chercheur comme Michael Hoey à dire de façon quelque peu péremptoire que « la linguistique de corpus n'est pas un sous-domaine de la linguistique mais la voie vers la linguistique » (propos rapportés par [Sampson, 2001, p. 6]). Il demeure bien une linguistique théorique, menant une réflexion à partir de quelques exemples obtenus par introspection, mais ce courant est à présent beaucoup moins représenté au sein du TAL[10].

Cet empirisme renouvelé n'est toutefois pas sans poser de multiples questions. Si la linguistique de corpus n'est pas une nouvelle branche de la linguistique, ce courant est-il homogène ou non ? Existe-t-il une théorie, ou un ensemble de théories assurant la cohérence de l'ensemble des expériences et des approches explorées au sein de la linguistique de corpus ? Il semble bien qu'il faille répondre par la négative à cette question, dans la mesure où la plupart des expériences se justifient par leur capacité à répondre à un problème, à obtenir des performances s'améliorant régulièrement, indépendamment de toute théorie sur la langue.

La même approche a donné des résultats valables dans des secteurs proches. Par exemple, la recherche en traitement de la parole repose en grande partie sur des techniques de traitement du signal qui ne disent rien (ou peu de choses) sur la langue, mais qui ont permis des progrès remarquables pour l'analyse de signal de plus en plus mauvaise qualité (l'analyse de conversations téléphoniques est aujourd'hui possible, alors que les premiers systèmes de transcription étaient naguère mono-locuteur et essentiellement

[10]Nous assimilons ici, sans doute un peu rapidement, linguistique de corpus et TAL. Il faudrait nuancer, dans la mesure où la linguistique de corpus désigne avant tout des études fondées sur corpus sans souci d'aboutir à des réalisations informatiques. A l'inverse, il existe de nombreuses approches en TAL fondées sur des corpus, sans que leurs auteurs ne se réclament de la linguistique de corpus [Léon, 2007a].

tournés vers la « parole dirigée »). Le TAL connaît donc aujourd'hui un état similaire, où l'avancée des techniques d'un côté et la pression des besoins de l'autre poussent au développement d'une « ingénierie des langues ». Ceci est une phase nécessaire et sans doute aussi la marque d'une certaine maturité du domaine, celui-ci étant suffisamment avancé pour donner jour à de réelles applications.

Nous pensons cependant qu'il est important de continuer à s'interroger sur les fondements théoriques du domaine, voire sur la notion de sens. Il est sans doute illusoire de chercher une théorie unifiante derrière la « linguistique de corpus » ; nous essaierons, de façon plus réaliste, de dessiner un cadre à nos recherches, permettant peut-être de mieux faire voir pourquoi nous avons tantôt choisi telle approche, tantôt telle autre.

1.2 Le sens, c'est l'usage !

Le renouveau de l'analyse sémantique, l'importance des corpus et des analyses contextuelles mettent au premier plan la notion d'usage. Cette notion amène assez logiquement à aller voir du côté de chez Wittgenstein : la philosophie de ce dernier éclaire d'un jour particulier la notion d'usage, de règle et de langage commun. Même si Wittgenstein ne développe pas une théorie linguistique à proprement parler, il a une conception de la langue qui nous semble très pertinente par rapport à la tâche que nous nous sommes fixée, à savoir faire une analyse partielle du sens des textes et produire, si possible, des outils permettant d'en rendre compte.

Dans cette partie, nous donnons un rapide résumé de quelques points essentiels de la pensée du philosophe. Nous examinons ensuite comment certains groupes de recherche en traitement automatique des langues dans le monde anglo-saxon, dès les années 1950, s'en font l'écho.

1.2.1 La grammaire du sens selon Wittgenstein

On trouve d'assez nombreuses références à Wittgenstein en linguistique. Celles-ci sont très focalisées et surtout, beaucoup plus rares dans la communauté du TAL [Wilks, 2005]. Les deux éléments les plus fréquemment repris sont la notion d'air de famille (« *family resemblance* ») et l'idée que c'est l'usage qui guide le sens. Il s'agit certes de points fondamentaux de la philosophie de Wittgenstein (que nous reprendrons à notre compte par la suite) mais d'autres l'ont dit avant lui, et peut-être de façon plus claire dans une perspective d'analyse linguistique[11]. Certains auteurs prennent cependant en compte plus fondamentalement la philosophie de Wittgenstein ; c'est notamment le cas à Cambridge (le CLRU — *Cambridge Language Research Unit* — étant longtemps dirigé par M. Masterman, une ancienne étudiante de Wittgenstein [Masterman, 2005]). Cela ne dit rien

[11] De la tradition philosophie indienne à travers l'école de Nyāya [Auroux, 1989] jusqu'à Saussure à travers la notion de parole [Bouquet, 2000].

de la pertinence de l'approche mais prouve qu'il peut valoir la peine de s'y intéresser, y compris dans une perspective de traitement automatique.

Avant d'aborder la pensée de Wittgenstein, il faut garder à l'esprit que celui-ci n'est pas un linguiste (et encore moins un linguiste informaticien !). Wittgenstein dit même clairement qu'il n'envisage le problème du langage qu'en tant qu'il éclaire des problèmes philosophiques (et aussi qu'il permet d'écarter un certain nombre de questions qui doivent être exclues du champ d'investigation de la philosophie, dans la mesure où un large pan de l'œuvre de Wittgenstein vise à montrer que certaines questions ne sont pas abordables par la langue, en tant que système de signes reflétant une pensée[12]).

Le point de départ bien connu de la pensée de Wittgenstein consiste à identifier le sens des mots à leur usage [1953, §66]. Wittgenstein affirme par là que la langue se donne à voir en elle-même ; elle ne cache pas le monde pur des Idées (contrairement à ce que décrit Platon) ni un langage universel (contrairement à une tradition bien établie allant au moins de Leibniz au... web sémantique dans son projet initial !)[13]. Les mots ont des sens complexes, variés, dépendant fortement du contexte. Ce qui peut apparaître comme un problème pour l'analyse automatique n'en est pas un du point de vue linguistique, bien au contraire : c'est cette variabilité qui fonde la richesse des langues. C'est aussi pour cela qu'une langue donnée[14] offre un système propre de représentation, dont aucun formalisme ni aucune logique ne peut complètement rendre compte (c'est un point sur lequel la pensée du philosophe ne varie guère d'un texte à l'autre).

> *En philosophie, nous comparons souvent l'usage des mots avec des jeux et des calculs dotés de règles stricts mais nous ne pouvons pas dire que quelqu'un qui utilise le langage doit jouer de cette sorte. (...) Tout au plus peut-on dire que nous construisons des langages idéaux. Mais le mot « idéal » peut être trompeur dans ce cas, car il semble dire que ces langages seraient meilleurs, plus parfaits que notre langage quotidien, que nous aurions besoin du logicien pour montrer*

[12] « Wittgenstein a exprimé sans aucune ambiguïté son peu de considération pour des analyses qui ne seraient pas motivées directement par la volonté de résoudre des problèmes philosophiques réels et profonds ». [Bouveresse, 2003]

[13] La notion de langue universelle varie bien évidemment selon les auteurs. Dans la tradition occidentale, les Classiques se réfèrent souvent à une langue adamique, c'est-à-dire à une protolangue hypothétique et universelle parlée par l'Humanité avant l'épisode de Babel. Leibniz est parmi les premiers à se dégager partiellement de cette tradition, au prétexte que cette langue adamique est impossible à dégager à partir des langues modernes. Il développe alors un projet important visant à redéfinir des concepts et des relations entre concepts non ambigus, afin de résoudre des problèmes de différentes natures, moraux, juridiques ou philosophiques [Leibniz, 2000]. Cette tradition perdurera en logique jusqu'au 19^{e} siècle, sous des tours divers, souvent moins ambitieux en apparence [Eco, 1994; Auroux, 1995]. On retrouve une partie de cette tradition dans le Wittgenstein du *Tractatus logico philosophicus* [1919]. C'est sur ces aspects que la réflexion du philosophe évoluera le plus profondément pour aboutir aux *Investigations philosophiques* [1953]. Ce sont les réflexions de ce dernier qui sont rapidement évoquées ici et qui ont inspiré une partie de cette étude.

[14] L'allemand ne marque pas l'opposition entre langue et langage, les deux mots correspondant à *Sprache*.

> *enfin aux hommes à quoi ressemble une phrase correcte.* [Wittgenstein, 1953, §81], voir aussi [Wittgenstein, 1919, 4.121]

On peut toutefois déduire de la lecture de Wittgenstein que ce n'est pas parce qu'il n'existe pas de langue idéale qu'il n'y a pas de concept dans la langue. Un concept correspond en fait à un sens (ou une famille de sens) fixé par l'usage et partagé par un ensemble d'individus. C'est parce qu'il existe des sens établis, partagés et non ambigus dans une situation donnée que les individus se comprennent (ou en tout cas, se comprennent assez pour échanger des informations avec un certain taux de succès). Le concept est donc ancré dans la langue et dans l'usage ; il dépend largement du point de vue, du contexte et de la situation[15].

Il faut sans doute relever que ces affirmations, qui peuvent sembler revêtir une certaine évidence aujourd'hui, n'allaient pas de soi dans les années 1920 : ce type de théorie sera essentiellement popularisé par les travaux de la philosophie du langage ordinaire, autour d'Austin notamment, dans les années 1950 [Austin, 1962].

La position de Wittgenstein ne correspond pas à un « relativisme » exacerbé. La vision du langage comme expérience personnelle doit être contrebalancée par l'argumentation du philosophe sur la notion de « langage public » [Gentner et Goldin-Meadow, 2003]. Il s'agit là d'un aspect fondamental de la pensée du philosophe : on observe effectivement, dans sa description du langage, un certain relativisme, dans la mesure où le sens dépend de l'intériorisation de l'énoncé mais la notion de consensus (ou d'inter-subjectivité) y est aussi fondamentale.

> *La nature subjective des concepts n'implique pas qu'il soit impropre de dire de deux personnes qu'elles "ont le même concept" ; conformément à mon explication du mot "concept", cela voudra dire qu'elles ont la même capacité mentale (...). Il y a bien entendu tous les degrés intermédiaires possibles entre les cas où nous dirions sans hésiter "ils ont des mots différents, mais ils les utilisent de la même façon ; ils ont le même concept", et les cas où nous dirions plutôt "ce n'est pas seulement une différence verbale, ils ont des concepts différents"* ([Geach, 1957], cité par [De Lara, 2005, p. 66]).

C'est parce qu'il y a des notions partagées de façon indiscutable, qu'il existe des situations (appelées *formes de vie* chez Wittgenstein) dans lesquelles on peut nommer des objets sans ambiguïté et que la compréhension est possible. Cette vision nous semble particulièrement éclairante pour le problème internalisme *vs* externalisme et n'est pas sans rappeler ces propos de Rastier :

> *Ni interne ni externe, la langue est ainsi un lieu du couplage entre l'individu et son environnement, parce que les signifiants sont externes (bien que reconstruits dans la perception) et les signifiés internes (bien que construits à*

[15] Wittgenstein est souvent rapproché de Malinowski. Celui-ci souligne par exemple que « la signification est inextricablement mêlée, et dépendante de l'activité en cours au moment où l'énoncé est prononcé ». Voir [Malinowski, 1923] et l'étude de Nyíri [1996].

> *partir d'une doxa externe). Comme le langage fait partie du milieu dans lequel nous agissons, c'est dans des pratiques diversifiées, dont témoignent les discours et les genres, que nous nous lions à notre environnement. Mais il est aussi peuplé de « choses » absentes, et dans l'expérience de l'altérité, du passé, de l'étranger, la culturalisation de l'enfant a lieu — non moins sinon plus que dans l'expression d'une expérience individuelle limitée au hic et nunc..* [Rastier, 2006]

Un ensemble de mots renvoie à un même concept s'ils partagent, non des propriétés communes, mais un certain « air de famille ». Wittgenstein utilise la métaphore bien connue des jeux pour expliquer cette notion [1953, §66]. Il existe des éléments (objets concrets ou pensées abstraites) qui forment des ensembles cohérents (des *catégories*) sans que l'on puisse dire exactement ce qui fonde cette cohérence. Quel rapport entre un jeu de marelle et un jeu d'échec ? Entre un jeu de fléchettes et un jeu de rôle ? Wittgenstein montre qu'il n'est pas simple de répondre à cette question et, d'une manière plus générale, qu'il est difficile de définir une catégorie par des propriétés communes partagées (ce qui peut poser problème si on croit en la notion de *prototype* ou de *parangon* [Rosch, 1973; Lakoff, 1987]). Il est ainsi toujours possible d'ajouter de nouveaux éléments à une catégorie donnée ; une catégorie a généralement des frontières floues et mouvantes, qui dépendent étroitement de la situation considérée. À l'inverse, elle doit être partagée entre individus de telle sorte que la communication soit possible. On verra que cette difficulté à catégoriser est récurrente et se pose constamment dans le cadre des applications présentées dans les chapitres suivants.

La communication implique des règles, ce qui amène Wittgenstein à aborder la notion de « jeu de langage » [1953, §7]. Un jeu de langage repose sur un ensemble de règles (et sur la possibilité de s'écarter des règles), qui semblent fixer le sens en fonction du contexte. La règle n'est pas une loi fixée une fois pour toutes : elle correspond au simple constat d'un sens fixé par l'usage (il ne s'agit bien évidemment pas sous la plume de Wittgenstein de désigner ainsi les règles de grammaire, au sens traditionnel du terme). Le jeu de mot apparaît naturellement quand deux règles peuvent s'appliquer dans une situation donnée, mais Wittgenstein vise une perspective plus large. Les jeux de langage décrivent des systèmes interdépendants, qui prennent sens par rapport à des situations données, avec des usages stabilisés.

> *Comprendre une phrase veut dire comprendre un langage. Comprendre un langage veut dire maîtriser une technique.* [Wittgenstein, 1953, §199]

Maîtriser une langue, c'est avoir la capacité de communiquer en respectant les usages de manière appropriée [1953, §66]. Mais l'usage repose en grande partie sur des règles implicites, écrites nulle part.

Pour Wittgenstein, la règle ne peut être dite, ou plutôt, elle peut toujours être contredite car elle ne correspond *in fine* qu'à l'usage. C'est pourquoi, par exemple, l'individu est

souvent incapable d'expliquer les règles suivies[16]. De là résulte aussi la difficulté à mettre au point des systèmes de compréhension, dans la mesure où l'usage n'est pas fixe, même si la variation peut être limitée dans des situations ou des cadres donnés.

1.2.2 Héritage philosophique et tradition linguistique anglo-saxonne

La pensée de Wittgenstein a connu un intérêt certain en linguistique à partir des années 1950, surtout dans la tradition anglo-saxonne. Le fait que le sens d'un mot apparaît en contexte (*by the company it keeps* pour reprendre l'expression de Firth [1957a]) est une intuition largement répandue dans le monde anglo-saxon. Il existe en effet outre-Manche une tradition bien établie de linguistique appliquée, à base d'observations sur corpus représentatifs, dont l'origine est largement antérieure à la généralisation de l'usage des ordinateurs. J. Léon [2007a] rappelle ainsi les études sur les co-occurrences au sein de la Bible, remontant au 18e siècle (travaux de Cruden, 1700-1770). La co-occurrence de mots dans un contexte donné permet de définir des familles paradigmatiques, c'est-à-dire des ensembles de mots de sens proches, qui partagent un certain « air de famille ». Maîtriser ce phénomène, c'est mettre au jour un système de règles valable dans un contexte donné.

Nous retiendrons trois écoles de pensée auxquelles nos travaux font écho. Toutes ne se réclament pas de Wittgenstein mais nous pensons que les trois se donnent à lire de façon pertinente à la lumière de la pensée du philosophe. Ces trois courants sont les suivants :

- les recherches autour de la notion de primitive sémantique par le groupe de Masterman à Cambridge, thème de recherche qui subira un essor massif sous des avatars divers, au sein de l'Intelligence Artificielle à partir des années 1960 jusqu'au web sémantique aujourd'hui (cf. la tâche d'annotation sémantique, chapitre 2).
- les recherches sur les sous-langages de Z. Harris, qui peuvent servir de base au calcul de classes sémantiques et de cadres de variation contraints (cf. chapitre 3). Même si Z. Harris ne fait pas appel à Wittgenstein dans ses travaux, on ne peut qu'être frappé par certains rapprochements avec la pensée du philosophe.
- les recherches sur l'analyse des collocations puis de la cohésion textuelle par Firth et ses successeurs, au premier rang desquels Halliday et Sinclair. Ces thèmes conduisent naturellement à aborder les notions d'architecture textuelle et de typologie de textes (cf. chapitre 4).

Il faut noter que ces courants ne sont bien évidemment pas indépendants les uns des autres. Les techniques d'analyse des collocations par des méthodes statistiques ont été largement explorées par le groupe de Masterman qui s'en est emparé pour identifier des primitives sémantiques au sein des textes. Firth a assisté aux premiers travaux du CLRU et aux discussions autour de la notion de langue universelle, même si c'était pour s'y opposer. L'approche de Harris mêlant analyse distributionnelle et sous-langage rappelle bien évidemment les travaux de Firth et de ses successeurs.

[16] Il s'agit pour une large part d'un processus naturel, voire inconscient [Le Du, 2005; Devitt, 2006]. La notion de règle, comme la relation entre Wittgenstein et l'inconscient ont été parmi les questions les plus débattues [Kripke, 1982 (1972); Bouveresse, 1992].

Firth et la notion de collocation

Nous nous appuyons ici essentiellement sur le recueil des œuvres de Firth [1957b] et sur l'étude de J. Léon « Meaning by collocation, the Firthian filiation of corpus linguistics » [2007b][17]. Deux éléments retiennent principalement notre attention quant à l'œuvre de Firth : la notion de collocation et la notion de (poly-)systémique.

Firth est connu pour avoir le premier décrit avec précision la notion de collocation, qui concerne non seulement la co-occurrence de deux mots mais aussi l'apparition conjointe de deux unités textuelles, de quelque nature que ce soit. Une citation de Firth est reprise dans nombre d'articles de linguistique dite de corpus : « *you shall know a word by the company its keeps* » (voir, par exemple, l'article de Church et Hanks [1989]). Firth a donné une définition plus précise de la notion de collocation :

> *Meaning by collocation is an abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea approach to the meaning of words. One of the meanings of* night *is its collocability with* dark, *and of* dark, *of course, collocation with* night. [Firth, 1957a, p. 196]

Pour Firth, la collocation forme un système et change le sens du mot isolé. La sémantique est d'une manière générale inséparable du contexte linguistique et de la situation d'énonciation : c'est pourquoi il promeut l'idée de systémique. La systémique prend en compte des facteurs extérieurs à la langue mais nécessaires à l'interprétation. Firth conteste le fait que la langue constitue un système cohérent que l'on pourrait saisir dans sa globalité : il promeut à l'inverse l'idée de langages restreints. Ce sont autant de micro-systèmes qui interagissent entre eux pour former une « poly-systémique » [Firth, 1957c].

Il convient donc d'étudier la notion de collocation et, plus généralement, la langue, à l'intérieur d'un « langage restreint » qui sera plus simple à cerner pour le linguiste. « *A restricted language can be said to have a micro-grammar and a micro-glossary* » [Firth, 1968]. On peut ici faire le rapprochement avec la notion de jeu chez Wittgenstein : un jeu comporte un ensemble de règles et un ensemble d'éléments que l'on doit maîtriser. La tâche du linguiste, selon Firth, est de mettre au jour la grammaire et le glossaire d'un langage restreint —, idée que l'on retrouve également, de manière très similaire chez Harris, à travers la notion de sous-langage.

J. Léon [2007b] souligne à juste titre que, si Firth a encouragé un modèle linguistique empirique visant à valider des hypothèses théoriques, il ne s'est guère penché sur la notion de corpus, sur les critères de constitution ou sur la notion de représentativité. Il ne mentionne pas non plus d'approche probabiliste, même si la notion de collocation porte en elle l'idée de probabilité. M.A.K. Halliday, au début des années 1960 (au sein du CLRU) reprend à son compte la notion de collocation : il lui donne un tour paradigmatique et étudiera la notion de thème et de cohésion textuelle à travers nombre de ses écrits [Halliday et Hasan, 1976; Halliday et Matthiessen, 2004].

[17] Sur la linguistique de corpus dans la période de l'après guerre en Grande-Bretagne, voir aussi [Williams, 2006].

Le calcul des collocations va se formaliser au niveau mathématique sous l'influence de Halliday puis de Sinclair. En particulier, la notion d'information mutuelle est en germe dans les travaux de Firth et sera progressivement formalisée par Sinclair [1966]. L'étude avant tout manuelle et intuitive de Firth dans les années 1930 à 1950 prend ainsi un tour beaucoup plus mathématique. Faute de moyens informatiques, Sinclair laissera cependant de côté ces recherches jusqu'à un retour en force dans les années 1980 et surtout 1990, alors que corpus et moyens de calculs sont disponibles (source : Léon [2007b]).

Le CLRU et la notion de primitive sémantique

Les travaux du Cambridge Language Research Unit (CLRU) sont directement en phase avec ceux décrits dans la section précédente : le CLRU est fondé et dirigé par Margaret Masterman (une ancienne étudiante de Wittgenstein à Cambridge) alors que Firth est encore actif (1955) ; à ses débuts, le CLRU a reçu la visite de Firth puis a accueilli M.A.K. Halliday de 1960 à 1966 (sur tous ces aspects, voir [Léon, 2002a]).

De manière moins anecdotique, même si le CLRU se démarque sur bien des points des travaux de Firth, la filiation avec Wittgenstein, la détermination du sens par l'usage et la priorité donnée à la sémantique face à la syntaxe sont partagés par les deux courants de recherche. Une différence majeure provient du fait que Firth était un pur linguiste, alors que Masterman s'intéresse de manière explicite au traitement automatique des langues : la traduction automatique (TA) est l'application phare de l'époque, qui a concentré l'essentiel des efforts du CLRU des années 1950 au début des années 1960 [Masterman, 2005]. L'intuition de Masterman (relativement originale à l'époque, surtout face au continent américain) est qu'il faut privilégier pour la TA une approche fondée prioritairement sur la sémantique et non sur la syntaxe. Mais ce point de départ n'est pas sans poser de multiples questions : qu'est-ce qu'un sens ou qu'un atome de sens ? Quelle représentation donner du texte ? Existe-t-il un mode de représentation sémantique indépendant de la langue ? Ces questions restent largement ouvertes cinquante ans après, faut-il le préciser...

Il serait illusoire de chercher des réponses définitives à ces questions mais les chercheurs de Cambridge, sous l'impulsion de Masterman, partent de la pensée de Wittgenstein pour guider leurs investigations. La langue est un système de signes qui fait sens en soi et qui ne peut être modélisé par un mode de représentation d'une autre nature. Cependant, l'ambiguïté inhérente au langage pose problème au TAL et il faut bien trouver un mode de représentation, une *interlingua* qui permette de rendre compte du sens, en résolvant les principales ambiguïtés et en rendant possible des équivalences entre langues. Pour Masterman, il est clair que cette *interlingua* ne peut être qu'un système de signes prenant sens par rapport au système linguistique lui-même.

La vue la plus « positiviste » de cet *interlingua* est sûrement celle de Richens, qui propose un système de primitives organisées en système (*Nude*) mais dont les fondements semblent mal assurés [Spärck Jones, 2000]. Masterman, au-delà de discussions sur la nature des primitive de Nude, prolonge ces recherches en exploitant le Roget, un thesaurus à partir duquel il est possible de dériver un réseau sémantique : les mots sont classés par rapport à une taxinomie où les éléments non terminaux sont appelés des têtes sémantiques. Un mot

peut apparaître sous plusieurs tête sémantique, s'il est polysémique. En reliant entre eux les mots partageant les mêmes types, on obtient des classes puis un réseau sémantique, certes un peu grossier, mais néanmoins exploitable. On s'éloigne ainsi du langage universel de Nude pour se rapprocher du mot de Firth « *you shall know a word by the company its keeps* ». Le réseau est toutefois fondé sur un thesaurus de langue générale et l'on sait que Firth, qui avait assisté à une réunion du CLRU autour de ces problèmes, a fortement critiqué cette approche (car elle s'oppose à l'idée de « langage restreint », *a priori* plus facilement abordable en linguistique).

La notion de primitive sémantique peut s'entendre dans deux sens différents : soit il s'agit du sens premier, originel, « pur » à l'image des Idées de Platon (approche que rejette vigoureusement Masterman). Soit il s'agit d'observer qu'un sens « prime » en fonction du contexte, sans que cela ne préjuge en quoi que ce soit de la situation de ce sens par rapport à une théorie de l'esprit. La première option est par exemple celle du mentalais de J. Fodor [1975] (ou la *lingua mentalis* de Wierzbicka [1980]). La deuxième voie, plus difficile et moins claire de prime abord, sera explorée au CLRU à travers les travaux de K. Spärck Jones [1964] et, surtout, de Y. Wilks [1975]. K. Spärck Jones propose dès les années 1960 une méthode de désambiguïsation sémantique fondée sur le contexte d'apparition des mots [Spärck Jones, 1964] : la méthode de Spärck Jones repose sur la notion de cohésion textuelle. La redondance implique l'usage de mots de champs sémantiques proches ou similaires, qui peuvent être identifiés à l'aide d'un réseau sémantique (lui-même issu du thésaurus anglais Roget). Par la suite, Yarowsky a proposé une méthode de désambiguïsation sémantique reprenant largement celle de Spärck Jones et évaluée avec succès sur des corpus variés [Yarowsky, 1992].

Y. Wilks pousse plus loin cette idée à travers la notion de préférence sémantique. Wilks part du principe que le sens du mot dépend étroitement de son entourage syntaxico-sémantique : le contexte permet de faire émerger le sens, par un système de préférences et d'attirances mutuelles entre primitives sémantiques. Il réévalue le poids de la syntaxe, largement sous-estimé auparavant au CLRU. Il montre aussi comment la sémantique peut guider la syntaxe, par exemple pour le rattachement prépositionnel, qui peut se calculer par le nombre de connexions entre mots inférés à partir d'une base de connaissances [Wilks *et al.*, 1985]. Toutefois, le lien avec la notion de genre et de sous-langage, tel qu'évoqué par Firth ou, à la même époque, par Harris, n'est pas vraiment présent dans les travaux du CLRU.

Harris et les sous-langages

Indépendamment des travaux présentés dans les sections précédentes, c'est bien évidemment à Harris que l'on songe quand on parle de sous-langage, d'analyse distributionnelle et de restrictions de sélection (voir [Habert, 1998] pour une bonne synthèse).

Dès les années 1950, Harris développe une théorie à la frontière de la théorie de l'information et de la linguistique [Harris, 1951, 1988]. Il part du principe que les mots n'apparaissent pas aléatoirement dans la langue : c'est le travail du linguiste que de déterminer les combinaisons possibles et impossibles. Appliquées aux sous-langages (médecine,

biologie), ces combinaisons sont très caractéristiques et peuvent s'exprimer sous forme de contraintes :

- Un sous-langage est marqué par un vocabulaire réduit et moins ambigu que la langue générale (« un sens par domaine » (qui sera étendu à « un sens par texte » chez Yarowsky [Gale *et al.*, 1992]) ;
- Il est marqué par des classes de mots qui co-occurrent ensemble et que l'on peut repérer par analyse distributionnelle (axe paradigmatique) ;
- La dépendance entre éléments est déterminée par une structure de phrase simple à partir de laquelle des variantes peuvent être calculées (axe syntagmatique).

On peut mettre la philosophie de Wittgenstein en regard de ce programme et constater un certain nombre de rencontres intéressantes : il est possible, sans tordre la pensée des uns ou des autres, de dire que les classes de mots partagent un air de famille et que leur organisation syntagmatique permet de décrire leur sens.

Les classes sémantiques ont des liens entre elles et forment des ensembles d'opérateurs et d'arguments. Il y a une co-détermination entre l'opérateur et ses arguments (l'opérateur est déterminé par la classe d'arguments qu'il accepte et *vice versa*). Comme le rappellent R. Dachelet [1994] et B. Habert [1998], la théorie de Harris s'accompagne d'une méthode d'analyse originale. Pour parvenir à identifier les classes d'opérateurs et d'arguments, il faut d'abord analyser le texte syntaxiquement, puis le normaliser pour le transformer en un ensemble de phrases simples, sans transformation. C'est sur cette base que peuvent être calculées les classes d'opérateurs et d'arguments, en fonction de l'analyse distributionnelle.

Ces principes d'analyse ont été appliquées dès les années 1970 à des textes de biologie par Z. Harris [1989] et surtout, à des textes médicaux à l'université de New York par l'équipe de N. Sager [1987]. Dans les années 1990, ce courant a reconnu un certain renouveau avec l'importance de corpus électroniques (pour une synthèse d'ensemble, voir [Nevin et Johnson, 2002], pour des travaux sur le français, voir [Dachelet, 1994] ou [Habert, 1998]).

Dès lors, on voit plusieurs points communs avec les travaux menés dans les années 1950 et 1960. L'analyse distributionnelle est par exemple l'élément fondamental de la méthode harrissienne et s'oppose à la notion de structure profonde explorée dans la même période par Chomsky ; la notion de sous-domaine (domaine de spécialité marqué par un « sous-langage ») est également un point fondamental. Enfin, les travaux de Z. Harris intègrent de façon très fine l'analyse syntaxique, ce qui rend l'approche beaucoup plus puissante que les travaux précédents.

1.2.3 Remarque sur les méthodes probabilistes en linguistique

Les sections précédentes ont montré, dès les années 1950, des approches empiriques, fondées sur l'analyse de corpus et privilégiant la sémantique par rapport à la syntaxe. Les techniques probabilistes sont le bras armé de ces études, afin de mesurer des classes de régularité, des restrictions de sélection et des relations d'ordre entre opérateurs et arguments. Ces approches ont toutefois été rejetées dans l'ombre à partir des années

1950, au profit d'approches formelles dans la lignée des travaux de Chomsky [1957]. La notion d'empirisme est alors fortement critiquée au profit d'un nouveau modèle fondé sur une série d'oppositions, comme compétence *versus* performance ou grammaticalité *versus* agrammaticalité [Chomsky, 1965]. Les raisons de l'opposition de Chomsky aux approches probabilistes sont rappelées par C. Manning [2003] :

- les méthodes probabilistes mélangent de façon erronée connaissances sur la langue et connaissances sur le monde : le nombre d'occurrences de *Paris* a des chances d'être supérieur à celui de *Pontault-Combault* dans un corpus constitué par exemple à partir du journal *Le Monde* mais ceci est dû à des facteurs pragmatiques et non à des facteurs linguistiques ;
- les méthodes probabilistes ne permettent pas de prédire l'agrammaticalité d'une séquence : les deux séquences « *Furiously sleep ideas green colorless* » et « *Colorless green ideas sleep furiously* » sont improbables, donc la probabilité d'apparition sera égale à 0 dans les deux cas, alors que la première séquence est agrammaticale et l'autre non. Comme le souligne C. Manning [2003], cette affirmation est bien évidemment fausse dans la mesure où les probabilités sont le plus souvent inférées sur des étiquettes et non sur une observation directe des formes de surface ;
- les méthodes probabilistes permettent de modéliser les réalisations de surface alors que ce sont les structures profondes du langage qui doivent être étudiées.

Une partie de ces objections tient sans doute au programme de Chomsky lui-même. Du point de vue de l'analyse sémantique, le fait de prendre en compte à la fois des facteurs linguistiques et des facteurs extérieurs à la langue n'est pas un problème en soi, bien au contraire : on sait que la pragmatique ou les connaissances sur le monde sont nécessaires pour inférer le sens d'un énoncé. La notion d'agrammaticalité est tout aussi contestable et contestée : beaucoup de séquences, notamment à l'oral, ne respectent pas la grammaire traditionnelle, mais sont pourtant parfaitement compréhensibles. On préférera sûrement donner une approximation de l'acceptabilité en se fondant sur la notion de probabilité d'une séquence (ce qui n'a d'ailleurs qu'un intérêt limité du point de vue de l'analyse sémantique). La notion de structure profonde est abstraite Chomsky [1966] : on ne peut donc s'appuyer sur cette notion si l'on considère que ce sont les manifestations du langage (*l'usage*) qui dirigent le sens. D'une manière générale, l'ensemble des objections de Chomsky repose sur l'affirmation d'une autonomie de la syntaxe, affirmation largement contestée aujourd'hui.

Dès la fin des années 1960, les fondements du modèle chomskyen sont attaqués de l'intérieur, au motif que le modèle proposé est par trop théorique et ne tient pas assez compte des données (cf. R.A. Harris [1993]). Plusieurs chercheurs comme Lakoff (mais aussi Postal, Ross ou Katz) contribuent à faire éclater le modèle chomskyen et développent des approches alternatives, à l'origine de courants comme la sémantique générative, puis la sémantique cognitive [Fuchs, 2004]. Ainsi, la sémantique cognitive nie le fait que le langage soit une activité autonome, propose de relier la faculté langagière à d'autres facultés cognitives comme la perception, la catégorisation ou la mémorisation.

Les recherches menées dans ce cadre, par Lakoff notamment, reviendront aux écrits de Wittgenstein pour mettre au premier plan la notion d'air de famille (qui sous-tend chez

Lakoff la notion de prototype [Lakoff, 1987]), ce qui lui permet d'expliquer les catégories lexicales aux bords *flous*, la relativité des jugements de grammaticalité et les phénomènes de polysémie et de métaphores (pour aboutir à la notion de *fuzzy grammar* [Denison *et al.*, 2006]). Ces approches, même si elles présupposent souvent des mécanismes cognitifs partagés et universels, montrent l'importance des phénomènes d'inter-subjectivité, de distance culturelle et d'histoire personnelle (en un mot, ce que Pears appelle la simple *nature humaine* [2007]) pour la conceptualisation.

1.3 Retour à l'analyse linguistique

Si nous nous fondons sur les approches décrites dans les sections précédentes, le TAL fait face à un paradoxe : d'un côté une sémantique fondée sur l'usage, variable en fonction de la situation ; de l'autre la nécessité de normaliser le texte pour en donner une représentation exploitable pour des traitements automatiques.

Cette tension entre deux pôles opposés marque les chapitres qui suivent. En fonction des besoins et des applications, ce seront tantôt des modèles précis, visant à normaliser *a priori* le contenu textuel, tantôt des analyses reposant sur des techniques plus dynamiques qui seront privilégiés. Nous verrons alors que l'adéquation des modèles est souvent discutable et que l'interprétation des approches dynamiques est souvent difficile. Ce questionnement est cependant, de notre point de vue, intéressant car il montre la complexité intrinsèque de la langue quand on se confronte à des applications réelles, au-delà de quelques exemples jouets.

Avant d'aborder ces points précis, nous examinons ici la notion de modèle d'application, qui évite certains problèmes liés à la confrontation du langage à la réalité : le modèle d'application est un premier niveau de représentation, qui fige une image simplifiée du monde extérieur. Nous donnons ensuite certains points qui nous semblent importants car ils fondent les analyses présentées dans les chapitres suivants (l'importance du corpus, des connaissances extérieures et la volonté de traiter les différents paliers du texte afin de montrer la transversalité des problèmes envisagés — sur tous ces points, voir p. 1.3.2).

1.3.1 La question de la référence

Les applications auxquelles nous nous intéressons, et qui ont été présentées brièvement en introduction (systèmes de résumé, d'extraction d'information, de questions-réponses), demandent une compréhension minimale du texte. Il faut par exemple fournir une représentation formelle partielle du texte si on veut pouvoir faire les inférences nécessaires pour répondre à une question donnée. La tâche revient donc, dans une certaine mesure, à gérer la variation linguistique afin de donner une représentation du texte qui permette de le manipuler.

On peut voir là un paradoxe. Les applications visées répondent à un modèle, celui-ci comporte des catégories prédéfinies et des relations entre ces catégories : l'enjeu applicatif

consiste justement à identifier au sein du texte des séquences correspondant à des catégories et à des relations préalablement identifiées. N'y a-t-il pas là une contradiction entre un modèle d'application qui semble figer le sens et la notion de sens émergent de l'usage comme nous l'avons vu *supra* ?

En s'inspirant de Wittgenstein, on peut dire qu'il n'y a pas de conceptualisation *a priori*, qui existerait en dehors d'une pratique sociale ou d'une situation donnée. L'impossibilité du langage privé montre d'ailleurs les limites de la relativité dans la relation de l'individu au langage : variation du sens certes, mais aussi stabilisation au sein de relations d'inter-subjectivité (la notion de consensus chez Wittgenstein), c'est-à-dire d'une communauté de pratiques.

Un modèle d'application est le résultat d'un travail de modélisation autour d'une pratique sociale, exprimée à travers un besoin utilisateur. Cette mise en perspective peut sembler bien terre à terre face aux objectifs de la philosophie de Wittgenstein (rappelons encore une fois que celui-ci ne s'intéressait pas au langage en soi mais simplement à l'éclairage que la question du langage donne à certains problèmes philosophiques). Il n'empêche qu'il me semble tout à fait pertinent de prendre les idées de Wittgenstein dans un sens concret et de les confronter à des problèmes, banals en apparence.

Il existe de toute manière une différence nette entre un modèle de la langue et un modèle d'application[18]. Il est bien connu, mais il n'est pas inutile de rappeler l'extraordinaire plasticité de la langue : il existe une multiplicité de façons de dire la même chose, ou, du moins, des choses légèrement différentes qui seront cependant interprétées d'une manière quasi identique. Mon interlocuteur peut utiliser des mots, des tournures, des enchaînements différents des miens. Je peux moi-même exprimer la même idée de deux façons différentes à une minute d'intervalle. Et, pour paraphraser ce qui a déjà été dit, cette plasticité n'est pas une limite du langage ; c'est au contraire cette plasticité qui fait la richesse de la langue, l'intérêt et la complexité de l'analyse sémantique.

Cette variabilité pose problème pour l'analyse automatique, dans la mesure où celle-ci doit pouvoir faire face à une multiplicité de formulations différentes. Face à cela, le modèle est doublement réducteur : en amont car il « écrase » la variation linguistique et, surtout, en aval car il ne dit rien de l'interprétation du texte.

Il faut souligner que cette simplification est précisément un des buts applicatifs que nous nous sommes assignés : identifier les différentes façons d'exprimer une idée afin de fournir des outils d'analyse efficaces (« l'énigme du même, de la signification même, l'introuvable sens identique, censé rendre équivalentes les deux versions d'un même propos », [Ricœur, 2006, p. 45]). Dans notre cadre, seul l'utilisateur a une connaissance suffisante du contexte et de l'application pour interpréter les résultats fournis par la machine (si tant est que ceux-ci soient corrects) et résoudre l'énigme du sens posée par Ricœur.

Les applications ne se préoccupent donc pas d'analyse de la référence en tant que telle. La modélisation préalable permet d'évacuer en grande partie cette question ; en phase

[18] Je considère ici que le modèle de l'application décrit les objets dans une application donnée, leurs relations et leurs mises en langue possibles. Un modèle de langue vise à rendre compte de la langue en elle-même, indépendamment des applications ; je ne fais pas appel à ce type de modèle dans ce mémoire.

d'utilisation, c'est à l'utilisateur de mettre en rapport le modèle de l'application avec le monde extérieur, à travers le travail d'interprétation. Ceci n'empêche pas de découvrir, par des moyens linguistiques, des configurations textuelles entraînant des changement de sens ou de référence (notamment quand le modèle permet de rendre compte de ces variations). Mais c'est au niveau textuel, par le repérage de configurations langagières particulières, que nous abordons ce problème.

1.3.2 Éléments pour l'analyse

Nous énumérons ici quelques principes sur lesquels nous nous appuyons pour les analyses qui seront développées dans les chapitres suivants. Il s'agit là de principes généraux qui ne dépendent pas des tâches considérées.

Le texte comme point de départ et point d'arrivée

Les applications sur lesquelles nous avons travaillé, qu'il s'agisse d'extraction d'information ou de systèmes de questions-réponses, visent à identifier au sein d'un corpus une information pertinente (par rapport à une requête ou un besoin précis). L'information n'est pas directement formalisée et on ne cherche pas à lui attribuer une valeur de vérité.

La référence n'est pas traitée en tant que telle dans les systèmes d'analyse sémantique de surface évoqués précédemment. Le modèle de l'application tient lieu de cadre et l'analyse linguistique ne dit rien quant à la validité du cadre envisagé. Le travail de modélisation est bien sûr très lié à la référence mais ce travail de modélisation proprement dit n'est pas abordé ici. Sa pertinence dépend d'une expertise qui précède (et dépasse) l'analyse linguistique.

Signalons par ailleurs que l'analyse de certains éléments parmi les plus complexes de la langue (comme les négations, les quantifieurs, les modaux...) est très largement négligée, et l'analyse fait peu de cas des nuances de sens. On peut regretter que l'analyse reste en surface et que le lien avec des techniques plus fines, qui pourraient mieux prendre en compte ces éléments parfois cruciaux, soit négligé (par exemple, ne pas traiter la négation amène évidemment à des contre-sens fréquents). On touche là aux limites de ce type de systèmes qui, de fait, délèguent une partie de la tâche à l'utilisateur.

L'interprétation pour donner du sens

Les systèmes d'extraction d'information fournissent traditionnellement une information brute, coupée de son contexte. Cette stratégie conduit à des problèmes évidents, dans la mesure où ils ne sont pas fiables à 100 % et où l'utilisateur ne peut vérifier l'information si elle n'est pas contextualisée.

Nous avons quant à nous toujours cherché à fournir un renvoi au texte, dans la mesure où l'outil aide le lecteur à chercher l'information pertinente mais ne peut se substituer à

lui pour vérifier l'information, la recouper ou la mettre en relation avec d'autres[19]. Même sur des textes techniques, le lecteur intervient avec sa subjectivité pour interpréter le texte [Rastier, 2001]. L'interprétation est une dimension qui échappe largement à la machine.

D'une manière générale, on pourrait imaginer des systèmes plus ambitieux que ceux mis en œuvre couramment (notamment pour l'analyse de la négation ou des modaux, comme nous l'avons vu au paragraphe précédent) mais on touche là à la limite des techniques actuellement utilisées. Soulignons quand même que, quelle que soit la finesse et la puissance des techniques utilisées, la dimension interprétative restera pour longtemps encore l'apanage de l'utilisateur.

Le corpus, représentatif d'une pratique

Nous avons longuement discuté tout au long de ce chapitre, à la suite de nombreux philosophes et linguistes, l'idée que le sens d'un mot ou d'une expression émerge du fait de son usage en corpus. Ce sont de nouveaux usages — l'emploi d'une expression dans un nouveau contexte par exemple — qui font évoluer le sens des mots.

Ce point de vue a des implications très directes sur le type d'analyse mise en œuvre : il est ainsi primordial de recourir au corpus pour identifier les éléments pertinents par rapport à une application donnée. Les techniques mises en œuvre sont de natures diverses : analyse manuelle de corpus, analyse distributionnelle, techniques d'apprentissage. Nous verrons plusieurs de ces techniques en fonction de la complexité de l'information recherchée, de la taille du corpus, de sa régularité ou de sa variété. Ces techniques visent à faire apparaître des « familles sémantiques », c'est-à-dire des ensembles de mots qui, dans un contexte donné, ont un sens proche, autrement dit, des mots qui partagent un air de famille. Toutes ces familles de sens n'ont pas la même fonction. Certains éléments jouent le rôle d'opérateur, les autres jouent le rôle d'argument [Harris, 1988]. Il est donc important, au-delà de la simple mise au jour de familles sémantiques, de s'intéresser à leur organisation sur un plan syntaxico-sémantique.

Comme le rappelle Harris, aussi bien l'acquisition de classes paradigmatiques que leur organisation sur le plan paradigmatique requiert une analyse préalable du matériau textuel. Le texte brut, parce qu'il comprend des incises et des expressions complexes sujettes à variation, est un objet qui ne se prête pas directement à l'analyse distributionnelle. Le texte doit donc être « préparé » et annoté, ce qui revient à le normaliser. Ce processus de normalisation n'est pas neutre : il implique de connaître *a priori*, au moins partiellement, la nature des éléments pertinents pour la tâche. Quand l'analyse distributionnelle est faite sur un corpus brut, elle donne généralement des résultats très pauvres et peu opérationnels.

[19] La mise en surbrillance de l'information recherchée est un moyen pratique de garder le contexte, même si cette technique rend difficile la mise en évidence de relations entre éléments repérés.

Les connaissance sur le monde, pour dépasser les limites du corpus

L'échec des premières recherches en Traduction Automatique — et plus généralement en Intelligence Artificielle — vient en grande partie de la sous-estimation de la part des connaissances sur le monde dans l'interprétation des textes. La représentation des connaissances est depuis devenu un des courants majeurs de l'IA et c'est enfoncer une porte ouverte que de souligner son importance pour la compréhension. C'est aussi pour cela que nous avons mis en avant dans ce chapitre les recherches menées au CLRU : il s'agit d'un des rares groupes à avoir souligné, dès les années 1950, l'importance de la notion d'*interlingua* pour guider l'analyse. Cette *interlingua* est en partie un langage intermédiaire exprimant des connaissances sur le monde[20].

Nous avons déjà souligné le fait que si le sens émerge de l'usage, celui-ci est aussi parfois insuffisant pour identifier le sens, par exemple si l'on dispose de trop peu de données pour en inférer des informations valables. Lors de l'analyse de textes, le système n'a souvent pas la possibilité de demander plus d'exemples à l'utilisateur (même si cette stratégie est parfois explorée à travers des systèmes interactifs). Il faut par ailleurs souligner que l'usage ne se limite bien évidemment pas au co-texte : la connaissance de la situation d'énonciation, du type de texte et du destinataire est fondamentale pour l'interprétation. Ce type de connaissances n'est pas directement appréhendable et doit être fourni soit *a priori*, soit dynamiquement, au système. Nous verrons, notamment au chapitre 4, que la situation et le genre textuel imposent des contraintes qui régissent l'agencement discursif.

On dispose aujourd'hui de bases de connaissances (connaissances linguistiques et/ou connaissances sur le monde) qui permettent en partie de contourner la difficulté. L'anglais est bien évidemment la langue la mieux dotée mais des données commencent à être disponibles pour le français (lexiques ou dictionnaires généraux, réseaux sémantiques, banques de termes spécialisés, *etc.*). Ces connaissances ne sont elles-mêmes pas parfaites, mais elles peuvent servir à guider l'analyse en identifiant des mots composés ou des termes techniques, voire des relations entre syntagmes. Ces ressources extérieures n'ont pas de statut spécial : elles divisent souvent le sens des mots de manière artificielle mais elles guident l'analyse en contexte et donnent souvent accès à des informations fondamentales pour l'analyse.

[20]Le rapport ALPAC sur la faisabilité d'un système de traduction automatique soulignait la complexité de la tâche en raison notamment de la polysémie et de la difficulté pour un système d'identifier le sens d'un mot au sein d'un texte donné. Il a pu être prouvé depuis que ce constat était contestable dans la mesure où il est souvent possible de déterminer le sens d'un mot d'après son contexte. Ainsi Y. Wilks [1975], définit des "formules sémantiques" permettant d'exprimer les jeux de sens préférentiels tels qu'on les trouve dictés par le contexte. Ces formules, codées manuellement, peuvent être en partie inférées à partir du corpus si on dispose de moyens adéquats (ce que Y. Wilks et son équipe feront dans les années 1980 et 1990 [Wilks *et al.*, 1996]). On peut aussi voir le dictionnaire génératif de J. Pustejovsky [1995] comme un prolongement de ces recherches tout comme les recherches, en France, de B. Victorri [1998]. Mais l'inférence du sens à partir des textes reste un processus lourd, encore mal maîtrisé et qui demande de larges masses de données. Or, beaucoup des applications auxquelles je me suis intéressé portent sur des domaines spécialisés, pour lesquels seuls des corpus de petite taille, souvent non annotés, sont disponibles.

Les différents paliers d'analyse

Je me suis focalisé dans cette partie essentiellement sur le sens des mots et sur leur mise en relation, ce que l'on peut assimiler à la notion de schéma prédicatif. Mais en filigrane, il a fréquemment été question de textes et de corpus : le sens d'une expression ou d'un item lexical n'a de réalité que par la variété de ses contextes d'usage, qui font apparaître des nuances de sens en fonction du contexte. Il s'agit donc d'un jeu d'oppositions qui se capte par une analyse globale du corpus faisant apparaître ici la présence d'un élément contextuel, là son absence.

De même, l'extraction d'information, domaine sur lequel j'ai tout d'abord travaillé, porte essentiellement sur la reconnaissance d'entités et de relations entre entités au sein de la phrase. Il s'agit donc d'une forme de compréhension très limitée du contenu textuel, qui doit être étendue si l'on souhaite parvenir à une vue plus globale de l'information véhiculée au sein d'un document. Je me suis alors intéressé à des approches prenant en compte des relations discursives et les caractéristiques typo-dispositionnelles des textes.

Ce découpage est toutefois relativement artificiel : il existe de nombreuses interactions entre ces différents paliers. Les entités nommées permettent ainsi de faire apparaître des liens entre documents, ce qui peut servir de base à la production de résumés multi-documents (cf. chapitre 2). L'analyse de grands blocs de texte repose aussi sur le repérage de marques lexicales de cohésion ou de rupture thématique ou rhétorique (cf. chapitre 4). On verra ici et là dans ce mémoire l'interaction des différents niveaux d'analyse mais il ne s'agit là que d'une esquisse tant cette voie semble riche.

1.3.3 Prendre du recul par rapport aux réalisations

J'examine dans cette section les raisons qui fondent cette étude, qui vise en définitive à décrire les rapports entre les modèles applicatifs et la réalité de la langue, que l'on cherche à appréhender le plus finement possible.

Questionner les modèles

Dans les chapitres qui suivent, je décris mes travaux et j'en examine les limites à la lumière des observations théoriques évoquées dans ce chapitre. Cependant, le projet que j'ai poursuivi ces dernières années n'a pas été celui d'une exploration continue et systématique sur le plan pratique d'un point de vue théorique, en fonction d'hypothèses fixées *a priori*. Cette approche serait sans doute plus conforme à une démarche scientifique classique, où l'on aborde un problème par la définition d'hypothèses en fonction d'un modèle que l'on cherche à évaluer, pour aboutir si possible à une nouvelle théorie.

Il n'aura pas échappé au lecteur que les différents aspects de la langue qui ont été examinés dans ce chapitre n'aboutissent pas à une théorie singulière. Dans le cadre des applications de TAL, la valeur du signe est inséparable du contexte et de l'interprétation qui peut en être donnée. Chaque application peut être vue comme un « jeu » particulier,

où le signe prend un sens particulier en fonction de règles (regroupées au sein du modèle que l'on se donne), et fournit un résultat plus ou moins conforme à la réalité.

Ce sont ces décalages entre les modèles proposés et la réalité de la langue qu'il me semble intéressant de continuer à explorer. Il est indéniable que l'ingénierie linguistique progresse et permet de répondre à des besoins précis[21]. Mon expérience passée, tant dans le monde industriel qu'universitaire, m'a amené à me confronter à des problèmes particuliers pour lesquels j'ai essayé de proposer des réponses efficaces. Dans quelle mesure ces réalisations correspondent-elles à la « *réalité linguistique* » ? Donnent-elles à penser sur la langue elle-même ? Quels décalages observe-t-on avec les observations formulées dans ce chapitre ? Il me semble qu'un regard critique sur les réalisations, ne portant pas tant sur leur côté pratique que sur ce qu'elles nous apprennent sur la langue (en étudiant leurs avantages et leurs limites), reste nécessaire. C'est cette mise en perspective j'ai essayé de mener à travers ce travail.

A ce propos, il me semble nécessaire de préciser que, contrairement à ce que l'on pourrait penser, une sémantique fondée sur l'usage n'implique pas obligatoirement des approches dynamiques, où le sens « émergerait » naturellement de l'analyse des différentes occurrences considérées. A l'inverse, comme il a été dit plus haut (et ce dès l'introduction), on sait que la langue est trop variée pour être appréhendée directement, sans précaution, par des moyens automatiques. Cette variation pose problème : c'est pour cela que Harris prône une normalisation des phrases avant de procéder à une analyse distributionnelle (celle-ci est sinon vouée à l'échec)[22]. La notion de normalisation étant elle-même vague, on retrouve alors rapidement les tensions entre le général et le particulier, entre l'ambiguïté de la langue et la nécessité de développer des types figés non ambigus[23].

Reconsidérer les catégories traditionnelles

En fonction de ces remarques, on ne s'étonnera donc pas de trouver dans ce qui suit des approches qui peuvent sembler contradictoires : la reconnaissance des entités nommées (chapitre 2) vise à catégoriser des séquences linguistiques en fonction de types figés et définis *a priori* ; l'analyse de texte médicaux (chapitre 4) vise de la même manière à instancier des modèles prédéfinis à partir de textes libres. A l'opposé, les expériences sur la métonymie des entités nommées (chapitre 2), sur la sous-catégorisation (chapitre 3) ou sur les typologies de textes (chapitre 4) explorent des techniques dynamiques, où l'on essaie de faire émerger des classes de comportement directement à partir des données, sans cadre prédéfini.

A y regarder de plus près, on verra des problèmes similaires émerger de ces approches très diverses, à commencer par la difficulté de toute tâche de modélisation et de catégori-

[21] Les campagnes d'évaluation, même si elles sont souvent contestables, permettent de mesurer des avancées dans des cadres précis.

[22] Par ailleurs le découpage des domaines est en ensembles significatifs (c'est-à-dire en classes d'usages particuliers) est également problématique.

[23] On retrouve bien évidemment ici des problèmes envisagés dans les travaux décrits *supra*, ceux du CLRU notamment.

sation[24]. Les séquences appelées entités nommées correspondent en fait à des classes mal définies, aux contours flous. Les textes médicaux sont hétérogènes et ne correspondent pas à des types précis. Les cadres de sous-catégorisation se laissent difficilement saisir car il n'y a pas de séparation claire entre arguments et modifieurs. Cela n'est pas sans rappeler la théorie du prototype qui a bien montré qu'il y avait des éléments plus caractéristiques que d'autres pour une classe d'éléments donnée [Rosch, 1973; Kleiber, 1990]. Il me semble important d'examiner les implications de ces remarques en ingénierie linguistique.

L'étude reviendra à plusieurs reprises sur les résultats respectifs des deux approches : analyse en fonction de types fixés *a priori* ou acquisition dynamique de classes de comportement. On essaiera le cas échéant de faire collaborer ces deux approches pour en voir l'apport respectif (voir par exemple les expériences sur l'acquisition de classes sémantiques au chapitre 3, où l'acquisition dynamique est guidée par des connaissances *a priori*). La mise en œuvre d'approches collaboratives n'est cependant qu'esquissée dans ce mémoire et on verra que de nombreuses pistes restent ouvertes, afin de mieux prendre en compte les particularités de la langue notamment.

Les travaux évoqués remettent parfois en cause des catégories et des problèmes classiques en linguistique. On verra par exemple que l'analyse statistique de corpus ne permet pas de retrouver l'opposition entre argument et modifieur autour du verbe. Mais cette distinction est-elle réellement fondée d'un point de vue linguistique[25] ? Ce questionnement porte aussi sur les stratégies d'évaluation. Le TAL a en effet adopté ces dernières années un mode d'évaluation positif (les cadres d'évaluation permettent de mesurer des avancées objectives des techniques et systèmes développés) mais aussi discutable (en laissant souvent penser que la solution de référence constituait une réalité objective). Il me semble à l'inverse nécessaire de prendre du recul pour s'interroger sur les réalisations, les modèles et les résultats obtenus, quelles que soient les approches adoptées, dans la mesure où chaque système vise à répondre à un besoin pratique et avéré.

1.4 Synthèse

J'ai exposé dans cette partie différentes considérations sur la langue, à partir de l'examen de travaux d'auteurs variés. Ces travaux ont été réinterprétés à l'aune de mes propres réalisations et fournissent un arrière-plan solide à partir duquel il est possible de jeter un regard mesuré sur mes réalisations passées. Il s'agit de voir en quoi celles-ci répondent à des besoins concrets et en quoi les résultats obtenus sont conformes ou non à l'image de la langue que l'on peut avoir.

[24]Notons cependant, comme les écrits de Wittgenstein semblent le suggérer, que l'hétérogénéité du signe et la liberté dans l'interprétation n'aboutissent ni à la subjectivité totale ni au relativisme forcené. Le linguiste (et l'utilisateur en général) a presque toujours une intuition claire sur le décalage entre le résultat obtenu et la réalité de la langue. L'application, parce qu'elle répond à un besoin particulier, est de ce point de vue extrêmement précieuse.

[25]Rastier a bien montré — pour le critiquer — le fondement ontologique de la grammaire [Rastier, 1999]

Chapitre 2

L'annotation sémantique

La compréhension de texte est généralement fondée sur un premier niveau d'analyse, visant à repérer et typer des séquences de textes pertinentes. La tâche de reconnaissance peut se borner à mettre en évidence les éléments à même le texte (par un marquage quelconque) : on parle alors d'annotation sémantique. Les éléments sont en général définis *a priori*, par le modèle de l'application visée.

Dans ce chapitre, nous nous interrogeons sur le statut de ces séquences et nous explorons la notion d'atome de sens. Nous nous focalisons ensuite sur le problème des entités nommées (noms de personnes, de lieux, d'entreprises...). Nous présentons un système de repérage opérationnel et nous en voyons ensuite les limites, notamment face aux phénomènes de glissements de sens comme la métonymie ou la métaphore.

2.1 Des atomes de sens ?

L'annotation sémantique est une tâche classique et bien connue du web sémantique. Quel contenu annoter ? Jusqu'où peut-on aller ? La tâche est-elle générique ? J'aborde ici ces questions.

2.1.1 Une normalisation nécessaire pour la compréhension automatique

Quasiment tout énoncé peut avoir plusieurs interprétations. Si le linguiste y voit généralement une richesse, certains auteurs comme Ogden et Richards [1923] y voient au contraire une source de malentendus et de conflits. De ce point de vue, il peut sembler vain de continuer à s'intéresser à la langue courante pour des traitements informatiques : il est alors tentant d'imaginer un nouveau langage, où chaque mot renverrait à une notion précise et nette, sans ambiguïté. C'est la tâche que s'était assignée Ogden à travers le développement du *Basic English* [Ogden, 1930], où 850 mots anglais sont censés suffire à

une communication simple et non ambiguë[26]. On retrouve ici toutes les utopies visant à développer des langages universels [Eco, 1994].

Cette idée est ensuite reprise en Intelligence Artificielle, sous des atours souvent cognitifs. On peut rappeler les positions de J. Fodor pour qui il existerait un langage de l'esprit, indépendant des langues parlées, constitué d'un ensemble de représentations cognitives universelles et innées : le « mentalais » (voir aussi la théorie de Wierzbicka, en partie inspirée de Leibiz [Wierzbicka, 1980]). Même si elle a une certaine logique et semble relativement confortable, cette position pose de nombreuses questions. Certaines peuvent sembler anecdotiques, comme la question de la création de nouveaux objets ou concepts (nos ancêtres avaient-ils déjà en tête le concept de grille-pain alors qu'ils ne maîtrisaient pas encore le feu ? [Wilks, 2005][27]). D'autres sont plus profonds, comme la connexion entre ces connaissances et la réalité de la langue : si celle-ci est tour à tour polysémique et ambiguë, il est logique de considérer que le langage mental puisse aussi rendre compte de ces aspects. On aboutit là rapidement à une aporie, comme le montre [Wilks, 2005], dans la mesure où l'ambiguïté reste présente là où on cherchait précisément à s'en débarrasser. On en déduit, avec Wittgenstein, que le seul mode effectif pour rendre compte du langage est le langage lui-même[28].

D'un point de vue informatique, on ne peut pas nier la nécessité de normaliser et de regrouper certains éléments par delà la variation langagière, fût-ce au prix d'une certaine simplification. L'annotation sémantique répond en partie à cet objectif. Il s'agit de reconnaître, au sein du texte, des éléments signifiants atomiques, qui pourront ultérieurement être mis en relation pour faire sens. L'hypothèse la plus fréquente est que ces éléments peuvent être nommés, qu'ils partagent des propriétés et peuvent être organisés dans des hiérarchies pour former des ontologies. Nous verrons *infra* les problèmes que pose ce point de départ simplificateur.

Il s'agit en effet de simplifier le problème de l'analyse de la langue par ordinateur : la machine doit résoudre de nombreuses ambiguïtés dont l'individu n'a même pas conscience. Il faut toutefois garder en tête le fait que les types d'éléments distingués répondent avant tout à des contraintes applicatives : celles-ci ne sont ni universelles ni fondées *a priori*, contrairement à certaines propositions souvent défendues au sein de la communauté du web sémantique.

[26] L'idée de définir des mots compliqués au moyen de mots plus simples est évidente et naturelle. Il ne s'ensuit évidemment pas que l'on puisse trouver un sous-ensemble de sens de base au moyen duquel tout le reste s'exprime.

[27] "It was this that drove Fodor (1975) to the highly implausible, but logically impeccable, claim that there is a language of thought predating real languages, and containing not primitives but concepts as fully formed as "telephone", on the ground that concepts cannot be built from or expressed by combinations of primitive concepts, and so must always be as wholes in any language of thought. This is, of course, the joke of a very clever man, but it is unclear what the alternatives can be, nor, more specifically, what an evolutionary computational theory of language can be."

[28] Il ne s'agit pas de nier l'intérêt d'autres modes de représentation comme la logique, mais beaucoup d'approches dites « formelles » me semblent déboucher sur des problèmes abstraits, qui n'éclairent guère notre connaissance du fonctionnement des langues.

2.1.2 L'annotation sémantique et le web

La volonté de créer un langage universel capable d'exprimer des connaissances générales est revenue en force dans les années 2000, à travers le web sémantique [Berners-Lee et Fischetti, 1999; Berners-Lee *et al.*, 2001]. Il s'agit d'associer des identifiants abstraits à des notions, puis de relier ces identifiants entre eux pour former un réseau de connaissances avérées, partagées et universelles. Cette vision du web sémantique peut être aisément critiquée tant elle semble reproduire les erreurs des premiers temps de l'Intelligence Artificielle [Sabah, 1988, 1989; Dreyfus, 1992]. La connaissance est relative à un point de vue, à une origine culturelle et elle évolue au cours du temps. Il est cependant vrai que certaines connaissances sont largement partagées, que n'importe quel individu sait faire des regroupements, des généralisations, même s'il a parfois du mal à définir exactement sur quelle base.

Comme le souligne K. Spärck Jones [2004], il existe différentes « visions » du web sémantique. La vision la plus ambitieuse consiste à élaborer un réseau universel de connaissances non ambiguës, avérées et interconnectées, permettant à des « solveurs » (modules d'inférence) de répondre à des questions, de résoudre des problèmes, voire d'interpréter des faits. Cette vision pose plusieurs problèmes, notamment :

- Le fait que ce modèle ignore la dimension sociale du web. Qui met en place ces connaissances, comment sont-elles validées ? Par quel domaine commencer ? Est-il possible d'obtenir un accord sur une description du monde, même partielle ? Et quelle granularité de description faut-il adopter ?
- Le statut de ces connaissances par rapport aux textes n'est pas entièrement clair. Celles-ci sont intégrées au texte quand il s'agit d'annotations, ou simplement liées au texte quand il s'agit de métadonnées. Il importe donc de s'interroger sur la nature du lien entre le texte et ces représentations conceptuelles. Peut-on mettre en correspondance des expressions employées en discours avec des connaissances statiques ? Comment gérer l'évolution du sens en fonction du contexte ?

Il est relativement simple de montrer — et l'histoire de l'Intelligence Artificielle le rappelle à l'envi — la difficulté de ces questions et le fait que l'on continue largement d'y buter. Il ne semble pas y avoir de progrès récents en linguistique ou en représentation des connaissances qui laissent entrevoir de réelle solution.

K. Spärck Jones [2004] souligne la difficulté de modéliser un ensemble de connaissances stables même pour des tâches *a priori* simples et bien balisées, comme l'indexation d'un fonds documentaire. La confrontation avec des documentalistes ou des bibliothécaires montre qu'il ne s'agit pas d'une tâche aussi simple qu'il y paraît, même s'il s'agit "juste" d'indexer des ouvrages (qu'il s'agisse d'une bibliothèque réelle ou virtuelle). L'indexation dépend du nombre de champs couverts, de leur stabilité, des intérêts des lecteurs ; les besoins peuvent être changeants, certains domaines disparaissent, d'autres apparaissent, *etc.*

De la même manière, la confrontation avec un système d'extraction d'information ou de questions-réponses est souvent révélateur. Même si le système répond parfois correc-

tement, il le fait d'une manière si mécanique qu'il ferait fuir n'importe quel utilisateur (à l'inverse du système Eliza [Weizenbaum, 1966] qui ne fournit pas de réponse précise mais semble beaucoup plus naturel en reprenant les mots de l'utilisateur, en s'intéressant à lui, bref, en étant coopératif). La réponse à une question n'est quasiment jamais un simple fait car l'utilisateur veut rebondir, avoir des compléments d'information, connaître le contexte d'un événement. Spärck Jones défend alors une position pragmatique et réaliste face au web sémantique : avoir un système ouvert, minimal et évolutif, qui puisse répondre aux besoins courants.

Dans ce qui suit, nous défendons l'idée que les entités nommées peuvent remplir ce rôle. La notion d'entité n'est pas vraiment fondée dans la théorie. Il s'agit d'une notion qui a émergé de la communauté travaillant sur des applications de compréhension de textes puis d'extraction d'information. Il s'agit d'éléments minimaux, se retrouvant dans de multiples applications, suffisamment génériques pour être en partie portables et réutilisables d'un domaine à un autre.

2.2 Les entités nommées comme éléments atomiques de sens

Cette section porte sur la notion d'entité nommées. Nous présentons tout d'abord une définition de la notion d'entités, les différentes classifications existantes et un système que nous avons développé initialement pour le français. Ce système, très classique, repose sur un ensemble de règles permettant de typer les séquences pertinentes suivant un jeu de catégories prédéfinies.

2.2.1 Retour sur la notion d'entité nommée

Les entités nommées (EN) désignent les noms de personnes, de lieux, d'organisations mais aussi les dates ou les unités monétaires. Voici la définition qui en a été donnée pour la campagne d'évaluation ESTER (Evaluation des Systèmes de Transcription Enrichie d'Emissions Radiophoniques, [Le Meur *et al.*, 2004]).

> *Les EN sont des types d'unités lexicales particuliers qui font référence à une entité du monde concret dans certains domaines spécifiques notamment humains, sociaux, politiques, économiques ou géographiques et qui ont un nom (typiquement un nom propre ou un acronyme). Une entité a généralement une existence relativement stable dans le temps, même si cette existence a un début (naissance, fondation, dépôt, formation...) et une fin (mort, dissolution, faillite, disparition...) et si l'entité évolue entre temps. Pour appréhender plus simplement cette notion, on s'appuiera de manière générale sur le principe du catalogue pour savoir si on a affaire à une EN. Ainsi si on peut aisément imaginer l'EN supposée comme étant une entrée d'un catalogue, annuaire, dictionnaire ou index alors celle-ci sera bien une EN. Les entités sont au cœur*

de la problématique de l'extraction de l'information d'un document. Par extension, on annotera les dates et les grandeurs physiques.

Comme on le voit dans cette tentative de définition, un moyen opérationnel utilisé aussi bien pour les campagne ACE (*Automatic Content Extraction*, [Doddington *et al.*, 2004]) que pour ESTER [Le Meur *et al.*, 2004] consiste à cerner la notion de nom de personne par une série d'exemples ou en ayant recours à la notion d'annuaire comme ci-dessus. Comme le souligne C. Chauviré [2003, p. 50] à propos du « second Wittgenstein » : « l'explication par les exemples ou par les échantillons est toute l'explication que peut donner un locuteur qui maîtrise un concept (c'est-à-dire sait l'appliquer à bon escient), et il n'en sait pas plus que cette explication, il ne possède aucun "universel" sous-jacent à ces exemplifications ou qui les transcenderait : les exemples constituent tout ce que le locuteur a à dire, étant entendu qu'il peut toujours donner des exemples supplémentaires le cas échéant ». Cette façon de caractériser les entités peut sembler insatisfaisant mais le propos de C. Chauviré rappelle implicitement que la classe considérée ne peut être définie par un ensemble de propriétés ; ses éléments partagent plutôt un « air de famille » (cf. chapitre 1, page 13).

Au delà, les tentatives de définitions insistent généralement sur le fait que les entités sont autonomes et monoréférentielles [Ehrmann, 2008] : *Jacques Chirac* désigne une personne précise ; le porteur du nom peut être identifié, même hors contexte. L'autonomie référentielle est nécessaire pour distinguer entre autres les entités des indexicaux (comme les pronoms « *je* » ou « *tu* » qui établissent un lien direct avec le locuteur, mais dont la référence dépend de la situation d'élocution). On notera toutefois que ces critères ne sont pas complètement opératoires. Un même nom peut tout d'abord désigner plusieurs personnes, ce qui contredit le critère de monoréférentialité. L'autonomie référentielle ne permet pas de distinguer les noms propres des expressions définies, alors qu'un syntagme comme *le Président de la République* a besoin d'être analysé pour recevoir une interprétation[29], contrairement à un nom comme *Jacques Chirac*. Enfin, les figures de style brouillent la référence, comme on le verra plus loin dans ce chapitre. Il ne s'agit évidemment pas de nier le caractère référentiel des entités, mais de montrer qu'elles se laissent plutôt saisir par leur place au sein du modèle d'application[30].

Alors que les noms propres ont longtemps été délaissés par les systèmes de traitement automatique (au prétexte qu'il s'agissait d'unités peu intéressantes qu'il suffisait de stocker dans un dictionnaire), le renouveau lié au travail sur corpus a révélé qu'il s'agissait en fait d'éléments clés pour l'analyse. Les conférences en extraction d'information (*Message Understanding Conferences*, [MUC7, 1998]) ont mis en avant plusieurs tâches génériques, au premier rang desquelles l'analyse des entités nommées. La tâche comporte en fait deux aspects : d'une part la reconnaissance des séquences pertinentes, d'autre part le typage en fonction d'une ontologie pré-établie. Sur des textes de type journalistique, les systèmes obtiennent généralement d'assez bons scores, avec un taux combiné de rappel et de précision supérieur à 0,90.

[29] Il faut notamment établir un référentiel temporel pour pouvoir interpréter le syntagme.

[30] Le modèle d'application définit un cadre qui restreint les possibilités d'interprétation et dans lequel les expressions définies peuvent parfois être assimilées à des entités, comme le propose [Ehrmann, 2008].

Du point de vue des applications informatiques, les entités nommées sont particulièrement importantes pour l'accès au contenu du document car elles forment les briques élémentaires sur lesquelles repose l'analyse (rappelons que les systèmes d'extraction d'information ou de questions-réponses ne visent en fait qu'une compréhension limitée, autour de certains éléments clés, à des fins de veille essentiellement ; il ne s'agit bien évidemment pas d'analyser ainsi n'importe quel type de texte). On ne se situe donc pas tant dans une perspective componentielle (réduction du sens jusqu'à des éléments atomiques) que dans une perspective constructionnelle (calcul du sens à partir d'éléments considérés comme atomiques). On verra par ailleurs dans la suite de l'étude que l'analyse ne peut pas être purement compositionnelle, dans la mesure où le contexte apporte son lot de contraintes et que le sens d'une expression ne peut pas toujours être simplement calculé à partir du sens de ses parties.

Le contexte influe sur l'interprétation des entités, ce qui empêche de les considérer comme des éléments directement référentiels et sans signification, contrairement à ce que proclament de nombreux auteurs. Comme le rappelle S. Leroy [2004b, p. 30], Mill [1995 (1843)] est celui qui a affirmé le plus clairement que « les seuls noms qui ne connotent rien sont les noms propres » et que ceux-ci n'ont à strictement parler « aucune signification. ». Ce point de vue est ensuite repris par Kripke [1982 (1972)] : ce sont les désignateurs rigides qui font référence aux objets du monde, organisés en ontologie. Le terme « entité nommée », copié de l'anglais *named entity*, reprend cette thèse : une entité n'a pas de contenu, il s'agit d'une simple étiquette sur un élément du monde. Même si Kripke et d'autres auteurs contemporains se placent dans une tradition logique, nous aurons l'occasion de montrer que cette tradition est très simplificatrice et ne correspond pas à la complexité langagière (voir section 2.4).

Dans un premier temps, les entités peuvent être assimilées aux noms propres. Toutefois, dès l'apparition du terme (concomitant des premières conférences en extraction d'information dans les années 1980), les dates, les données monétaires et d'autres éléments chiffrés sont ajoutés à la liste. Le terme prend un tour de plus en plus applicatif : il peut s'agir de noms de gènes et de protéines pour une application de biologie, de noms de maladies et de pathologies en médecine, voire de termes dans certains cas. La notion d'entité est alors clairement équivalente à un élément atomique de sens pertinent pour une application donnée. La notion d'entité n'est donc pas une notion générique mais dépend largement du modèle de l'application.

Intuitivement, on peut avoir l'impression que les noms de personnes ou les noms de lieux ont généralement une place plus centrale que les noms de protéines. Il s'agit en fait d'un artefact lié à la notion de situation commune. Ainsi, le web étant un média sur lequel on peut trouver des informations de toute nature, la notion de nom propre y a une place centrale. Dans le cadre d'une base spécialisée comme Medline (base de données de bio-médecine), les noms de protéines ou de gènes jouent le même rôle.

2.2.2 Systèmes de repérage et de catégorisation des entités nommées

Les entités nommées sont traditionnellement typées suivant une hiérarchie pré-établie. Nous examinons ce type de hiérarchie et les problèmes posés par les textes.

Hiérarchies de types d'entités

Sous l'influence des conférences américaines d'évaluation MUC, les travaux en extraction d'entités nommées ont d'abord été effectués sur des textes journalistiques ou des dépêches d'agence. Dans ce cadre, l'identification des entités nommées inclut trois types d'expressions :

- ENAMEX : les noms propres incluant les noms de personnes, de lieux et d'organisations.
- TIMEX : les expressions temporelles comme les dates et les heures.
- NUMEX : les expressions numériques telles que les expressions monétaires et les pourcentages.

Les hiérarchies comportent ainsi, pour les plus simples, une douzaine de types de base (feuilles de la hiérarchie) mais ont souvent besoin d'être étendues pour couvrir de nouveaux besoins (de nouvelles tâches ou de nouveaux domaines). Ainsi, il n'est pas rare de faire face à des hiérarchies de plus de 200 éléments, comme on peut le voir sur la figure 2.1 [Sekine *et al.*, 2002] (voir [Ehrmann, 2008] pour une étude très détailée de différents cadres d'annotation proposés pour la tâche).

Repérage et classification des entités nommées

De nombreux travaux ont porté sur l'identification des noms propres dans des textes journalistiques, notamment les *Message Understanding Conferences* [MUC7, 1998]. La reconnaissance des entités nommées à partir de textes écrits est actuellement la tâche d'extraction d'information qui obtient les meilleures performances. Celles-ci sont évaluées en utilisant des mesures classiques comme P&R (aussi appelée F1), correspondant à la moyenne harmonique de la précision et du rappel [31]. Sur des textes journalistiques, les taux obtenus sont quasiment comparables à ceux des humains, de l'ordre de 0,90 P&R (le score dépendant étroitement des types d'entités considérés). Ces résultats dépendent bien évidemment des particularités de la langue analysée : la tâche est beaucoup plus compliquée pour l'arabe par exemple, qui ne possède pas la notion de majuscule et dont certains éléments sont fortement ambigus (prénoms quasi systématiquement semblables à des noms communs, *etc.*).

Deux grands types d'approches sont généralement suivis pour l'identification des entités : une approche linguistique « de surface » et une approche probabiliste. L'approche linguistique est fondée sur la description syntaxique et lexicale des syntagmes recherchés :

[31] Soit la formule suivante : P&R $= \frac{2*Précision*Rappel}{Precision+Rappel}$

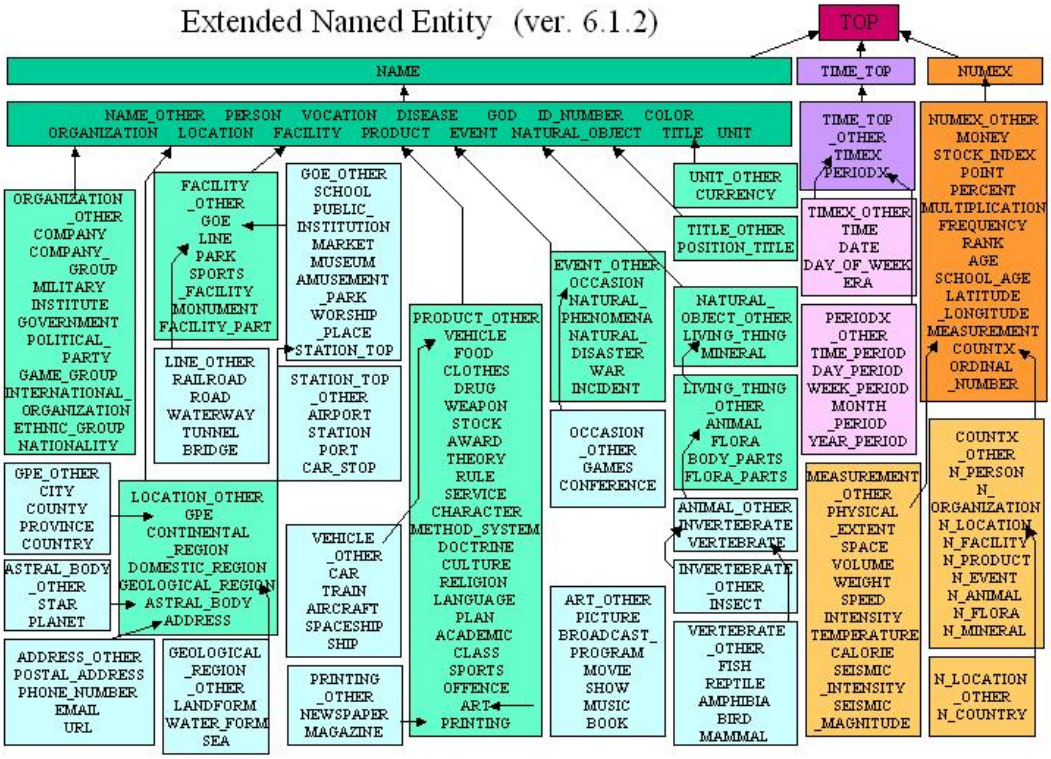

FIG. 2.1 – Hiérarchie d'entités définie par S. Sekine

les syntagmes intéressants sont repérés par des règles de grammaire fondées sur des marqueurs lexicaux (ex. *Mr* pour *Mister* ou *Inc.* pour *Incorporated*), des dictionnaires de noms propres et des dictionnaires de la langue générale (essentiellement utilisé en "négatif", pour repérer les mots inconnus) [Aberdeen *et al.*, 1995; Grishman, 1995; Appelt et Martin, 1999]. De son côté, l'approche probabiliste utilise un modèle de langage entraîné sur de larges corpus de textes pré-étiquetés. Cette approche est particulièrement robuste lorsque les textes sont bruités, c'est pourquoi la grande majorité des systèmes dédiés à l'oral adoptent une telle approche [Kubala *et al.*, 1999] .

2.3 TagEN, un système de repérage des entités nommées

J'ai développé un premier système de reconnaissance des entités nommées durant ma thèse, effectuée à Thales Recherches et Technologies. A mon arrivée au LIPN, j'ai développé en collaboration avec différentes personnes (notamment J.-F. Berroyer, qui a effectué un stage de Master sous ma direction en 2004) un nouveau système appelé TagEN, essentiellement pour les besoins internes du laboratoire. Ce système, fondé avant tout sur une analyse au moyen de transducteurs, n'est pas original en soi mais il offre une base nécessaire à de nombreuses applications. Il a par la suite été étendu à d'autres langues

et évalué au cours de campagnes officielles. Il a enfin été intégré à plusieurs chaînes de traitement linguistique, notamment dans le cadre du projet européen ALVIS.

2.3.1 Principes généraux

Les spécifications du système TagEN correspondent à celles du système mis au point au cours de ma thèse, et décrites dans [Poibeau, 2003]. L'idée principale est de développer un système hybride, en ayant recours à la fois à une base de règles et à des techniques d'apprentissage. Les règles permettent d'encoder des informations de façon lisible, efficace et facilement modifiable. Les techniques d'apprentissage permettent d'augmenter la couverture de manière automatique ou semi-automatique.

Aspects logiciels

Le système de repérage est dans un premier temps fondé sur des lexiques et des grammaires codées sous forme d'automates à nombre fini d'états. Même si les techniques d'apprentissage fonctionnent bien pour la reconnaissance des entités, j'ai choisi d'employer des automates dans la mesure où ceux-ci peuvent garantir un bon taux de reconnaissance et une bonne lisibilité des ressources. Il est en effet primordial que des linguistes ou des analystes puissent intervenir sur le système pour le faire évoluer, l'adapter et le spécialiser en fonction de nouveaux besoins. Les automates sont particulièrement adaptés dans la mesure où il s'agit d'un modèle de traitement efficace et très approprié au traitement de courtes séquences de texte.

Nous avons alors eu recours à la boîte à outils Unitex[32]. Il s'agit d'un logiciel libre facilement adaptable et intégrable à une chaîne de traitement. Unitex comporte un ensemble de programmes permettant de manipuler des transducteurs récursifs. Le logiciel comprend en particulier des programmes pouvant être appelés indépendamment pour la compilation et la minimisation de dictionnaires, le calcul de l'union ou l'intersection de graphes, *etc.*

TagEN encapsule les modules Unitex afin d'offrir un programme « appelable » en ligne de commande avec différentes options quant au format d'annotation (annotation sous formes de balises XML intégrées au texte ou déportées en dehors du document), aux entités prises en compte (annotation classique de type MUC ou spécialisées pour un domaine donné), ou encore quant à la langue du texte analysé (principalement anglais ou français, mais voir aussi 2.3.3).

Ressources

Les ressources sont de deux types : d'une part des dictionnaires, de l'autre des grammaires.

[32] `http://www-igm.univ-mlv.fr/~unitex/`

En plus des dictionnaires fournis avec Unitex, il est nécessaire de développer des dictionnaires spécifiques, encodant les informations pertinentes pour l'analyse des entités nommées. Il s'agit bien évidemment de listes de noms et de prénoms, de noms de lieux, *etc.* : ces listes doivent être importantes pour assurer au mieux la couverture (voir [Poibeau, 2003] pour plus de précision) : il est aujourd'hui possible de trouver sur le web des listes de plusieurs centaines de milliers d'items désignant des entités. Des listes trop importantes génèrent toutefois beaucoup de bruit, la plupart des entités considérées étant fortement ambiguës. L'analyse de corpus particuliers exige en outre une adaptation des ressources (par exemple, l'analyse récente de textes portant sur l'Iran a amené à ajouter de nombreuses entrées aux dictionnaires initiaux).

Les grammaires de reconnaissance sont codées, comme nous l'avons vu dans la section précédente, sous forme de transducteurs récursifs. Le graphe 2.2 fait ainsi appel à trois sous-graphes (en grisé) pour la reconnaissance des noms de personnes, de lieux et de sociétés.

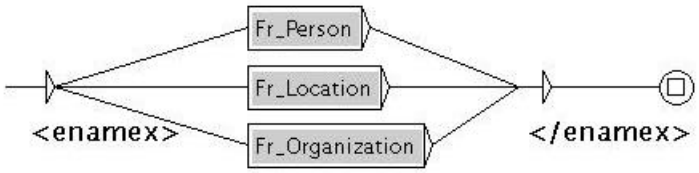

FIG. 2.2 – Graphe Unitex pour la reconnaissance des entités nommées

Chaque sous-graphe décrit lui-même un ensemble complexe d'éléments qui peuvent correspondre à des items lexicaux ou, plus souvent, à des étiquettes provenant des dictionnaires (voir la figure 2.3 pour un exemple de graphe, en l'occurrence le graphe des dates).

Pour plus de précisions sur les ressources, nous renvoyons encore une fois à [Poibeau, 2003] qui fournit une description beaucoup plus détaillée de la grammaire représentée sous forme de transducteurs.

Le système a été conçu de manière à ce qu'il soit aisé d'y ajouter ses propres ressources. Par exemple, si on travaille en biologie, on sera essentiellement intéressé par des noms de gènes et de protéines. Si on travaille sur Wikipedia, on sera davantage intéressé par des entités de type « classiques », à savoir des noms de personnes, de lieux, de sociétés. On voit là tout l'avantage d'avoir un système paramétrable, avec des ressources codées en clair et ne demandant pas d'entraînement sur un corpus annoté préalablement. Ce travail d'adaptation a été largement exploré dans le cadre du projet européen ALVIS, qui visait précisément à développer des chaînes de traitement robustes pour différents domaines de spécialité.

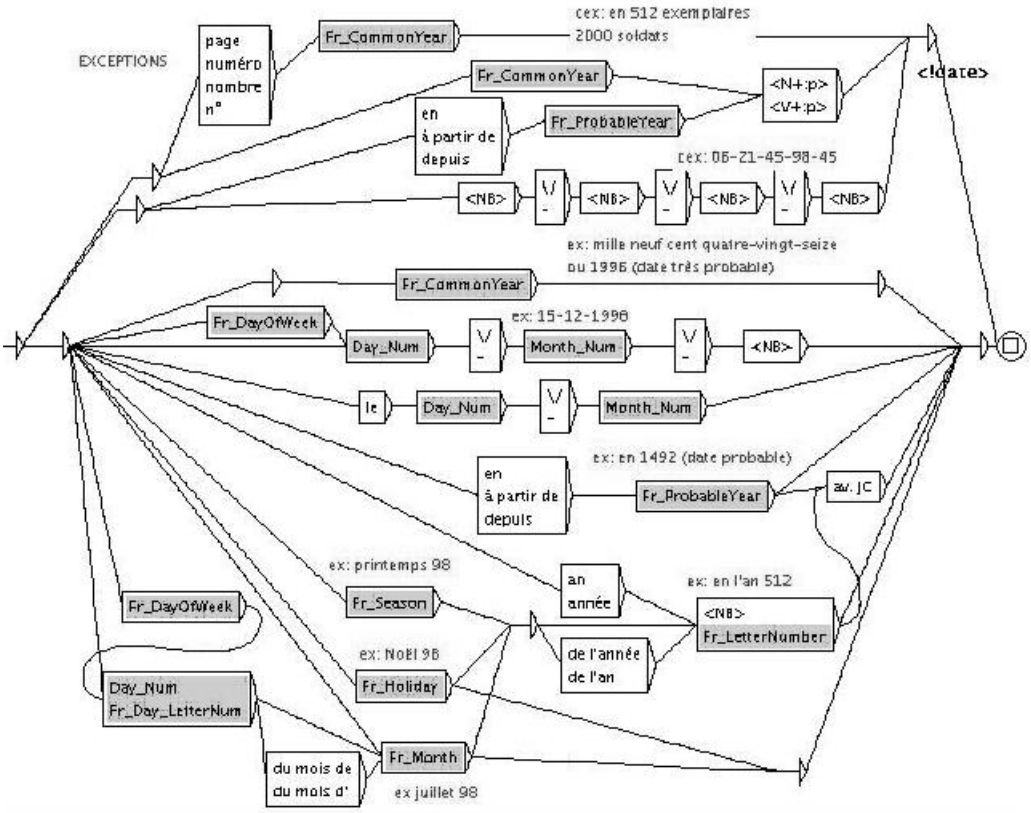

FIG. 2.3 – Graphe Unitex pour la reconnaissance des dates (les exceptions permettent de reconnaître explicitement des séquences ambiguës qui ne sont pas étiquetées en tant qu'entités).

Augmentation de la couverture par acquisition à partir de corpus

Il est possible d'étendre et d'améliorer le système initial par des heuristiques simples. J'ai montré comment le *typage dynamique des mots inconnus* au sein de séquences analysées pouvait se révéler pertinent. De la même manière, on sait que l'analyse de certains mots inconnus d'après le contexte d'apparition est très pertinente (par exemple, un mot inconnu au sein d'une énumération de noms de lieux a de fortes chances d'être un nom de lieu) [Poibeau, 2003].

Pour certaines séquences fortement ambiguës, si un sens prédomine au sein d'un corpus, j'ai montré qu'il était intéressant de limiter l'ambiguïté en attachant par défaut le sens le plus probable. Dans ce cas, certaines heuristiques peuvent aider à la *correction d'erreur* (pour re-typer l'entité en contexte, si certains indices laissent penser que l'entité est utilisée avec un sens différent du sens par défaut).

Des techniques d'apprentissage plus avancées seraient en outre efficaces pour augmenter la couverture. Nous explorons actuellement le couplage d'un étiquetage initial réduit mais fiable, avec des techniques d'apprentissage pouvant s'appuyer efficacement sur cette base (au moyen de *Conditional Random Fields* par exemple, dans le cadre du projet ANR

CROTAL, 2008-2009). Les CRF, en modélisant des dépendances à longues distances (ce que les autres modèles graphiques ne permettent pas de prendre en compte), peuvent permettre de typer des séquences en fonction d'un contexte élargi. Il a été montré que de tels contextes pouvaient être intéressants pour identifier de nouvelles séquences (en fonction de la place dans la phrase ou dans des structures types comme des énumérations par exemple) [Finkel *et al.*, 2005].

2.3.2 Évaluation et participations à des campagnes

Le système développé à Thales Recherches et Technologies avait été évalué dans le cadre d'applications d'extraction d'information [Poibeau, 2003], notamment sur le corpus MUC-6, qui est en quelque sorte un corpus que l'on peut qualifier de "stéréotypique". J'avais alors montré que ce système obtenait des performances proches de celles de plusieurs systèmes ayant participé à MUC-6, ce qui n'est guère étonnant dans la mesure où l'on trouve facilement sur le web des ressources adaptées à ce corpus.

Un travail plus important a été effectué en 2001 avec Leila Kosseim pour évaluer les performances sur plusieurs types de corpus, afin d'étudier la portabilité du système et la variabilité de l'expression des entités en fonction du corpus [Poibeau et Kosseim, 2001]. L'étude avait montré, sans surprise, que le système fonctionnait mieux sur des corpus réguliers (textes de journaux, dépêches d'agence) que sur des textes moins soignés (mails, transcription de l'oral). Les techniques de correction d'erreur et, surtout, de typage dynamique sont relativement efficaces pour améliorer les performances sur les corpus plus variés.

Le système développé en 2003 a été évalué en interne d'abord, puis à travers la participation à des campagnes d'évaluation (EQueR, ESTER) [Delbecque *et al.*, 2005], enfin à travers la participation à des projets comme ALVIS [Hamon *et al.*, 2005]. La campagne ESTER portait en fait sur l'analyse de l'oral [Gravier *et al.*, 2004] et TagEN, développé pour de l'écrit, n'était pas vraiment adapté (le logiciel ne comprend aucun mécanisme pour les particularités de l'oral comme les disfluences). Cette tâche était largement expérimentale (il s'agissait de la première campagne de ce type pour le français, écrit comme oral) et seulement trois systèmes participèrent. TagEN se classa en deuxième position, sachant que les trois systèmes avaient des performances très proches. Des traitements spécifiques pourraient largement améliorer ces résultats sur des données si particulières.

TagEN a enfin été intégré à des chaînes de traitement, notamment dans le cadre du projet européen ALVIS. Ce projet visant à traiter des documents sur le web, il exigeait des modules de traitement efficaces. Pour d'évidentes raisons d'efficacité, l'analyse linguistique ne peut pas être envisagée sur des textes tout venant sur le web. Elle est en revanche justifiée pour des domaines particuliers. C'est typiquement le cas de *crawlers*, qui sont en fait des « moissonneurs » spécialisés pour lesquels il peut être utile de développer des ressources propres. C'est dans ce cadre qu'ont été faites des expériences sur Wikipedia ou Medline, montrant que l'outil était capable d'annoter efficacement de grandes bases de données textuelles.

2.3.3 Modules multilingues

Des extensions multilingues ont été écrites dans le cadre d'un cours donné annuellement à l'Institut National des Langues et Civilisations Orientales (INaLCO). Sur deux ans[33], plusieurs groupes d'étudiants ont développé des ressources linguistiques permettant d'avoir des embryons de systèmes pour une dizaine de langues, parmi lesquelles l'allemand, l'arabe, le chinois, l'espagnol, le finnois, le japonais, le malgache, le polonais, le russe et le suédois [Poibeau, 2003].

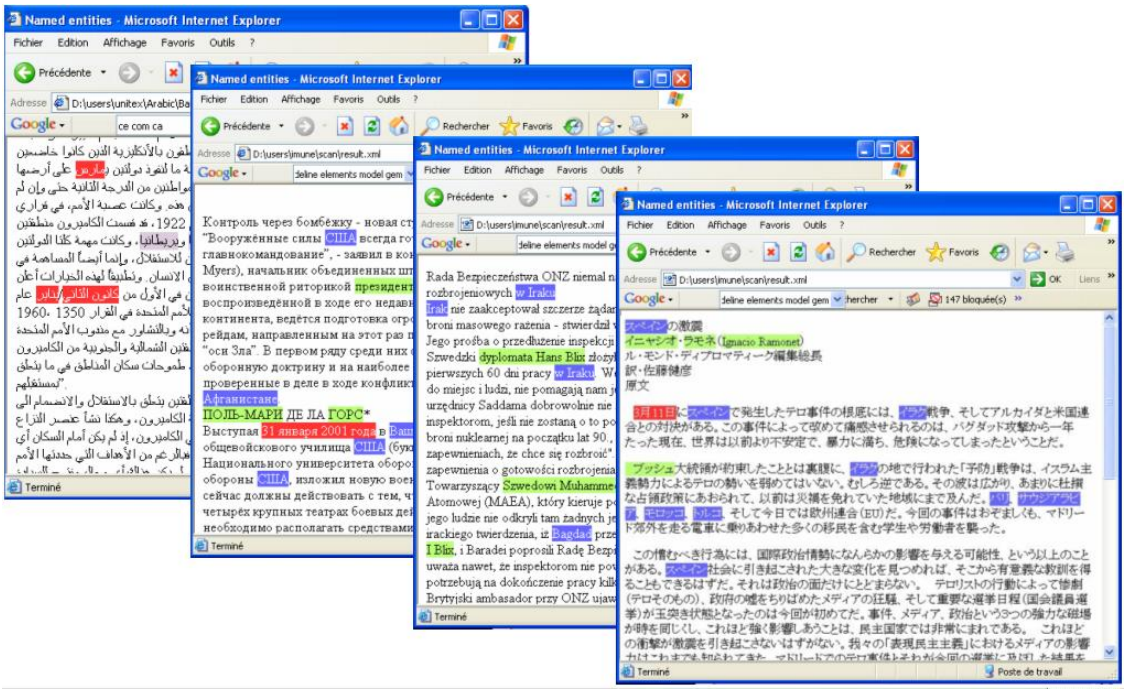

FIG. 2.4 – Modules multilingues (ici, des exemples de textes en arabe, russe, polonais et japonais)

On remarquera que certaines de ces langues ne sont pas indo-européennes. Certaines sont agglutinantes (comme le finnois), d'autres ont un jeu de caractères qui n'est pas latin (russe, chinois, japonais) ou s'écrivent de droite à gauche (arabe). Toutes ces langues peuvent être traitées avec notre architecture initiale, quelquefois en modifiant l'usage normal d'Unitex (en intégrant par exemple un analyseur morpho-syntaxique indépendant, comme Chasen pour le japonais, cf. `http://chasen.naist.jp/hiki/ChaSen/`).

2.3.4 Désambiguïsation des entités

Une fois que des entités ont été repéré, il est utile de les désambiguïser. Le contexte est parfois suffisant pour distinguer différentes occurrences, comme par exemple *Leclerc*

[33] Il s'agit essentiellement de projets de l'année 2002-2003 et 2003-2004. Certains étudiants ont depuis repris les premières implémentations, notamment pour le japonais.

dans *le général Leclerc*, *l'avenue Leclerc* ou les *centres commerciaux Leclerc*[34]. Au sein d'un même type, les systèmes de reconnaissance essaient parfois de distinguer différents « rôles » attachés à une entité, indépendamment de l'identité référentielle des éléments concernés (*Chirac* en tant que maire de Paris, président de la république ou modeste retraité ; voir [Jacquet et Ehrmann, 2006; Ehrmann, 2008] pour des résultats intéressants obtenus à partir d'une analyse distributionnelle de grands corpus).

Dans tagEN, si aucune règle ne peut s'appliquer en fonction du contexte, le mot est étiqueté avec un sens par défaut (nom de personne dans le cas de *Leclerc*). La désambiguïsation est un phénomène complexe et difficile à traiter quand il s'agit de noms du même type, réellement ambigus.

L'analyse de la position des occurrences à l'intérieur d'un document peut parfois suffire, par exemple si on peut identifier que *Chirac* reprend *Jacques Chirac* et non *Bernadette Chirac*. Plus globalement, l'analyse de grands corpus donne des pistes pour la désambiguïsation : le repérage des contextes communs et le calcul de co-occurrences autour des entités permet de les regrouper de manière pertinente.

Cette tâche a été très étudiée car elle répond à des besoins bien identifiés en recherche d'information. On peut ainsi mentionner les campagnes *Web people search* (`http://nlp.uned.es/weps/`), dont le but est de classer des pages web contenant un nom de personne ambigu en autant de « paquets » qu'il y a de personnes différentes nommées (par exemple, distinguer autant de *Mark Johnson* qu'il y a de personnes différentes désignées dans les pages web ramenées par un moteur de recherche en tapant simplement le nom comme requête). Ce type de développement n'a pas encore été intégré à TagEN.

2.4 Difficultés et limites de la catégorisation

Nous montrons dans cette partie les limites de l'approche traditionnelle, en grande partie fondée sur l'hypothèse que chaque entité a un référent stable indépendamment du contexte. Nous montrons ici la grande instabilité référentielle des entités nommées et les conséquences pour l'analyse automatique.

2.4.1 Instabilité référentielle des entités nommées en contexte

La catégorisation traditionnelle des entités repose sur une hypothèse de stabilité référentielle : elles pourraient recevoir un type rendant compte de leur référent indépendamment du contexte. Mais, contrairement à ce qui est communément admis, il est facile de montrer que le référent est instable et dépend fortement du contexte [Leroy, 2004b]. Les dates se confondent ainsi fréquemment avec des événements :

[34]Même si les systèmes ne s'attachent généralement pas à ce genre de détails, on peut au passage remarquer ici le cas *l'avenue Leclerc*, où *Leclerc* désigne bien le général ; c'est le type plus global (personne, lieu, *etc.*) qui intéresse le plus souvent les systèmes de repérage automatique.

> *Le 11 septembre 2001 a représenté un tournant dans l'histoire américaine.* (Elie Wiesel, site `www.france-amerique.com`)

Il est parfois difficile de classer les noms d'organisations, qui peuvent être catégorisés comme institution, communauté d'individus ou encore bâtiment.

> *Le journal télévisé a eu lieu hier en direct de l'ONU.*
> *L'ONU était en grève hier.*
> *L'ONU a fêté ses 50 ans.*
> *L'ONU n'acceptera pas une attaque frontale de l'Irak.* (forum du Monde)

On retrouve le même phénomène pour les noms de lieu [Lecolle, 2006] :

> *L'Europe veut garder la tête du FMI.* (Libération, 10 mars 2004)

Les noms de personnes sont encore plus labiles. Nous n'insisterons pas sur les exemples où le nom de personne est en fait devenu la désignation d'une entreprise, d'un stade ou d'un lieu quelconque.

> *Une rencontre d'un niveau technique assez médiocre à l'Abbé Deschamps.* (stade d'Auxerre, in *Journal L'équipe*)

Mais, même sans cela, les exemples de catégorisation changeante sont légion. Un nom de personne peut référer à une œuvre, à un objet ou à tout autre élément ayant un lien direct avec la personne en question [Godard et Jayez, 1993].

> *George Sand est sur le troisième rayon à partir du bas.* (cf. [Fauconnier, 1984])
> *Pierre est garé en face.* (cf. [Cadiot et Visetti, 2001])

Il existe par ailleurs des phénomènes plus rares mais bien attestés où un nom propre, particulièrement un nom de personne, ne renvoie pas au référent traditionnel. C'est par exemple le cas de l'antonomase, largement étudiée par S. Leroy [2004a] : dans *mon oncle est un vrai Harpagon*, le nom propre précédé du déterminant ne réfère bien sûr pas à un personnage réel s'appelant Harpagon (il s'agit dans ce cas d'une métaphore et non d'une métonymie dans la mesure où il n'y a pas de lien direct entre les deux éléments comparés implicitement — dans ce qui suit, je propose un moyen de traiter les cas de métonymie, qui sont les plus fréquents en corpus) [Lecolle, 1999].

On voit à travers ces exemples rapides que les entités nommées ne se comportent pas fondamentalement différemment des autres unités linguistiques. L'exemple le plus connu est certainement celui du *Prix Goncourt* introduit initialement par D. Kayser [1988]. Celui-ci distingue plusieurs sens différents, suivant qu'il s'agisse du prix, de sa valeur monétaire, du livre qui a obtenu le prix, de l'institution ou encore de l'auteur.

2.4.2 Entités et contenu sémantique

La section précédente a mis en avant plusieurs exemples difficiles à caractériser. Comment annoter les séquences en contexte, ce dernier leur faisant parfois perdre leur identifiant référentiel ? Poser cette question pousse à revenir sur la définition de la notion d'entité.

M.-N. Gary-Prieur [1994] propose d'opérer une distinction entre le sens et le contenu du nom propre[35]. Se fondant sur une terminologie empruntée à G. Kleiber [1981], Gary-Prieur reprend l'idée que le sens d'un nom propre réside dans la notion de « *prédicat de dénomination* » : le nom propre désigne une personne (ou un lieu, *etc*). Mais si le sens est « *une propriété qui caractérise le nom propre en tant qu'unité de la langue* », le contenu correspond à « *un ensemble de propriétés du référent initial qui interviennent dans l'interprétation de certains énoncés contenant ce nom* »[36].

Ainsi, pour reprendre les exemples précédents, à partir du nom d'une organisation donnée, il est possible d'avoir accès aux personnes travaillant pour cette organisation, au lieu où se situe cette organisation, *etc.* Le sens de phrases comme *Le journal télévisé a eu lieu hier en direct de l'ONU* ou *L'ONU était en grève hier* s'explique alors aisément. Il est même possible de fournir des gloses explicitant le sens des phrases visées (par exemple *Le journal télévisé a eu lieu hier en direct du siège de l'ONU*).

L'explication du phénomène est donc liée à un changement de référence. Le contexte linguistique joue donc un rôle important dans le choix des propriétés mises en évidence. Le contenu du nom propre est inséparable de son insertion dans un discours donné : le contexte permet de choisir un aspect particulier du nom propre (« *certaines propriétés en relation directe avec le contexte* », [Gary-Prieur, 1994]).

La notion de propriété oblige à revoir la thèse du nom propre vide de sens, telle qu'on l'a vue sous la plume de Mill ou de Kripke. K. Jonasson [1994] propose de rassembler les différents aspects du nom propre en remontant à un point de vue cognitif. Le nom propre permettrait d'isoler un individu au sein d'une classe (une personne particulière au sein de la classe des personnes, un lieu parmi la classe des lieux, *etc.*). Comme le dit Jonasson : « *le fondement cognitif du nom propre correspond à son association directe avec la mémoire stable à [une entité*[37]*] et non à un concept embrassant un nombre infini d'occurrences particulières* ». Elle parle aussi de « parcours interprétatif » afin de cerner le sens du nom propre en contexte.

L'analyse en terme de référence ne doit pas faire oublier que ces considérations ne sont guère productives pour l'analyse automatique. Si on veut essayer de traiter le problème, il

[35] La question de la référence des noms propres a été beaucoup étudiée, tant en philosophie qu'en linguistique. Nous ne reviendrons pas sur les très nombreux travaux en linguistique [Gary-Prieur, 1994; Jonasson, 1994; Kleiber, 1981, 1994] ou en logique [Mill, 1995 (1843); Kripke, 1982 (1972)] pour n'en citer que quelques-uns. Les ouvrages de Gary-Prieur [1994] ou de Jonasson [1994] par exemple contiennent de bons états de l'art, y compris pour les approches logiques.

[36] L'idée qu'un nom propre réfère non pas à une entité mais à l'ensemble des propriétés de cette entité est déjà présente chez Montague (voir par ailleurs les présentations de Chambreuil [1991] ou Galmiche [1991]).

[37] Jonasson emploie le terme « particulier » pour désigner une entité.

faut impérativement s'attacher à la découverte de configurations linguistiques particulières permettant d'inférer un emploi métonymique.

2.4.3 Analyse automatique de la métonymie

Le nombre important de métonymies en corpus a poussé quelques auteurs à envisager un traitement automatique du phénomène. Il s'agit en premier lieu des travaux de Nissim et Markert [2003], qui ont proposé un algorithme de résolution original qui a obtenu les meilleurs résultats pour cette tâche. Plus récemment, une campagne d'évaluation a été organisée lors de SEMEVAL 2007 (*Metonymy Resolution at SEMEVAL 2007*, `http://www.comp.leeds.ac.uk/markert/MetoSemeval2007.html`). Enfin, le corpus français ESTER [Le Meur *et al.*, 2004] permet de développer des travaux sur le français. Nous donnons ci-après un aperçu de nos travaux dans ce domaine.

Description de la tâche

Comme on l'a vu, les entités nommées, en contexte, peuvent subir un changement de référent, du fait notamment de la métonymie. On part du principe qu'il est possible de reconnaître manuellement ces cas, de les annoter et de les typer. La tâche consiste alors à développer des systèmes permettant de reconnaître automatiquement les cas de métonymie et, éventuellement, de typer de manière fine ces cas en fonction de leur nouveau référent. Les campagnes d'évaluation permettent de disposer actuellement de plusieurs jeux de données pour cette tâche.

Corpus disponibles

Il existe des corpus annotés permettant d'étudier le phénomène de la métonymie pour les entités nommées, aussi bien pour le français que l'anglais. On dispose pour les deux langues de données d'entraînement et de données de test.

Le corpus anglais est composé de 2 000 occurrences de noms de pays annotés, tirés du British National corpus (BNC) [Markert et Nissim, 2007]. L'annotation est proposée avec une granularité variable, allant d'un niveau grossier (seuls les emplois littéraux *vs* non littéraux sont considérés) jusqu'à une annotation fine, où le type de métonymie est indiqué (par exemple, `place-for-people` désigne le fait que le nom de pays est employé pour désigner ses habitants). A noter que pour SEMEVAL, les organisateurs fournissaient en outre différents types d'annotations liées au corpus (parties du discours, *etc.*).

L'annotation fine s'appuie principalement sur les types de transfert suivants : `place-for-people` (le nom de lieu désigne un ensemble de personnes), `place-for-event` (le nom de lieu désigne un événement, qu'il soit ponctuel ou périodique), `place-for-product` (le nom de lieu désigne un produit). Seul le premier cas semble vraiment productif (98 %

des cas de métonymie)[38]. Notons aussi des cas de lectures mixtes (à la fois littéral et métonymique)[39] : il s'agit généralement de cas de co-présence, où différentes parties du co-texte se réfèrent à des sens différents du même mot [Godard et Jayez, 1993; Gayral *et al.*, 2001].

Le corpus français provient de la campagne d'évaluation ESTER [Le Meur *et al.*, 2004]. Il est composé de transcriptions manuelles d'émissions radiophoniques (journaux radiodiffusés de radios françaises et marocaines). Les noms de lieux (noms de pays, de villes, *etc.*) y sont répartis en trois classes : `gsp.loc` (le référent est un nom de lieu), `gsp.pers` (le référent est un groupe de personnes), `gsp.org` (le référent est une organisation). Le corpus compte un peu plus de 1 600 éléments de type `gsp`. La distinction entre `gsp.pers` et `gsp.org` ne semble pas évidente : de fait, même pour des humains, choisir la bonne catégorie peut s'avérer problématique. Ceci pose évidemment un problème quant à l'évaluation et la réalité des distinctions supposées !

Système développé et performances

Le système développé repose essentiellement sur des heuristiques et la reconnaissance de configurations linguistiques particulières repérées à partir de corpus d'entraînement [Poibeau, 2006]. Un rapide examen d'exemples annotés montre des configurations fortement récurrentes et stéréotypées. Par exemple, l'emploi d'une entité précédée de *dans* (ou *in* en anglais) correspond quasi systématiquement à un emploi dit littéral.

Pour l'anglais, faute de temps et afin de respecter les contraintes de SEMEVAL, nous n'avons pas utilisé d'analyseur syntaxique ni de ressources autres que le corpus brut. L'idée que nous voulions explorer était d'identifier si on pouvait se passer de syntaxe et se contenter d'une analyse distributionnelle fondée uniquement sur les formes de surface. Se fondant sur aussi peu de connaissances *a priori*, le système développé a obtenu d'assez mauvais résultats [Poibeau, 2007]. L'expérience révèle, en négatif, l'importance de la syntaxe et de la sémantique pour la tâche visée ; il est par ailleurs possible d'en tirer un certain nombre d'enseignementts :

- environ la moitié des cas analysés sont fortement contraints et peuvent être étiquetés avec succès avec une stratégie aussi fruste. Il s'agit généralement d'emplois littéraux facilement reconnaissables, car précédés d'une « préposition de lieu » (plus exactement une préposition peu ambiguë introduisant généralement un complément de lieu) ; cette stratégie permet d'éliminer de nombreux cas de lecture littérale et donc d'obtenir un corpus plus équilibré entre lecture littérale et lecture métonymique.

[38] Il n'est pas toujours évident de déterminer si on a affaire à une métonymie ou non. Ainsi, on peut dire que même si des noms communs comme *poubelle* ou *camembert* sont issus de noms propres, ce lien n'est plus motivé tellement l'emploi du nom commun est dominant. Une large partie des locuteurs ignore d'ailleurs l'origine de ces noms communs.

[39] L'exemple fourni par les auteurs dans leur guide d'annotation est le suivant : *they arrived in Nigeria, hitherto a leading critic of* (*Ils arrivèrent au Nigeria, jusque là bastion critique de...*) où *Nigeria* a une valeur littérale — le nom est précédé de la préposition *in* —, les organisateurs de la tâche proposent aussi d'y voir un exemple de `place-for-people` car ce sont les habitants qui sont critiques.

- quelques (rares) cas de métonymie sont analysés avec succès grâce à l'analyse distributionnelle (notamment les emplois du domaine sportif, du type « la France a gagné la coupe du monde », marqués par un vocabulaire relativement stable et spécifique contenu dans des structures syntaxiques, elles, très variables ;
- les autres cas (qui se répartissent alors en nombre à peu près égal entre emplois métonymiques et non métonymiques) ne peuvent pas être analysés sans recours à la syntaxe.

Les résultats obtenus sont de l'ordre de 0,35 P&R pour les emplois non littéraux (38 % de précision et 31 % de rappel) et 0,84 P&R pour les emplois littéraux[40]. Ces résultats, modestes et globalement décevants, permettent malgré tout de repérer quelques éléments d'analyse simples à mettre en œuvre et fiables pour repérer facilement un grand nombre de cas d'emplois littéraux, là où d'autres systèmes (certes plus performants) utilisent d'emblée une stratégie complexe et peu lisible (e.g. [Nissim et Markert, 2003]).

Pour le français, le système est fondée sur une analyse plus complexe, faisant appel notamment à la reconnaissance de schémas syntaxico-sémantique. Les configurations linguistiques pertinentes sont analysées et codées manuellement ; on utilise pour cela le logiciel Unitex.

Schématiquement, l'analyse est fondée sur les règles suivantes (les règles 1 et 3 reprennent les observations faites sur l'anglais) :

1. reconnaissance de configurations morpho-syntaxiques simples pour les emplois littéraux (e.g. présence d'une préposition de lieu) ;
2. reconnaissance de configurations morpho-syntaxiques simples pour les emplois métonymiques (e.g. entité sujet d'un verbe de « parole », comme *dire*, *proclamer*, *répéter*, *etc.*) ;
3. reconnaissance d'autres emplois métonymiques par analyse distributionnelle (mots discriminants dans le contexte, valable notamment pour la reconnaissance de contextes sportifs) ;
4. heuristiques pour la reconnaissance d'autres contextes métonymiques (restriction des emplois métonymiques aux noms de pays et de capitale) ;
5. les autres occurrences sont marquées comme littérales.

Les performances obtenues avec ce système pour les emplois métonymiques sont 66 % de précision et 64 % de rappel. L'ordre de grandeur est comparable à celles obtenues pour l'anglais par Markert et Nissim [2007]. La comparaison est toutefois difficile dans la mesure où il ne s'agit pas de la même langue ni du même type de données (noms de pays uniquement pour l'anglais, noms de lieux variés pour le français ; dans ce dernier cas, le

[40]Du fait de la très grande disparité du nombre d'occurrences entre emplois littéraux et non littéraux (environ 80 % des occurrences pour les premiers), il y aurait une forte prime à la découverte de règles fiables, même peu productives, permettant le repérage d'emplois non littéraux. Hélas, aucune règle de ce type n'a encore été mise à jour et la stratégie par défaut, qui consiste à tout annoter comme littéral, n'est pas si facile à mettre en défaut.

fait de restreindre les emplois métonymiques aux noms de pays et de capitales permet un net gain dans les performances).

A noter qu'une équipe de Xerox a employé dans le cadre de SEMEVAL une analyse distributionnelle sur un corpus beaucoup plus large que celui fourni par les organisateurs [Brun *et al.*, 2007; Ehrmann, 2008]. Ce type d'analyse permet d'améliorer la couverture initiale obtenue par une analyse à base de patrons syntaxico-sémantiques mais nécessite un corpus important (le taux d'amélioration dépend évidemment de la grandeur des corpus utilisés ; elle est importante dans le cas de SEMEVAL, dans la mesure où le corpus d'origine était de taille modeste).

Lors de SEMEVAL 2007, d'autres équipes ont utilisé une approche à base d'apprentissage pour tout le processus et ont obtenu des performances équivalentes aux meilleurs systèmes développés manuellement [Farkas *et al.*, 2007]. Ce type d'approches nécessite des volumes de données annotées importants. On peut cependant en inférer que des régularités sont identifiables automatiquement à partir de gros corpus : il existerait donc bien des configurations linguistiques, partiellement régulières, permettant de repérer les usages métonymiques.

2.4.4 Commentaires sur les expériences

Les emplois métonymiques représentent environ 20 % des emplois, aussi bien pour les corpus français que pour les corpus anglais étudiés. Il ne s'agit donc pas d'un emploi majoritaire, mais d'un phénomène fréquent qui ne pose en soi aucun problème d'interprétation au lecteur (il s'agit en quelque sorte d'un phénomène "transparent" dont le lecteur n'a pas conscience à la lecture du texte).

En revanche, la reconnaissance automatique de ce type d'emploi est difficile. Les résultats rapportés ci-dessus en témoignent : les systèmes peinent à aller au-delà de 60 à 70 % d'emplois métonymiques reconnus comme tels dans les textes. En fait, au-delà de certaines configurations plus ou moins aisément reconnaissables, les systèmes échouent à reconnaître les lectures métonymiques non stéréotypées et fortement variables, donc difficilement prédictibles *a priori* par introspection (c'est-à-dire sans étude sur corpus). Aucune description complète et satisfaisante du phénomène en jeu n'a encore pu être donnée d'un point de vue linguistique. Les systèmes de génération d'usages figurés sont encore largement déficients, même si on voit apparaître ici et là des tentatives intéressantes [Veale *et al.*, 2006].

L'analyse fine de la métonymie (que nous n'avons pas abordée) est encore plus difficile, dans la mesure où les emplois types (noms de lieu pour désigner un ensemble d'habitants par exemple) sont écrasants (le type `place-for-people` représente ainsi plus de 98 % des emplois métonymiques de noms de lieu dans le corpus SEMEVAL). Aucun système n'a obtenu de performances satisfaisantes pour cette tâche lors de SEMEVAL 2007 ; les volumes de données fournis étaient de toute manière trop faibles pour obtenir une évaluation fiable et les dictinctions trop fines semblent souvent discutables.

Les derniers commentaires que l'on fera portent d'ailleurs sur la difficulté — ou plutôt l'ambiguïté — de la tâche, même pour un humain. Certains cas peuvent être facilement identifiés car ils correspondent à des schémas syntaxico-sémantiques stéréotypés comme on l'a déjà vu mais d'autres posent de nombreux problèmes lors de l'annotation. Il existe en effet des cas mixtes, où plusieurs lectures co-existent comme dans l'exemple de *Nigeria* décrit ci-dessus.

Le point qui nous semble le plus problématique est celui de l'annotation fine. Les types proposés reposent souvent sur une analyse référentielle visant à reconstituer du sens, là où la figure de style (la métonymie) permet au contraire une prise de distance par rapport à la référence. Certains aspects (que l'on peut chercher à caractériser en parlant de focalisation, de propriété ou de trait sémantique) de l'entité sont accessibles mais celle-ci ne perd pas sa caractérisation initiale. Il nous semble que c'est pour cela que l'on peut avoir affaire à des lectures mixtes, parfois floues et peu caractérisables en termes de types sémantiques.

2.5 Perspectives

L'analyse ci-dessus a montré les limites d'un traitement trop précis sur le plan référentiel des entités nommées. Un traitement des phénomènes de glissement de sens sur le plan linguistique (et plus seulement référentiel) semble donc plus approprié mais aucune théorie ne donne d'explication réellement exploitable d'un point de vue automatique. Au-delà de quelques cas isolés, il n'est d'ailleurs pas certain que le traitement des emplois métonymiques soit réellement utile d'un point de vue applicatif.

Plutôt que de persévérer dans cette voie, il nous a semblé plus approprié de voir en quoi les entités nommées étaient révélatrices du contenu même des documents. Ainsi, des textes contenant les mêmes entités ont des chances de parler du même événement, surtout s'il s'agit de dépêches d'agence. C'est le sujet de la thèse d'Aurélien Bossard [2008], qui essaie de voir comment exploiter ces éléments d'information relativement isolés, pour des tâches nécessitant normalement davantage de connaissances, comme la catégorisation de textes ou le résumé multi-documents. Dans ce cadre, le poids des différentes entités, leur répartition et leur distribution sont des éléments fondamentaux qui doivent être modélisés de façon extrêmement précise.

Cette thèse montre aussi l'imbrication des niveaux linguistiques (cf. chapitre 1, page 28). Les entités nommées, au niveau local, permettent de donner une idée des liens entre documents en dessinant un réseau de relations complexes, au niveau global. Le réseau peut à son tour donner lieu à des représentations variées (graphes, cartes) et offrir des moyens originaux d'accès à l'information. Si nous n'avons guère exploré les outils de cartographie jusqu'à présent, il s'agit d'une perspective intéressante et offrant des débouchés très directs à ces travaux.

2.6 Synthèse

J'ai présenté dans ce chapitre mes travaux concernant l'analyse des entités nommées. Comme on a pu le voir, ces éléments tiennent une place prépondérante dans nombre d'applications de compréhension de textes. J'ai détaillé le système TagEN, développé au LIPN, qui permet une analyse robuste et relativement performante pour la tâche visée. Dans un second temps, j'ai montré les limites liées à la tâche elle-même quand on doit faire face à des glissements de sens. Les systèmes sont alors relativement inopérants[41], face à des phénomènes encore mal maîtrisés et peu formalisés au niveau linguistique.

[41] On peut aussi penser que c'est la tâche qui est mal définie, dans la mesure où, comme nous l'avons souligné, certaines distinctions sont difficiles à faire même par un humain. Il n'empêche qu'il y a là un problème de fond, dans la mesure où l'on perçoit des nuances de sens sans arriver à les décrire de façon suffisamment fine et précise.

Chapitre 3

Rôles sémantiques et relations entre entités

Le repérage des entités au sein d'un texte est souvent une première étape d'analyse avant des traitements plus complexes. Ainsi, l'extraction d'information ou les systèmes de questions-réponses exigent de mettre en relation les entités, afin d'identifier leur rôle par rapport à un événement donné. L'analyse des relations sémantiques nécessite d'explorer deux axes complémentaires :

- Un axe « paradigmatique » : la sémantique de la relation est généralement « portée » par un prédicat, correspondant à une famille lexicale plus ou moins large (par exemple, pour exprimer la notion d'*achat*, on est bien évidemment intéressé par des mots ou des expressions tels que *acquérir*, *acheter*, *faire l'acquisition de*, *etc.* L'analyse exige donc de repérer ces éléments lexicaux dont le sens est jugé équivalent dans le cadre d'une application donnée.
- Un axe « syntagmatique » : la sémantique des arguments dépend bien évidemment de leur rôle par rapport au prédicat. Il faut donc mettre en œuvre une analyse syntaxiqueafin de découvrir les relations entre mots au sein de la phrase (par exemple, identifier l'*acheteur*, le *vendeur*, l'*objet* d'une transaction exige de repérer la fonction syntaxique de ces éléments par rapport au prédicat).

En combinant ces deux axes, on cherche à repérer des paraphrases, c'est-à-dire des phrases ou des structures prédicatives exprimant une même relation entre des éléments identiques. Il existe une littérature abondante sur la paraphrase en linguistique [Fuchs, 1994] mais nous en restons ici à un point de vue essentiellement applicatif.

J'ai initialement abordé ces questions lors de ma thèse, dans le cadre d'une application d'extraction d'information. L'analyse des schémas prédicatifs était relativement simplifiée, dans la mesure où il s'agissait d'une tâche analogue à de l'étiquetage sémantique, ensuite filtré par un jeu de contraintes syntaxiques locales (pour déterminer la nature du lien entre le complément et la tête du syntagme considéré) [Poibeau, 2003]. J'ai depuis complété ce travail en collaboration avec Cédric Messiant, à travers une exploration de l'interface

syntaxe-sémantique pour acquérir semi-automatiquement des comportements syntaxico-sémantiques particuliers à partir de gros corpus.

3.1 Sur la notion de prédicat

Il est important de définir d'abord ce que l'on entend par prédicat. Il s'agit d'une notion bien connue et très étudiée en linguistique mais qui n'est pas sans soulever des questions, sur sa nature, sur son lien à la logique ou encore sur ses rapports avec l'analyse linguistique traditionnelle.

3.1.1 Considérations générales

On s'inspire ici des travaux déjà anciens de Tesnière [1959], qui nous semblent bien adaptés à nos besoin. Tesnière a notamment proposé de représenter les relations de dépendance à travers la notion de *stemma*.

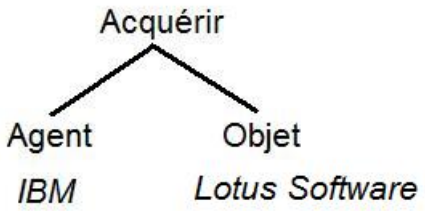

FIG. 3.1 – Un stemma *à la* Tesnière, correspondant à la phrase *IBM a acquis Lotus Software*.

La figure 3.1 montre les relations entre le prédicat verbal (*acquérir*) et les deux éléments régis : *IBM* et *Lotus Software*.

Un prédicat correspond donc à un stemma où un noyau supérieur régit différents éléments. Le nombre des éléments régis définit la *valence* du prédicat (la valence est de 2 sur la figure 3.1 dans la mesure où deux éléments sont régis par le verbe). Tesnière inclut le sujet dans la valence du verbe, contrairement à l'analyse traditionnelle qui isole le sujet du groupe verbal, ou qui s'appuie sur une partition thème-rhème. On adopte ici la représentation de Tesnière : que ce soit pour un système d'extraction d'information ou de questions-réponses, la distinction n'a pas lieu d'être dans la mesure où l'on cherche avant tout à caractériser la sémantique des différents éléments régis, qu'ils occupent la position de sujet ou non.

Les éléments régis sont appelés *actants*. Un actant est défini par Tesnière comme étant un « être ou [une] chose qui, à un titre quelconque, et de quelque façon que ce soit, même

au titre de simple figurant et de la façon la plus passive, participe au procès ». Il existe ainsi des schémas mono-actanciels, bi-actanciels (par ex., *IBM a racheté Lotus Software* où les deux actants sont *IBM* et *Lotus Software*) ou tri-actanciels (par ex., *Lenovo a racheté à IBM sa division PC* où les trois actants sont *Lenovo*, *IBM* et *sa division PC*). Déterminer la nature et la fonction des actants est un problème difficile : il n'existe pas une liste de fonctions stable et faisant autorité [Mel'cuk, 2004; Levin et Rappaport Hovav, 2005]. Nous nous intéressons ici plutôt à des rôles sémantiques, c'est-à-dire à une description très précise de la fonction dépendant étroitement de l'application visée.

Les actants sont distingués des circonstants, qui ne participent pas de manière directe au procès. On sait que la distinction entre actant et circonstant est un problème difficile (par exemple dans *IBM a racheté Lotus Software pour 3,5 milliards de dollars*, la somme d'argent est *a priori*, dans la grammaire traditionnelle, un circonstant mais ceci est discutable d'un point de vue applicatif). Nous aurons l'occasion d'y revenir.

Le principal avantage des schémas prédicatifs ainsi définis au moyen des notions de valence et de rôle sémantique réside dans le fait qu'on peut représenter les relations de dépendance indépendamment des réalisations syntaxiques de surface. Il est même possible de définir des familles de prédicats (familles sémantiques) qui partagent la même structure actancielle. La théorie de Tesnière intègre donc la notion intuitive de paraphrase : la théorie a donc une dimension sémantique, par delà l'aspect formel. Cela a été critiqué par certains syntacticiens "purs" (rejetant la sémantique à l'extérieur de leur théorie) mais ne nous dérangera bien évidemment pas pour notre travail. En revanche, nous sommes quant à nous directement confronté aux réalisations de surface qu'il va nous falloir essayer de dépasser pour les généraliser.

Les travaux de Tesnière préfigurent les grammaires de cas de C. Fillmore [1968]. Ce dernier a depuis montré la limite de la notion de cas et propose aujourd'hui, à travers FrameNet, une ressource beaucoup plus riche, où chaque argument du prédicat reçoit un rôle sémantique précis [Fillmore, 1982]. En l'absence de ressources de ce type pour le français, il semble nécessaire de voir la meilleure façon d'acquérir des connaissances lexicales semi-automatiquement, à partir de corpus.

3.1.2 Stratégie d'acquisition automatique à partir de corpus

L'acquisition de schémas prédicatifs vise à produire une base de connaissances, contenant un ensemble de prédicats (noms ou verbes exprimant une notion donnée, comme la notion d'achat) et leurs dépendants possibles.

J'ai proposé, d'abord à travers ma thèse puis à travers différents travaux effectués depuis que je suis au CNRS, un ensemble de méthodes permettant l'acquisition de schémas prédicatifs. Au niveau syntaxique, la méthode proposée est purement endogène et dynamique : c'est l'usage des unités en corpus qui permet d'en inférer le comportement. On travaille sur de gros corpus afin d'acquérir des connaissances « larges » sur la langue. Au niveau sémantique, pour le typage des arguments, il a semblé plus intéressant de considé-

rer des restrictions de sélection précises dans le cadre d'applications données, plutôt que d'attribuer des rôles sémantiques trop généraux aux unités considérées.

L'acquisition de familles de comportement syntaxique repose sur l'analyse de corpus bruts mais, selon la démarche de Harris (cf. chapitre 1, page 20), il est nécessaire de normaliser préalablement les données afin d'échapper au moins partiellement aux variations de surface. Les compléments du verbe sont ensuite extraits et classés suivant la force de leur lien avec le verbe, représentée par une distribution de probabilités. Je détaille ces aspects dans la section suivante.

Pour l'analyse des restrictions de sélection, l'utilisateur doit disposer d'un corpus d'entraînement puis fournir un ou des exemples représentatifs ; ceux-ci sont ensuite généralisés de manière interactive. Le processus est donc dirigé vers un besoin particulier et n'a pas vocation à produire une ressource plus générale (voir section 3.3).

3.2 Acquisition de cadres de sous-catégorisation

De nombreux travaux ont été réalisés ces dernières années autour de la notion de cadre de sous-catégorisation. Un cadre de sous-catégorisation inclut des informations sur le nombre et la nature des arguments, y compris sur un plan sémantique (on parle aussi de cadre de valence, cf. [Tesnière, 1959; Iordanskaja et Mel'cuk, 2000; Mel'cuk, 2004]). Déterminer le nombre d'arguments d'un prédicat exige une analyse fine des compléments, afin de distinguer les arguments des modifieurs.

3.2.1 Positionnement

La modélisation des cadres de sous-catégorisation a longtemps été une tâche manuelle [Gross, 1975], à l'heure où les ordinateurs restaient des moyens d'expérimentation relativement peu accessibles[42]. Le Laboratoire d'Automatique Linguistique et Documentaire (LADL) de M. Gross a ainsi abrité de la fin des années 1960 à la fin des années 1990 des équipes de linguistes chargés de mettre au point des dictionnaires électroniques pour le français. Cependant, la définition manuelle de lexiques riches est une tâche coûteuse et difficile, sujette à erreurs et à incohérences. Elle ne permet pas l'adaptation rapide des ressources à un nouveau domaine ou à une tâche ciblée.

Une autre approche consiste à acquérir les lexiques et notamment les informations de sous-catégorisation à partir d'un corpus arboré [O'Donovan *et al.*, 2005; Kupść, 2007], qu'il faut élaborer au préalable. Cette contrainte limite beaucoup l'intérêt de la méthode,

[42] Notons cependant qu'au même moment, Z. Harris et N. Sager exploraient des techniques automatiques — notamment l'analyse distributionnelle — pour identifier des contextes pertinents et des comportements syntaxiques caractéristiques [Sager *et al.*, 1987; Harris, 1988; Nevin et Johnson, 2002]. Le fait de travailler en grande partie par introspection est un choix revendiqué de M. Gross, qui ne se différencie pas de Chomsky sur ce point ("*Maurice Gross advocated a subjective method with a collective control : empirical observations are performed through introspection by a team of native speaker linguists. This is how the Lexicon-Grammar tables of French verbs were constructed*". [Laporte, 2005]).

dans la mesure où de tels corpus sont peu disponibles et fort coûteux à constituer. De plus, cette approche ne permet pas d'adapter facilement la ressource à un nouveau corpus et entraîne souvent un problème de rappel, dans la mesure où les corpus arborés sont encore de taille modeste.

Pour pallier les défauts des approches précédentes, des travaux visant l'acquisition automatique de cadres de sous-catégorisation à partir de corpus bruts (ou annotés automatiquement) ont été menés dès les années 1990 [Manning, 1993; Brent, 1993; Briscoe et Carroll, 1997]. La plupart de ces travaux portent sur l'anglais mais il existe aussi des travaux sur d'autres langues comme, par exemple, l'allemand [Schulte im Walde, 2002]. Jusqu'à récemment, aucune expérience de ce type n'avait été menée sur le français, probablement en raison de l'absence de larges corpus et d'analyseurs syntaxiques performants ; la première tentative est probablement celle de Chesley et Salmon-Alt [2006], qui ne porte que sur une centaine de verbes. Plusieurs tentatives sont actuellement en cours pour acquérir de manière manuelle ou semi-automatique un lexique à large couverture, contenant des informations de sous-catégorisation pour le français (voir par exemple, le *Lefff* [Sagot *et al.*, 2006]).

Avec Cédric Messiant, nous avons développé un système d'acquisition de cadres de sous-catégorisation à partir d'un corpus brut analysé automatiquement [Messiant *et al.*, 2008]. L'architecture de notre système s'inspire des travaux menés à Cambridge pour l'anglais [Briscoe et Carroll, 1997][43]. Les travaux sur l'anglais ont en effet montré qu'il était possible d'acquérir des lexiques à large couverture (le lexique *VALEX* contient 6300 verbes anglais [Korhonen *et al.*, 2006a]), couvrant différentes catégories morpho-syntaxiques (verbes, noms, adjectifs [Preiss *et al.*, 2007]) et utilisables dans diverses applications (acquisition de classes sémantiques [Korhonen et Briscoe, 2004]).

Toutefois, la situation pour le français n'est pas la même que pour l'anglais : beaucoup de données et d'outils sont disponibles pour l'anglais. L'équipe de Cambridge suppose *a priori* connue la liste des comportements syntaxiques possibles des unités visées. La tâche consiste alors à essayer de repérer, parmi une liste de cadres possible, celui qui correspond au comportement d'un item lexical en contexte. A l'inverse, nous partons du point de vue qu'il n'y a pas de raison de considérer comme acquis *a priori* l'ensemble des comportements syntaxiques possibles. Nous proposons une démarche inductive qui vise à construire dynamiquement les cadres à partir des observations en corpus.

L'approche est donc constructionnelle et repose essentiellement sur une hypothèse statistique, à savoir que les arguments sont essentiels à la signification, donc généralement présents autour du verbe, alors que les modifieurs sont plus variables, donc statistiquement moins significatifs. Cette approche ne permet cependant pas directement de retrouver les distinctions traditionnelles qui sont fondées sur des critères linguistiques et non statistiques. A l'issue de ce type de démarche, il sera donc important d'examiner l'écart constaté par rapport à la grammaire du français classique et voir si notre approche permet de décrire les unités linguistiques sous un autre jour.

[43]Ces recherches s'insèrent dans le cadre d'un projet entre l'Université de Paris-Nord et l'Université de Cambridge (PHC Alliance 2007-2009).

Notons toutefois que les techniques statistiques opèrent des distinctions entre compléments sur la base de la fréquence : il reste donc à voir dans quelle mesure ces distinctions recoupent celles de la grammaire traditionnelle entre arguments et modifieurs.

Dans ce qui suit, nous présentons une expérience faite à partir d'un corpus journalistique (10 ans du journal *Le Monde*), mais des expériences sur d'autres corpus, notamment technique, contribuent à montrer l'intérêt de la méthode pour acquérir des comportements syntaxiques variés en fonction du domaine considéré. Nous revenons dans la section 3.2.2 sur quelques questions théoriques liées à la notion de sous-catégorisation, en particulier la distinction argument/modifieur. La section 3.2.3 décrit ensuite le fonctionnement du système d'acquisition développé et de ses différents modules. L'expérience d'acquisition du lexique de sous-catégorisation et son évaluation sont traitées dans la section 3.2.4. Enfin, nous situons notre travail par rapport aux travaux existant dans le domaine dans la section 3.2.5.

3.2.2 La sous-catégorisation, une notion floue

Les cadres de sous-catégorisation d'un prédicat décrivent l'ensemble des constructions syntaxiques possibles pour ce prédicat. La plupart des travaux existants sur la sous-catégorisation distinguent les arguments, qui doivent faire partie des cadres, et les circonstants (appelés aussi modifieurs), qui doivent en être exclus. Les approches fondées sur une acquisition automatique ont tendance à « éclater » les cadres, dans la mesure où les arguments optionnels contribuent à identifier des structures de surface distinctes (on pourra ainsi trouver un cadre intransitif pour un verbe transitif dont le complément est optionnel). Un travail de restructuration des cadres obtenus est nécessaire pour obtenir un résultat plus compact et plus proche de l'analyse traditionnelle[44].

Les linguistes ont longtemps étudié ce problème et proposé des batteries de tests permettant de faire la différence entre les deux types de compléments (ainsi, les modifieurs pourraient plus facilement être effacés, déplacés, *etc.*), sans qu'aucun critère réellement discriminant ne puisse être finalement défini [Levin et Rappaport Hovav, 2005]. Les critères proposés ne sont par ailleurs pas toujours faciles à mettre en œuvre dans le cadre d'une analyse automatique. Cette distinction est pourtant fondamentale lorsqu'on traite de sous-catégorisation verbale. Christopher Manning synthétise de façon assez éclairante cette question [Manning, 2003] :

> *There are some very clear arguments (normally, subjects and objects), and some very clear adjuncts (of time and 'outer' location), but also a lot of stuff in the middle. Things in this middle ground are often classified back and forth as arguments or adjuncts depending on the theoretical needs and convenience of the author.*

[44]Notons toutefois que cette étape de restructuration n'est pas triviale dans la mesure où des verbes homonymes comme *voler* se distinguent par leurs schémas de sous-catégorisation respectifs : « *l'oiseau vole* » est intransitif, contrairement à « *Perugia a volé la Joconde* ». Une étape de désambiguïsation sémantique serait nécessaire pour distinguer ce type de cas [Korhonen *et al.*, 2003].

Il existerait donc plutôt un continuum entre arguments et modifieurs, ce qui permet de mieux représenter le caractère plus ou moins obligatoire des « arguments » des verbes (voir aussi [Lazard, 1994, p. 80], [Pottier, 1992, p. 124–127], [Rastier, 1998]). Il faut désormais trouver un moyen de représenter ce continuum en termes de cadres de sous-catégorisation. La réponse apportée par Manning est de représenter la sous-catégorisation comme une distribution de probabilité sur les cadres :

> *Rather than maintaining a categorical argument / adjunct distinction and having to make in/out decisions about such cases, we might instead try to represent subcategorization information as a probability distribution over argument frames, with different verbal dependents expected to occur with a verb with a certain probability.*

Les lexiques acquis à partir de corpus s'enrichissent ainsi de probabilités. A chaque construction de chaque verbe (ou locution verbale) est attachée une probabilité qui peut aider, par exemple, à guider un analyseur syntaxique probabiliste à évaluer des hypothèses concurrentes (l'ensemble des cadres d'un verbe donné a une probabilité égale à 1). La distribution de probabilités peut en outre permettre de repérer des ensembles d'unités lexicales ayant le même comportement (c'est-à-dire ayant une distribution de cadres proche). On peut dès lors chercher à vérifier si les comportements syntaxiques correspondent à des familles sémantiques homogènes, suivant en cela une hypothèse posée par B. Levin [1993].

L'information contenue dans ces probabilités est évidemment plus riche que la distinction binaire argument/circonstant, ce qui peut poser problème lors de l'évaluation. Dans la plupart des travaux précédents [Chesley et Salmon-Alt, 2006; Preiss *et al.*, 2007], les cadres de sous-catégorisation sont comparés à un « *gold standard* », autrement dit à une ressource lexicographique de référence, définie manuellement, souvent issue d'une source papier et sans indication probabiliste. Les deux jeux de données sont donc de nature différente et la question de la pertinence de cette méthode d'évaluation mérite d'être posée. Pour comparer les sorties probabilistes du système d'acquisition automatique avec les données binaires[45] issues d'un « *gold standard* », il faut donc consentir à une perte d'information sur les probabilités obtenues. D'autres méthodes d'évaluation sont envisageables, comme, par exemple, le calcul de l'apport de l'utilisation du lexique dans une application, notamment dans un analyseur syntaxique [Poibeau et Messiant, 2008]. Nous y reviendrons.

3.2.3 Le système d'acquisition de cadres de sous-catégorisation

LexSchem est un lexique acquis automatiquement à partir de corpus brut grâce à un système développé par Cédric Messiant, appelé ASSCI. Le corpus est d'abord étiqueté et analysé syntaxiquement par l'analyseur Syntex [Bourigault *et al.*, 2005]. Notre système

[45] Binaire au sens où, dans un « *gold standard* » traditionnel, un cadre est valide ou non. Une construction syntaxique peut le cas échéant être définie comme rare mais les analyseurs syntaxiques ne sauront que faire de cette indication si elle n'est pas « probabilisée ».

d'acquisition automatique de cadres de sous-catégorisation est composé de trois modules, exécutés en série sur les sorties de l'analyseur syntaxique [Messiant *et al.*, 2008] :

1. Un module d'extraction des verbes et de leurs dépendances ;
2. Un module de construction des cadres candidats à partir des informations extraites par le premier module ;
3. Un module de filtrage des cadres qui rejette les candidats non pertinents en se basant sur leur fréquence dans le corpus.

Traitements préliminaires

Le corpus choisi pour l'acquisition des cadres de sous-catégorisation doit d'abord être analysé par Syntex, analyseur syntaxique développé à Toulouse par Didier Bourigault [Bourigault *et al.*, 2005]. La lemmatisation et l'étiquetage morphosyntaxique sont préalablement réalisés par le TreeTagger[46].

Nous avons choisi Syntex car il s'agit d'un analyseur non lexicalisé[47], procédural à cascade. L'analyseur a obtenu de bonnes performances sur des textes variés lors de la campagne *EASY*[48]. Comme le dit D. Bourigault, Syntex effectue une analyse en dépendances et ne se base sur aucune théorie particulière. Les catégories morphosyntaxiques et les relations syntaxiques s'appuient sur « la grammaire traditionnelle ». Syntex étiquette chaque mot de la phrase analysée avec son recteur (gouverneur syntaxique dont le mot dépend) et ses régis (dépendances). Cependant, l'analyseur ne fait pas de différence entre les arguments et les modifieurs des verbes : Syntex lie tous les syntagmes propositionnels au verbe (cf. la figure 3.2).

Le module d'extraction

Le module d'extraction prend en entrée les sorties de l'analyseur Syntex (figure 3.2). Il y repère les verbes qu'on a choisi de traiter (typiquement, tous les verbes dont le nombre d'occurrences est suffisamment élevé pour obtenir des résultats fiables) et il en extrait les dépendances. A ce niveau, les informations extraites comprennent le lemme du verbe et ses régis (ses arguments potentiels, sujet exclu). Pour chacun de ces régis, on stocke la nature de sa relation de dépendance au verbe (par exemple, objet, attribut du sujet ou syntagme prépositionnel) et sa catégorie morphosyntaxique[49].

Certaines erreurs récurrentes (repérées manuellement) de TreeTagger ou de Syntex sont traitées par ce module. Par exemple, dans le syntagme « *Le programme d'armement* », le

[46] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

[47] Mises à part quelques rares exceptions, comme les verbes à contrôle, qui sont gérés au sein d'une liste ; il existe une dizaine de telles listes, toutes composées de moins d'une quinzaine d'éléments.

[48] `http://www.limsi.fr/Recherche/CORVAL/easy/`. Les scores et les rangs de Syntex pour cette campagne sont disponibles à `http://w3.univ-tlse2.fr/erss/textes/pagespersos/bourigault/syntex.html\#easy`

[49] Pour le détail des différentes relations de dépendance et des catégories morphosyntaxiques repérées par Syntex, voir `http://www-etud.iro.umontreal.ca/~demorali/syntex/syntex_format.html`

Le chasseur de jadis attrapait les oiseaux au filet avant de les engraisser.

```
DetMS|le|Le|1|DET;2|
NomMS|chasseur|chasseur|2|SUJ;5|DET;1,PREP;3
Prep|de|de|3|PREP;2|
Adv|jadis|jadis|4||
VCONJS|attraper|attrapait|5||SUJ;2,OBJ;7,PREP;8,PREP;10
DetMP|le|les|6|DET;7|
NomMP|oiseau|oiseaux|7|OBJ;5|DET;6
Prep|à|au|8|PREP;5|NOMPREP;9
NomMS|filet|filet|9|NOMPREP;8|
Prep|avant de|avant de|10|PREP;5|NOMPREP;12
Pro|le|les|11|OBJ;12|
VINF|engraisser|engraisser|12|NOMPREP;10|OBJ;11
Typo|.|.|13||
```

FIG. 3.2 – Analyse par Syntex d'une phrase exemple (la 5e ligne générée par Syntex correspond au verbe *attraper*, qui gouverne quatre éléments : le sujet, l'objet et deux compléments prépositionnels ; chaque élément est indexé par sa position relative ; chaque relation est nommée (`SUBJ`, `OBJ`...) et pointe vers l'objet cible référencé par son index).

mot *programme* est étiqueté `Verbe` par le TreeTagger, ce qui provoque des erreurs dans les analyses de Syntex qui se répercutent dans nos cadres.

Ce module a pour objectif de rassembler les informations qui permettront au module suivant de construire les cadres de sous-catégorisation candidats pour chaque verbe.

Le module de construction

Le module de construction prend les sorties du module précédent (figure 3.2) et construit les cadres de sous-catégorisation candidats. Pour chaque syntagme dépendant du verbe, son étiquette syntaxique (celle qui sera utilisée dans les cadres) est déterminée sur la base de sa relation avec le verbe et de la catégorie morphosyntaxique de sa tête (voir figure 3.3).

Le chasseur de jadis attrapait les oiseaux au filet avant de les engraisser .

```
SN_SP[avant de]_SP[à]
```

FIG. 3.3 – Cadre candidat construit à partir de la phrase exemple (dans cet exemple, le verbe *attraper* aurait comme compléments un complément d'objet direct (`SN`), suivi de deux complément indirect (`SP`), le premier introduit par la locution *avant de*, le second par la préposition *à*. Chaque complément est séparé du précédent par le caractère _. Il s'agit bien évidemment d'une analyse erronée à ce stade, avant la phase de filtrage.

Des règles de construction des catégories syntaxiques doivent être écrites, ce qui implique des choix linguistiques sur la profondeur des étiquettes à attribuer aux dépendances. Nous avons finalement choisi six étiquettes :

- SN : syntagme nominal ;
- SINF : subordonnée infinitive ;
- SP[*prep*+SN] : syntagme prépositionnel où *prep* est la préposition "tête" ;
- SP[*prep*+SINF] : subordonnée infinitive introduite par une préposition ;
- SA : syntagme adjectival ;
- COMPL : subordonnée ;

Ces étiquettes et les choix qui en découlent sont largement dépendants de l'application dans laquelle les cadres seront utilisés. Le module peut être facilement adapté aux besoins des utilisateurs.

Le module de filtrage (élimination des cadres erronés)

Les cadres candidats obtenus par le module de construction doivent ensuite être filtrés. En effet, à ce stade, il reste beaucoup de cadres erronés qu'il faut éliminer. Ces erreurs sont souvent dûes à des erreurs du taggeur ou de l'analyseur syntaxique.

A. Korhonen a montré qu'on obtenait les meilleurs résultats en utilisant un simple seuil sur les fréquences relatives des cadres candidats [Korhonen *et al.*, 2000]. La fréquence relative d'un cadre pour un verbe donné est calculée en divisant le nombre d'occurrences de ce cadre pour le verbe en question par le nombre total d'occurrences de ce verbe.

$$rel\_freq(scf_i, verb_j) = \frac{|scf_i \cap verb_j|}{|verb_j|}$$

Le seuil auquel est comparé ce ratio est fixé empiriquement. Nous avons choisi d'utiliser des seuils différents en fonction des situations. Par exemple, un seuil plus élevé est utilisé pour la forme intransitive des verbes, du fait que, dans certains cas difficiles (par exemple en cas d'incises), Syntex ne prend pas de décision et n'attache rien au verbe.

Lorsqu'un cadre est rejeté par le module du fait d'une probabilité trop faible (inférieure au seuil fixé), une procédure vise à le faire correspondre automatiquement à un autre cadre. Pratiquement, le système essaie de le « réduire », c'est-à-dire de supprimer un syntagme prépositionnel du cadre initial, afin de le renvoyer vers un cadre existant ou de créer un nouveau cadre. Les fréquences relatives des cadres candidats sont alors calculées à nouveau et un deuxième filtrage est effectué (le seuil utilisé pour ce deuxième filtrage est plus faible). Cette méthode permet de filtrer les cadres trop peu fréquents (souvent erronés) sans perdre toute l'information contenue dans ces cadres.

Ainsi, pour le verbe *attraper*, le cadre candidat `SN_SP[avant de]_SP[à]` est comparé au seuil. Sa fréquence relative est inférieure au seuil, il est donc éliminé. Ce cadre contient des syntagmes prépositionnels, l'algorithme va donc le réduire. Les deux cadres suivants sont alors possibles :

```
SN_SP[avant de]
SN_SP[à]
```

L'algorithme cherche laquelle des deux prépositions apparaît le plus fréquemment dans le contexte du verbe ; il s'agit ici du cadre avec le complément introduit par *à*. Le cadre retenu est alors `SN_SP[à]`. La fréquence relative de ce cadre sera calculée puis comparée au seuil. Ici, le cadre ne sera pas éliminé donc `SN_SP[à]` est le cadre de sous-catégorisation correspondant au verbe *attraper* dans la phrase exemple. On note que le cadre ainsi obtenu conserve la notion d'instrument à l'intérieur de la sous-catégorisation.

3.2.4 Expérience

Le système décrit dans la section précédente a été utilisé pour extraire un lexique de cadres de sous-catégorisation à partir d'un corpus journalistique. Nous avons évalué ce lexique en le comparant avec un « *gold standard* », malgré les réserves que nous avons formulées sur ce type d'évaluation (voir section 3.2.2). Des expériences sont en cours pour permettre d'évaluer également l'utilité des informations statistiques obtenues à partir du corpus.

Corpus utilisé

Le corpus choisi pour l'acquisition du lexique LexSchem est un corpus composé des articles du quotidien *Le Monde* sur 10 ans (1991-2000) acquis auprès de l'agence ELRA[50]. Ce choix comporte un double avantage : il s'agit du plus gros corpus pour le français (200 millions de mots) qui soit suffisamment « propre » pour limiter les erreurs d'analyse (Syntex a obtenu sur ce type de corpus des scores de précision et de rappel respectivement de 0,76 et 0,58 à la campagne EASY) et il s'agit d'un corpus journalistique, qui est donc relativement hétérogène (il traite aussi bien de sport que de politique, d'économie que de vie quotidienne). Nous avons limité l'extraction de cadres de sous-catégorisation aux 3 268 verbes ayant plus de 100 occurences dans le corpus.

Le lexique LEXSCHEM

LexSchem est accessible au format XML à l'adresse suivante : `http://www-lipn.univ\-paris13.fr/~messiant/lexschem/lexschem.xml`. Le lexique comprend une entrée par verbe (balise `<verb lemma="">`, le lemme du verbe correspond à la valeur de l'attribut `lemma`). Chaque verbe a un ou plusieurs cadres de sous-catégorisation (balise `<scf>`). Chaque cadre de sous-catégorisation est composé d'un identifiant (balise `<id>`), de son « patron » (balise `<pattern>`), de sa fréquence relative (balise `<relfreq>`), de son nombre d'occurrences dans le corpus d'acquisition (balise `<freqcnt>`) et de cinq exemples (balises `<example>`).

[50] La version du corpus utilisée avait été nettoyée par B. Habert (alors au LIMSI) et l'analyse par Syntex fournie par D. Bourigault (alors à l'ERSS).

Le lexique est également consultable via une interface web à l'URL suivante : `http://www-lipn.univ-paris13.fr/~messiant/lexschem`. Elle permet de consulter, pour chaque verbe, la liste de ses cadres de sous-catégorisation ainsi que les probabilités qui sont associées à chaque cadre. Cinq exemples extraits du corpus sont également consultables pour chacun des cadres de chaque verbe. Le fait d'extraire les exemples en même temps que les cadres de sous-catégorisation facilite la validation du lexique et le repérage d'erreurs à partir d'usages attestés.

Protocole d'évaluation

L'évaluation du lexique a été effectuée en comparant les cadres obtenus pour 25 verbes[51] avec les entrées du *Trésor de la Langue Française informatisé (TLFi)* [52]. Les verbes ont été choisis autant pour leur diversité syntaxique et sémantique que pour la variabilité de leur nombre d'occurrences dans le corpus (de 200 à plus de 100 000 occurrences).

Les scores de précision, de rappel et de F-Mesure[53] sont ensuite calculés automatiquement. Notons que l'emploi de ces termes peut toutefois être trompeur : en fait, ces mesures permettent juste de donner une idée du nombre de schémas de sous-catégorisation partagés ou non entre les différentes ressources. Une étude des résultats plus minutieuses doit ensuite être menée pour déterminer d'une part les erreurs d'analyse, d'autre part la nature des schémas de sous-catégorisation spécifiques à chaque ressource[54].

[51] La liste des 25 verbes est la suivante : *aimer, apprendre, chercher, comprendre, compter, concevoir, continuer, croire, donner, exister, jouer, montrer, obtenir, offrir, ouvrir, posséder, proposer, refuser, rendre, s'abattre.*

[52] Accessible à l'url suivante : `http://atilf.atilf.fr/tlf.htm`

[53] Pour calculer ces scores, nous avons besoin de compter les nombres de :
- Vrais positifs (VPs) : ce sont les cadres de sous-catégorisation correctement identifiés par le système (les cadres qui se trouvent à la fois dans notre lexique et dans le *TLFi*) ;
- Faux positifs (FPs) : ce sont les cadres de sous-catégorisation incorrects (ils sont proposés par le système mais absents du *TLFi*) ;
- Faux négatifs (FNs) : ce sont les cadres de sous-catégorisation corrects mais "oubliés" par notre système (absents de notre lexique mais présents dans le *TLFi*).

Les scores de précision et de rappel sont calculés comme suit :

$$Précision = \frac{Nombre\ de\ VPs}{Nombre\ de\ VPs + Nombre\ de\ FPs}$$

$$Rappel = \frac{Nombre\ de\ VPs}{Nombre\ de\ VPs + Nombre\ de\ FNs}$$

Le score de F-Mesure permet de combiner les scores de précision et de rappel dans un même indicateur, ce qui facilite la comparaison de résultats entre différents travaux :

$$F - Mesure = \frac{2 * Précision * Rappel}{Précision + Rappel}$$

[54] Les ressources peuvent ne pas concorder si un schéma a été attribué à un verbe par erreur. Mais *LexSchem* ou le *TLFi* peuvent aussi contenir des schémas correspondant à des emplois spécifiques à un

	# verbes	# cadres	Préc.	Rappel	F-Mesure
Preiss *& al.*, 2007	6300	168	0,82	0,59	0,69
Chesley & Salmon-Alt, 2006	104	27	0,87	0,54	0,67
LexSchem	3268	336	0,79	0,55	0,65

TAB. 3.1 – Comparaison avec les travaux récents

Analyse des résultats

Nous avons calculé les scores de précision, de rappel et de F-Mesure pour les 25 verbes choisis en comparant les entrées de chaque verbe dans le TLFi avec leurs entrées dans LexSchem. Nous avons obtenu une précision de 0,79 et un rappel de 0,55. La F-mesure correspondante est 0,65. Ces résultats sont légèrement inférieurs aux travaux équivalents récents (voir tableau 3.1).

Nous avons mené une étude des cas de désaccords entre les deux ressources. L'origine des erreurs peut être multiple. Le TreeTagger provoque des erreurs en étiquetant des noms comme verbes ; Syntex effectue des erreurs d'attachement qui se retrouvent ensuite dans nos cadres. Ces erreurs sont à la fois des "oublis" d'attachement (Syntex ne trouve pas un argument qui devrait être relié au verbe) et des attachements qui n'ont pas lieu d'être (Syntex attache au verbe un argument qui ne devrait pas l'être). Si le deuxième type d'erreur peut être corrigé par notre système (via le module de filtrage notamment), nous ne pouvons rien faire contre les oublis d'attachement. Ces oublis arrivent fréquemment, en particulier lorsqu'il y a incise ou inversion de l'ordre des arguments. Le module de construction des cadres de sous-catégorisation effectue lui aussi des erreurs : les informations pour reconstituer les cadres sont parfois insuffisantes ou ambiguës. Enfin, le module de filtrage effectue de nombreuses erreurs en écartant des cadres valides et en acceptant des cadres incorrects. Ces erreurs pourraient être corrigées en améliorant la technique de filtrage ou en trouvant des seuils plus "précis".

Même si les performances sont comparables aux autres travaux équivalents, un grand nombre d'erreurs subsistent. Il est alors légitime de se demander si un lexique avec de tels taux d'erreurs est utilisable pour des applications de TAL. Une première réponse est apportée par John Carroll *et al.* qui ont montré qu'on pouvait améliorer les performances d'un analyseur syntaxique en le lexicalisant, même si le lexique utilisé contient un taux d'erreurs élevé [Carroll *et al.*, 1998]. Plus récemment, il a été montré que des données, même bruitées, permettent d'obtenir des classes sémantiques satisfaisantes, en particulier sur des domaines spécialisés pour lesquels des ressources générales comme WordNet ne sont pas appropriées [Korhonen *et al.*, 2006b]. Signalons par ailleurs que des techniques efficaces de recherche d'erreurs à partir de corpus ont récemment été mises au point, en repérant par exemple des configurations phrastiques contradictoires avec les informations lexicales connues [Sagot et Villemonte de la Clergerie, 2006]. Enfin, une validation manuelle, même s'il s'agit d'une étape coûteuse, peut être mise en place ; cette validation

corpus donné. L'acquisition permet alors de mesurer l'« écart linguistique » sur le plan lexical entre deux corpus.

reste moins coûteuse que le développement complètement manuel d'une ressource. Notre approche garantit en outre une certaine homogénéité dans la stratégie de construction du lexique.

Les résultats sont surtout intéressants pour mettre à jour une ressource existante. L'analyse de ce que l'on a appelé des "faux positifs" est particulièrement précieuse dans ce cas : les faux positifs correspondent en effet à des schémas trouvés en corpus et ne figurant pas dans la ressource de référence. La validation de cet ensemble de données est relativement peu coûteuse (une liste d'exemples attestés est automatiquement fournie pour chaque schémas de sous-catégorisation) et permet une mise à jour à partir de corpus qui ne serait pas possible avec une approche purement manuelle (le nombre de verbes considéré est trop important et les corpus trop gros pour être exploités manuellement).

3.2.5 Commentaires sur les expériences

L'évaluation présentée dans la section précédente a été faite à partir du TLFi, essentiellement pour des raisons pratiques (disponibilité de la ressource, neutralité théorique). Ce choix n'est cependant pas sans poser problème, dans la mesure où un nombre non négligeable de faux négatifs (cadres "oubliés" par le système) correspondent en fait à des emplois anciens ou littéraires enregistrés dans le TLFi mais absents du corpus *Le Monde*. L'utilisation d'une ressource comme DicoValence [van Den Eynde et Mertens, 2006] ou le Lexique-Grammaire [Gross, 1975] aurait en partie permis de résoudre cette difficulté mais ces ressources étant fondées sur des théories particulières (par exemple, l'approche pronominale [van Den Eynde et Blanche-Benveniste, 1978] dans le cas de DicoValence), l'interprétation des informations qui y sont décrites n'est pas triviale[55].

Le choix de la ressource de référence dépend aussi de la finalité de l'étude. On peut ainsi considérer le choix du TLFi comme une manière de mesurer l'écart entre une ressource fondée sur le français littéraire et une ressource fondée sur un corpus journalistique. On constate alors, par exemple, que *Le Monde* a en définitive peu d'emplois spécialisés qui ne se trouvent déjà enregistrés dans le TLFi. En revanche, ce dernier inclut de nombreux emplois anciens et littéraires qui ne sont plus en usage. Ce type d'analyse ne peut bien sûr se contenter de chiffres donnant une vue macroscopique comme nous le faisons ici mais doit se doubler d'une étude de type lexicographique fine.

Nous avons depuis entamé une évaluation à plus large échelle sur la base de 300 verbes également présents dans TreeLex [Kupść, 2007]. TreeLex est un lexique dérivé automatiquement du corpus arboré de Paris 7 [Abeillé *et al.*, 2003], lui-même composé d'articles issus du journal *Le Monde* annotés à la main. Les résultats de l'évaluation avec TreeLex sont meilleurs en ce qui concerne le rappel (par rapport à l'évaluation avec le TLFi), ce qui n'est pas surprenant dans la mesure où les deux corpus sont d'origine identique. Une étude fine par type de cadres et par type de verbes, tenant compte de la fréquence et de la variabilité des emplois est en cours.

[55] Un travail de « traduction » des tables du Lexique Grammaire en un format plus facilement utilisable par des applications de TAL a été entrepris au LORIA [Gardent *et al.*, 2005].

Les étapes suivantes consistent à affiner la méthode d'acquisition à partir de corpus. Il faudrait en effet mieux tenir compte de la dispersion des compléments en fonction de la nature de la préposition tête (plus un complément est employé avec des verbes différents, plus il a tendance à constituer un modifieur). Il faudrait aussi tenir compte de la variabilité des noms apparaissant en position d'argument : le repérage de régularités à ce niveau doit permettre l'identification de structures figées ou semi-figées, à l'image des travaux de C. Fabre et D. Bourigault [2008]. Il est enfin possible d'améliorer la couverture en ayant recours au web (pour la langue générale) ou en baissant le nombre d'occurrences pour l'acquisition, en faisant attention à ne sélectionner que des phrases simples dont l'analyse est quasi-sûre.

Il faut ensuite valider l'approche à travers des applications et des cadres pratiques. Le lexique ainsi obtenu à partir du journal *Le Monde* peut permettre d'améliorer l'analyse de Syntex sur des corpus journalistiques. En effet, quand les phrases sont trop complexes, Syntex laisse certains compléments "libres" si aucun indice fiable ne permet de proposer un rattachement assez sûr (rappelons que l'analyseur n'est pas lexicalisée). Dans la mesure où LexSchem fournit des indications fiables quant à la sous-catégorisation des verbes, certains problèmes de rattachement pourraient ainsi être résolus.

Au delà, c'est sur des corpus de spécialité que l'approche est intéressante, dans la mesure où elle permet d'acquérir des types de comportements syntaxiques spécifiques à partir de corpus annotés. Une étude est en cours pour le domaine de la médecine.

Enfin, l'approche peut être adaptée pour traiter différentes parties du discours. Cédric Messiant a récemment adapté la méthode pour traiter les schémas de sous-catégorisation autour du nom. Cette étude est menée en collaboration avec des linguistes de l'Université Lille 3 (A. Balvet, R. Marin) qui se servent des résultats fournis par le système ASCCI comme une première étape permettant d'accéder très rapidement à un très grand nombre de données. Un important travail manuel est ensuite nécessaire si on souhaite obtenir une ressource ayant un taux d'erreur très marginal. Il semble toutefois que l'approche fournit une aide précieuse pour la constitution initiale de ressources ou pour compléter une base déjà existante.

3.3 Acquisition semi-automatique de familles sémantiques

En dehors des informations sur la sous-catégorisation syntaxique, il est important d'avoir des connaissances précises sur le rôle des arguments autour du prédicat. C'est dans ce but que j'ai exploré différentes stratégies pour l'acquisition semi-automatique de classes sémantiques et de restrictions de sélection. Les expériences décrites dans cette partie reposent toutefois sur des corpus beaucoup plus modestes que ceux utilisés pour l'acquisition de schémas de sous-catégorisation. Un travail reste à faire pour coupler les deux approches, comme mentionné dans la discussion qui clôt ce chapitre.

Les travaux décrits ici reposent sur des techniques d'apprentissage symbolique d'une part, et sur un réseau sémantique à large couverture d'autre part.

3.3.1 Acquisition automatique de classes par apprentissage symbolique interactif

La première expérience repose sur l'acquisition de connaissances à partir de corpus, par apprentissage symbolique interactif. J'ai utilisé pour cela un outil appelé Asium, développé par David Faure dans le cadre de sa thèse sous la direction de C. Nédellec, au Laboratoire de Recherche en Informatique (LRI) d'Orsay [Faure et Nedellec, 1998; Faure, 2000]. Asium permet l'acquisition de classes sémantiques à partir de corpus représentatifs. Il a pour objectif d'aider un utilisateur à créer un ensemble de classes sémantiques pertinentes par rapport à une tâche et à les classer hiérarchiquement.

Asium a été choisi essentiellement pour sa méthode de validation interactive. L'approche repose sur une analyse distributionnelle, permettant de générer des classes de mots apparaissant dans des contextes similaires (les classes de base). La mesure de similarité d'Asium permet de calculer le recouvrement entre classes et de proposer le regroupement de classes ayant une similarité supérieure à un seuil fixé par l'expert [Faure, 2000]. Les classes de base sont successivement agrégées par une méthode coopérative, ascendante en largeur d'abord afin de former les concepts de l'ontologie niveau par niveau. Ce processus de catégorisation ne se contente pas d'identifier des listes de noms apparaissant dans des contextes similaires, mais il augmente ces listes par induction. En effet, le rassemblement de deux classes de base (`C1` et `C2`) trouvées après deux contextes différents `CTXT1` et `CTXT2` permet d'autoriser les séquences non attestées `CTXT1 C2` et `CTXT2 C1`. Par exemple, imaginons que les mots *acquisition* et *fusion* aient été regroupés au sein d'une même classe. Si *procéder à une acquisition* est attesté, le système va induire par généralisation que *procéder à une fusion* est une structure valide, même si elle n'est pas attestée en corpus. Asium requiert, à ce stade, l'intervention d'un expert du domaine pour vérifier les classes une à une, par mesure de similarité décroissante, jusqu'à atteindre le seuil en-deçà duquel les classes ne sont plus présentées à l'expert.

Les classes de base fournies par Asium nécessitent un important travail de validation. Dans nos expériences, du fait de la faible quantité de corpus disponible, le seuil retenu pour les classes de base est obligatoirement bas et les rapprochements entre classes s'opèrent sur un nombre d'éléments communs relativement peu important.

Asium utilisait initialement une approche non supervisée, prenant en compte toutes les structures de dépendance au sein du texte. Cette stratégie oblige à valider ou à refuser un nombre important de classes qui ont des éléments communs mais qui ne sont pas pertinentes pour le domaine. Asium a ensuite été modifié afin de focaliser la recherche sur les seules classes comprenant des éléments du domaine, définis par un ensemble de mots placés dans un fichier jouant le rôle de filtre.

Cette stratégie débouche sur un apprentissage de type supervisé, ce qui améliore notablement la pertinence du résultat, à condition de définir à chaque fois un ensemble de

mots pertinents. Suivant la classe sémantique que l'analyste cherche à modéliser, cet ensemble de mots n'est pas le même. Il y a donc là un facteur certain de subjectivité : la définition du filtre est une opération subjective qui influe notablement sur les résultats [Faure et Poibeau, 2000].

Par rapport à un mode non supervisé, la mise en place de processus de contrôle (filtres de mots clés) contribue à améliorer la qualité des résultats et, surtout, fait décroître le temps de mise au point des ressources. La taille des corpus utilisés en extraction (quelques milliers de mots le plus souvent, dans le contexte industriel des expériences mentionnées[56]) ne permet cependant pas d'obtenir des classes de base réellement satisfaisantes. Ici, les filtres sont de simples listes de mots pertinents : il s'agit sans doute d'une ressource trop peu informative pour permettre une amélioration vraiment significative des résultats. Par contraste, on peut noter les expériences de Morin et Jacquemin [1999] fondées sur des données beaucoup plus structurées (des thésaurus), afin de guider et d'affiner les résultats de l'apprentissage. Dans notre cas aussi, la mise en œuvre de telles données permettrait sans doute une acquisition de meilleure qualité.

3.3.2 Utilisation d'une ressource linguistique générale : le réseau sémantique de Memodata

Afin de contraster l'expérience faite avec Asium, j'ai procédé à la même manipulation en partant d'un réseau sémantique couvrant la langue générale. Pour ce faire, j'ai retenu le réseau de la société Memodata qui a développé depuis plus de 10 ans une ressource très complète, accompagnée de fonctionnalités diverses, accessibles via une interface de programmation (API, *application Programming Interface*) [Dutoit, 2000].

Le Dictionnaire Intégral est le nom du réseau sémantique développé par Memodata. Il est fondé sur la notion de mots-sens : les éléments de base du réseau sont les différents sens associés à un mot (une chaîne de caractère). Ce réseau est riche d'environ 186 000 mots-sens : il était donc comparable, au moment des expériences, au Wordnet anglais [Fellbaum, 1998], mais il en diffère notablement par la façon dont l'information y est structurée. En effet, sa structuration n'est pas seulement fondée sur une relation « *est-un* » et sur des présupposées psychologiques, mais aussi sur une décomposition des mots-sens en sèmes (analyse componentielle), structurés par des liens typés (pour plus de détails, voir [Poibeau et Dutoit, 2008]).

Le réseau intègre de nombreuses relations entre des éléments de catégories syntaxiques différentes. Cet aspect est bien évidemment primordial pour la tâche qui nous intéresse, puisqu'il nous faut par exemple avoir accès aussi bien au nom *achat* qu'au verbe *acheter* pour identifier la notion d'acquisition. La couverture du Dictionnaire Intégral est plus large que celle du réseau sémantique au format EuroWordnet existant pour le français, une partie de l'information contenue dans EuroWordnet provenant d'ailleurs des bases de

[56] Ce type d'expériences contraste avec celles qui sont faites aujourd'hui à partir du web, pour lesquelles les masses de données traitées sont souvent beaucoup plus importantes. Il existe cependant toujours des besoins industriels spécifiques sur des quantités de données limitées.

Memodata (le reste du réseau EuroWordnet français a été produit en majeure partie par l'Université d'Avignon, responsable du projet pour la France) [Vossen, 2002].

3.3.3 Évaluation et comparaison des deux approches

Les deux approches ci-dessus ont été comparées dans le cadre de ma thèse. Le but était alors de développer un système d'extraction d'information pour le domaine financier (pour un site boursier en ligne : `http://www.firstinvest.com`). Il fallait notamment trouver des informations sur les achats, ventes et fusions de sociétés, ce qui nécessite l'élaboration de classes sémantiques comportant des termes généraux (*acheter*, *vendre*, *etc.*) mais aussi plus techniques (*faire une OPA*, *procéder à un échange d'océanes*[57], *etc*).

Comme la première approche (par apprentissage) est fondée sur un corpus représentatif, les classes obtenues sont bien ciblées. L'approche souffre toutefois de deux défauts majeurs : le nombre d'erreurs d'analyse oblige à un important travail manuel pour valider les résultats. D'autre part, les mots ayant un faible nombre d'occurrences sont assez largement ignorés, du fait que la méthode est très sensible à la fréquence.

L'utilisation d'un réseau sémantique comme celui de Memodata est complémentaire. Les classes fournies par le réseau sont riches et généralement pertinentes, mais des éléments importants sont manquants, notamment la plupart des termes techniques. L'avantage majeur d'une ressource comme Le Dictionnaire Intégral est son immédiate disponibilité : l'utilisateur n'a pas besoin de définir un protocole expérimental et de valider des classes sémantiques de manière interactive, il lui suffit de parcourir le réseau et de sélectionner les éléments qui lui semblent intéressants.

3.3.4 Définition d'une méthode hybride

Globalement, les résultats obtenus sont complémentaires. Asium permet de bien "cibler" les éléments fréquents, qui par chance comportent souvent les principaux termes techniques d'un domaine donné. Le Dictionnaire Intégral permet de compléter cette analyse en ajoutant des éléments plus rares mais néanmoins pertinents. Une approche hybride a alors été mise au point : après chaque itération avec Asium, les liens de synonymie et d'hyponymie du Dictionnaire Intégral sont parcourus sans récursion (on n'a donc accès ainsi qu'aux items lexicaux ayant un lien direct avec les mots déjà présents dans la classe). Les nouveaux mots glanés ainsi sont proposés à l'utilisateur qui peut choisir de les retenir ou non. A l'issue de ce processus, on observe que la couverture obtenue est meilleure qu'avec chaque ressource utilisée séparément. Bien sûr, comme dans toute expérience sur corpus, il demeure des éléments absents et des éléments ambigus, ce qui empêche d'obtenir des performances parfaites.

Soulignons enfin la subjectivité de la tâche. On est ici très proche de l'annotation sémantique entrevue au chapitre précédent (cf. chapitre 2, page 34). Une classe sémantique

[57] Terme financier désignant des obligations convertibles en actions nouvelles ou existantes.

n'a pas de contours nets : la décision consistant à inclure un mot ou non dépend de l'appréciation de l'analyste chargé de mettre au point le système. Même si l'évaluation reste possible par rapport à une référence ou une application donnée, celle-ci reste relative à l'expérience décrite et est relativement peu généralisable.

3.4 Discussion et perspectives

Les développements présentés dans cette section sont encore à un état exploratoire. Le véritable enjeu consiste en effet à évaluer les cadres de sous-catégorisation dans le cadre d'applications réelles, afin de déterminer leur apport. Il faut ensuite continuer ces expériences et proposer de nouveaux algorithmes permettant d'enrichir les ressources acquises [Saint-Dizier et Viegas, 1995; Saint-Dizier, 1999].

Une fois toutes les informations mises en relation, il est possible de se rapprocher d'une ressource comme VerbNet, à condition de compléter l'acquisition automatique par un important travail de vérification et de correction manuelle. Au niveau théorique, nous gardons en arrière-plan l'idée d'explorer l'interface syntaxe-sémantique, comme exposé plus haut. D'une manière générale, la comparaison avec des ressources comme FrameNet [Baker *et al.*, 1998] ou VerbNet [Kipper-Schuler, 2003] (directement inspiré des travaux de B. Levin [1993]) seraient intéressants. En filigrane se pose le problème de la comparaison de classes syntaxico-sémantiques dans des langues différentes.

Au-delà, il me semble que ces expériences posent deux questions majeures. D'une par le rapport entre ces schémas prédicatifs de nature linguistique et la modélisation de type logique qu'on peut en tirer. D'autre part, la signification fondamentale des données statistiques obtenues par analyse de corpus.

Sur le premier plan, aussi bien Tesnière que Fillmore sont très clairs : la notion de prédicat linguistique est beaucoup plus complexe que son équivalent logique. Les alternances, les transformations et les glissements de sens rendent la sémantique lexicale infiniment plus complexe que le calcul des prédicats en logique. Cet aspect est détaillé dans A. Lemaréchal [1995] mais l'auteur souligne aussi l'intérêt d'une représentation sous forme de prédicat logique, pour comparer des langues en apparence très différentes. C'est une perspective que je compte explorer à l'avenir.

L'aspect statistique a un intérêt certain pour modéliser les relations entre prédicats et arguments. Comme nous l'avons souligné, ce mode de représentation permet de dépasser l'opposition binaire et quelque peu réductrice entre arguments et circonstants. Cependant, cette représentation n'est pas sans poser des questions : quelle est la nature de la nouvelle représentation obtenue (cf. chapitre 1, page 29) ? Est-elle valable au-delà d'un domaine, voire d'un corpus donné ? Sur un plan cognitif, peut-elle être mise en rapport avec une dimension intrinsèquement statistique de la langue, dimension qui se manifesterait aussi bien dans les phénomènes d'acquisition que de compréhension [Bybee, 2006] ?

Une partie des erreurs du système provient du fait que certaines rencontres de surface fréquentes entre verbe et complément ne correspondent en fait pas à des arguments. On

est alors dans des cas de collocations, voire de colligations, pour reprendre la terminologie de Firth (cf. chapitre 1, page 18). Dans quelle mesure ces rencontres de surface sont-elles pertinentes ? Que disent-elles sur notre façon d'appréhender la langue ? Ce sont des questions ouvertes qui touchent aux préoccupations de la linguistique cognitive[58]. Elles posent la question du rapport entre des approches linguistiques comme celles que nous avons évoquées autour de Fillmore ou Goldberg et des approches informatiques à base de gros corpus comme celles que nous avons commencé à explorer.

3.5 Synthèse

J'ai détaillé dans cette partie différentes expériences visant à acquérir des informations de nature prédicative à partir de corpus. J'ai d'abord présenté une expérience portant sur de très gros corpus et visant à acquérir des informations de sous-catégorisation pour un grand nombre de verbes. La suite présente des travaux portant sur des corpus de taille plus modeste : il s'agit de mettre au point, à des fins d'extraction d'information essentiellement, des techniques efficaces pour acquérir des informations sur le rôle sémantique des arguments et l es restrictions de sélection. Ces expériences, encore en cours, sont amenées à être développées dans les années qui viennent pour enrichir la couverture et affiner les informations acquises à base de corpus.

[58] D. Legallois [2006; 2008] a étudié ces questions en détail et propose un rapprochement qui me semble très pertinent entre la tradition cognitive américaine — notamment les grammaires de construction [Goldberg, 1995, 2006] — et la linguistique de corpus de tradition anglaise venant à la suite de Firth [Firth, 1957b; Palmer, 1968]. Ces deux courants semblent s'être assez largement ignorés. Pourtant, la linguistique cognitive a proposé des analyses intéressantes de certains phénomènes mis en avant, sous une lumière assez différente, par la linguistique de corpus.

Chapitre 4

Analyse et typologies de documents procéduraux

Les analyses présentées jusqu'ici visaient essentiellement le niveau du mot et du syntagme. Mais ce sont vers les textes que sont principalement tournées les applications : il faut faciliter l'accès aux documents quand ils sont sur support informatique, permettre une prise de connaissance rapide du contenu et naviguer rapidement à l'intérieur de ceux-ci pour y trouver l'information pertinente.

Ceci est particulièrement vrai des documentations techniques. Il existe des besoins avérés et exprimés de manière récurrente, dans divers cadres professionnels, en matière d'informatisation de la documentation. La notion de documentation technique est large et variée : il peut s'agir de descriptions de mécanismes ou de techniques, de procédures ou de textes réglementaires, *etc.*

Au sein de l'équipe Représentation des Connaissances et Langage Naturel (RCLN), l'étude de textes de spécialité a toujours tenu une place privilégiée. C'est ainsi que je me suis intéressé à des textes de médecine, pour proposer une aide à la modélisation fondée sur une pré-formalisation du contenu linguistique. La thèse d'Amanda Bouffier au LIPN, sous la direction de Daniel Kayser et de moi-même, s'inscrit dans ce cadre dans la mesure où elle porte sur la modélisation de textes procéduraux ayant trait au domaine médical, les Guides de Bonnes Pratiques. Ce type de documents a été retenu car il se prête bien à des analyses discursives sur du texte intégral, parce qu'un modèle formel standardisé existe et que la tâche répond à un vrai besoin en informatique médicale [Bouffier, 2007; Bouffier et Poibeau, 2007a,b].

Nous nous sommes alors interrogés sur les moyens de poursuivre et de généraliser ce travail. Les textes concernés, marqués par la présence de procédures, forment-ils une classe de documents particuliers et homogènes ? Cette classe est-elle repérable par des moyens automatiques ? Correspond-elle à un type ou à un genre donné ? D'une manière plus générale, nous nous sommes interrogés sur les rapports entre le local et le global, c'est-à-dire sur la façon de lier entre elles des analyses plus ou moins locales pour proposer des parcours de lecture dynamiquement adaptables, tout en restant cohérents.

Plus que les autres encore, ce chapitre doit beaucoup à des projets et des collaborations avec des collègues linguistes ou informaticiens. Outre la thèse d'Amanda Bouffier déjà mentionnée, la réflexion sur les genres et les types textuels a été en grande partie menée dans le cadre du projet TEXTCOOP[59], avec Françoise Gayral, Maria Zimina-Poirot puis Marie-Paule Jacques. Ce chapitre reprend donc des éléments de réflexion collectifs mais j'en assume la mise en perspective et bien évidemment les éventuelles erreurs.

4.1 Modélisation d'un genre de textes particulier : les Guides de Bonnes Pratiques

Les guides de bonnes pratiques (désormais GBP) sont des textes procéduraux appartenant au domaine médical. Ils occupent une place particulière à l'intérieur de ce domaine, à la fois par les situations de communication dans lesquels ils s'inscrivent et par leur fonction. Il s'agit de textes produits par les autorités sanitaires[60], s'adressant à des médecins pour les guider dans leur diagnostic et les traitements à prescrire. On peut les considérer ainsi comme des « dispositifs de communication socio-historiquement codifiés » pour reprendre les termes de D. Maingueneau [1996] et les rattacher à ce que Biber appelle des registres [Biber, 1995]. Des contextes de production aussi stéréotypés ont des conséquences à la fois sur l'organisation de tels textes (régularité dans la structuration du contenu et dans la structuration visuelle) et sur l'expression de leur contenu (vocabulaire, formes employées, *etc.*).

L'étude présentée dans cette section est fondée sur le travail de thèse d'Amanda Bouffier, effectuée au LIPN sous la direction de D. Kayser et de moi-même depuis la fin 2004.

4.1.1 La notion d'architecture textuelle

L'inscription des GBP dans un genre particulier rend possible la mise au jour de traits spécifiques, tant dans la structure visuelle que dans le contenu, permettant de repérer des blocs informationnels représentatifs. Comme l'ont montré M.-P. Péry-Woodley [2001] ou J.-M. Adam [2005], plusieurs plans d'analyse coexistent.

- Le premier plan concerne le texte indépendamment de la situation dans laquelle il a été émis. Il s'agit de mettre à jour son organisation, sa structuration en différentes unités, l'agencement de ces unités entre elles et leurs relations respectives.
- Le second restitue le texte dans sa situation d'énonciation et met en jeu des notions telles que l'intentionnalité de l'auteur, les actes de langage, les fonctions réalisées par le texte ou seulement par certaines de ses portions. Il insiste sur la fonction énonciative que chaque portion possède et sa relation aux autres.

[59] Projet ANR RNTL (2006–2008) dont les partenaires, outre le LIPN, sont l'IRIT et Sinequa (`http://www.textcoop.org/`).

[60] HAS, Haute Autorité de Santé ; ANAES, Agence Nationale d'Accréditation et d'évaluation en Santé

Il est clair que ces deux plans ne sont pas indépendants et que des indices de divers ordres jouent pour permettre de les repérer dans les textes. La « signalisation » de ces plans dans la surface du texte est multiple.

Elle peut concerner la seule mise en forme matérielle du texte qui n'est là considérée que du point de vue de son inscription physique, concrète sur un support matériel [Pascual, 1991; Virbel, 1985; Luc et Virbel, 2001]. Beaucoup de marques visuelles, qu'elles soient typographiques (gras, italique, *etc.*) ou dispositionnelles (présence de tabulation, sauts de lignes, *etc.*) fonctionnent comme des « instructions de lecture » qu'il serait dommage de ne pas utiliser. Les GBP sont ainsi structurés autour de séquences conditions/recommandations (du type *cas à traiter/traitement à suivre*) ; l'identification de ces séquences se fonde essentiellement sur la présence d'énumérations, le changement de paragraphes et les signes de ponctuation. Les titres jouent aussi un rôle majeur M.-P. Jacques et J. Rebeyrolle [2006] montrent qu'outre leur fonction de segmentation, de regroupement et de hiérarchisation, les titres servent, sur le plan notionnel, à donner des indications sur le contenu et participent ainsi à la construction du monde que le texte relate.

Cette signalisation repose aussi sur la présence de connecteurs et de marqueurs linguistiques. Suivant les études considérées, les auteurs peuvent en rester à un niveau purement descriptif ou chercher à automatiser le repérage. Pour illustrer le premier cas, citons M. Charolles [1997] qui s'intéresse à certaines expressions, souvent des syntagmes prépositionnels en position initiale et détachée, considérées comme introductives de cadres de discours. Pour lui, ces expressions délimitent un univers de discours spatial (*en France*), temporel (*en 1981*) ou un espace de discours (*selon Chirac*) et permettent de regrouper des contenus propositionnels dans un bloc homogène dont la sémantique dépend de celle de l'introducteur. Comme nous le verrons ultérieurement, les segments des GBP que nous cherchons à repérer sont le plus souvent introduits par un type d'expressions particulières et nous emprunterons à Charolles les notions de cadre et de portée.

Citons aussi Pascual et Péry-Woodley [1995] qui s'intéressent à certains verbes performatifs. Ceux-ci jouent un rôle pragmatique : des verbes comme *organiser*, *définir*, *conclure*, *commenter*... jouent un rôle méta-linguistique permettant de typer des segments de textes. Pour les GBP, nous verrons que certains verbes marquant une action sont ainsi de bons indices.

4.1.2 Présentation du corpus

Les GBP se présentent comme des catalogues de situations cliniques auxquelles sont associées des recommandations. Le peu d'impact de ces textes sur les pratiques des médecins a poussé à un effort en vue de leur informatisation. Le passage à un support informatique est censé favoriser la prise de connaissance des GBP par les médecins en situation de consultation, et donc contribuer à améliorer leur prise en compte par les médecins.

Les GBP informatisés doivent offrir différents moyens d'accès conviviaux, textuels ou non. Les systèmes de consultation reposent sur des bases de connaissances et des systèmes experts conçus à partir de l'analyse et de la modélisation des GBP [Musen *et al.*, 1996;

Boxwala *et al.*, 2004]. Cette étape est, à l'heure actuelle, lourde et coûteuse. Pour fournir une étape intermédiaire entre le texte brut et le modèle formel, le modèle documentaire GEM (Guideline Elements Model, `http://gem.med.yale.edu/`, [Shiffman *et al.*, 2000]), élaboré par l'université de Yale en coordination avec de nombreux groupes d'experts, est reconnu comme un standard international, indépendant de la langue. Ce modèle[61] a pour but de représenter et structurer les éléments textuels pertinents pour la modélisation. Cette étape intermédiaire représente déjà une aide à la modélisation mais le passage entre le texte et GEM est manuel. Le travail décrit ici vise à automatiser partiellement ce passage.

Pour ce faire, un corpus composé de 25 GBP publiés par l'ANAES (Agence Nationale d'Accréditation et d'Evaluation en Santé) et l'AFSSAPS (Agence Francaise de Sécurité Sanitaire des Produits de Santé) entre 2000 et 2005 a été assemblé. Ces GBP portent sur la prise en charge de diverses pathologies (par ex. diabète, hypertension, asthme *etc.*) et sur la pratique d'examens (par ex. endoscopie digestive basse). Le corpus, homogène du point de vue du style d'écriture, représente environ 250 pages imprimables (environ 150 000 mots). On peut trouver ces guides aux adresses `http://www.anaes.fr` et `http://affsaps.sante.fr`.

4.1.3 La segmentation des guides, un problème de portée

L'objectif est de fournir une aide à la modélisation en proposant une représentation structurée « à la GEM » des GBP. Comme les GBP sont fortement structurés autour de couples conditions/recommandations, le travail consiste à isoler les segments qui correspondent aux conditions et aux recommandations puis à lier ces unités entre elles [Bouffier, 2007].

La tâche serait relativement simple si chaque condition était suivie dans la même phrase de la recommandation à laquelle elle est liée de manière linéaire ; mais différents phénomènes textuels rendent la tâche complexe, la portée de la condition pouvant dépasser la phrase à laquelle elle appartient comme sur la figure 4.1.

Dans l'exemple 4.1, on peut voir la présence de trois segments « condition » (en gras). Le segment « *chez le sujet non immunodéprimé* » exprime une condition qui doit être liée non seulement à la recommandation qui le suit dans la même phrase (« *des biopsies coliques nombreuses et étagées sont recommandées* ») mais également aux conditions et recommandations exprimés dans les phrases suivantes. Ce segment a une portée dite « étendue ».

Dans ce qui suit, on appelle segment élémentaire une séquence de texte de taille égale ou inférieure à la phrase. On distingue d'une part les « segments condition » (en gras sur la figure 4.1), d'autre part les « segments recommandation » (segment élémentaire mentionnant une action que le médecin doit effectuer, en italique sur la figure 4.1). La portée d'un segment condition est représentée par un cadre qui est soit « minimal »

[61] Il s'agit en fait d'une définition de type de documents (DTD) XML, c'est-à-dire d'une grammaire décrivant les entités possibles et leur agencement autorisé.

IV-En cas de diarrhée chronique

Chez le sujet non immunodéprimé, en cas d'aspect macroscopique normal de la muqueuse colique,

des biopsies coliques nombreuses et étagées sont recommandées... Les biopsies isolées sont insuffisantes... L'exploration de l'iléon terminal est également recommandée (grade C).

En cas d'aspect normal de la muqueuse iléale ... ,

la réalisation de biopsies n'est pas systématique (accord professionnel).

Chez le sujet immunodéprimé,

il est nécessaire de réaliser des biopsies systématiques ...

FIG. 4.1 – Plusieurs cadres imbriqués

lorsqu'il ne porte que sur un segment recommandation faisant partie de la même phrase que son introducteur, soit « étendu » lorsqu'il intègre plusieurs segments recommandations (en général plusieurs phrases).

La principale difficulté est de délimiter ces cadres, autrement dit, de segmenter le texte en cadres conditionnels dans la mesure où leur clôture n'est pas explicite. Notons que cette clôture existe : le cadre devrait en théorie être borné par une limite claire, même si celle-ci reste implicite, car le médecin doit entreprendre une série d'actions thérapeutiques précise en fonction d'un ensemble de conditions. Il peut y avoir des choix, des alternatives mais cette série d'actions est néanmoins précise et finie. La tâche est distincte du repérage de cadres spatio-temporels, dans des textes narratifs par exemple : il est connu que les adverbiaux temporels ouvrent souvent des cadres dont les limites ne sont pas claires (on peut même parfois penser que l'adverbial ouvre plutôt une « perspective » temporelle ou spatiale, sans réelle marque de fin). Les tâches de segmentation le prouvent aisément : il existe une grande variabilité entre annotateurs sur des textes narratifs, chaque proposition de découpage pouvant être justifiée *a posteriori* [Bilhaut, 2006].

Notre objectif est donc de proposer une méthode de segmentation fondée sur des indices présents dans le texte et qui soit, si possible, généralisable à d'autres textes dits procéduraux. La méthode de segmentation des GBP opère en deux temps : dans un premier temps, on procède au repérage des segments élémentaires condition et recommandation. Ces segments élémentaires sont les briques de base pour calculer, dans un second temps, les cadres conditionnels, leur portée et leur éventuelle imbrication.

4.1.4 Stratégie d'analyse

La notion de portée se prête bien à l'analyse et on en connaît de multiples exemples en linguistique, aussi bien en syntaxe qu'en sémantique. On a ainsi vu fleurir ces dernières années un nombre important d'études linguistiques sur la place et la portée des adverbiaux dans les textes [Charolles et Péry-Woodley, 2005]. La modélisation des GBP passe bien évidemment par là, dans la mesure où les conditions que l'on cherche à repé-

rer correspondent généralement à des compléments adverbiaux dont il faut déterminer la portée. Le traitement automatique de ces questions reste cependant en retrait, à quelques exceptions près [Bilhaut, 2006].

Les études sur la position des adverbiaux laisse apparaître les tendances suivantes :
- lorsque l'adverbial est détaché en tête d'une séquence (énumération, paragraphe, *etc.*), il a tendance a porter sur toute la séquence qui suit ;
- lorsque l'adverbial est intégré à la phrase, il ne porte que sur la phrase courante.

Ce constat, déjà effectué par plusieurs auteurs (par ex. [Charolles *et al.*, 2005]) dans d'autres cadres — adverbes temporels, adverbes spatiaux — a été confirmé par une étude manuelle d'Amanda Bouffier sur le corpus préalablement constitué [Bouffier, 2006]. Cette structuration par défaut n'est pourtant pas sans exceptions. A. Bouffier constate notamment les deux cas suivants :

- la portée d'un adverbial peut être étendue (c'est-à-dire aller au-delà de sa frontière par défaut) si des marqueurs de cohésion sont présents dans le texte, au-delà de la frontière par défaut ;
- la portée d'un adverbial peut être réduite (c'est-à-dire s'interrompre en deçà de sa frontière par défaut) si des marqueurs de rupture sont présents dans le texte.

Les éléments marquant la cohésion et la rupture au sein du discours sont innombrables et ont fait l'objet de nombreuses études. Nous nous sommes fondé notamment sur les analyses de Halliday et Hasan [1976] et Halliday et Matthiessen [2004]. Les phénomènes de reprise ou de variation lexicale (par des synonymes, des hyponymes ou des hyperonymes), les déterminants démonstratifs et les pronoms sont autant de marqueurs de cohésion. Les conjonctions et les adverbiaux adversatifs (*mais*, *à l'opposé*, *d'un autre côté*, *etc.*) sont quant à eux des indices connus de rupture. Leur repérage reste un processus long et fastidieux qui peut en partie être automatisé (par le repérage de régularités en tête de phrase par exemple) mais qui reste encore largement manuel.

Il est ensuite apparu nécessaire d'appuyer l'observation linguistique sur une mesure objective de l'utilité relative de chaque trait linguistique retenu, et ses éventuels liens avec d'autres traits. Pour ce faire, chaque trait est modélisé de façon uniforme et un arbre de décision est induit. Les arbres de décision étant organisés à partir du trait le plus discriminant, on peut ainsi hiérarchiser les différents traits. Le processus permet également de tester chaque trait indépendamment des autres afin de déterminer son intérêt pour la tâche (mesure de gain d'information) et les corrélation de distribution entre traits pertinents [Bouffier, 2008].

Cette approche rappelle d'autres expériences dans le domaine. Il existe des études linguistiques qui vont dans le même sens, mais qui portent sur d'autres types d'adverbiaux [Bilhaut, 2006; Charolles *et al.*, 2005; Le Draoulec et Péry-Woodley, 2005]. Ces études sont souvent confrontées au problème de fin de cadre déjà évoqué : la fin du cadre n'est pas claire et partant, il est difficile de déterminer la valeur exacte des marqueurs dits de cohésion ou de rupture (c'est-à-dire leur pertinence pour la tâche).

Sur le plan automatique, l'approche peut être rapprochée de l'algorithme *shift/reduce* de D. Marcu [1999]. Celui-ci cherche à repérer automatiquement la structure de textes

en s'appuyant sur la *Rhetorical Structure Theory* (RST, [Mann et Thompson, 1988]). Il « empile » des séquences tant qu'il repère des marques de cohésion (dans un cadre légèrement différent puisqu'il s'appuie sur la RST) et passe à la séquence suivante (*shift*). Quand intervient une marque de rupture, il « clôt » le niveau d'analyse courant (*reduce*).

4.1.5 Architecture et implémentation

L'analyse s'effectue en trois étapes [Bouffier et Poibeau, 2007a].

1. Lors d'une première étape, les marqueurs pertinents pour l'analyse sont identifiés et annotés au sein du texte. Cette annotation concerne la structure matérielle (repérage des titres, paragraphes, énumérations) et divers éléments lexicaux pertinents pour la suite de l'analyse (verbes d'action, marqueurs de condition et de recommandations, marqueurs de cohésion et de rupture).
2. Lors d'une deuxième étape, les segments de condition et de recommandation sont identifiés. Ceux-ci correspondent le plus souvent à des phrases ou à des propositions à l'intérieur des phrases. L'analyse se fonde sur le repérage de faisceaux d'indices identifiés à l'étape précédente (introducteurs de conditions d'un côté, présence de verbes d'action et de structures argumentales types de l'autre).
3. Lors de la troisième étape, l'analyse de la portée est effectuée, suivant la technique indiquée dans la section précédente, en procédant d'abord à une analyse par défaut et en corrigeant celle-ci quand une exception apparaît.

L'analyse reposant sur des faisceaux d'indices et des règles complexes et non sur une approche clairement procédurale, une architecture de type tableau noir a été adoptée. Celle-ci permet de faire collaborer des agents, chaque phénomène linguistique étant analysé de manière autonome et incrémentale. Les annotations sont stockées dans une base (le tableau noir) unique, à laquelle chaque agent a accès. Les étapes (2) et (3) sont effectuées par des agents qui se déclenchent automatiquement à partir du moment où leurs conditions d'application sont remplies. Un superviseur contrôle le processus et résout les conflits (quand, en fonction d'une configuration donnée, deux agents peuvent intervenir).

4.1.6 Évaluation

Les outils aidant à la formalisation des GBP doivent être capables de proposer une analyse du guide proche de celle d'un expert. Pour appréhender leur performance, il convient de comparer leurs résultats à ceux d'un groupe d'experts sur des guides n'ayant pas servi lors de la phase de développement du système.

Élaboration de la référence manuelle

Un certain nombre de GBP ont été annotés manuellement pour servir de base à l'étude et pour jouer le rôle de "référence" lors de l'évaluation. L'évaluation est faite avec cinq GBP

	Conditions	**Recommandations**	**Portée**
# référence	278	401	679
# trouvés	228	349	
# corrects	211	340	531
Rappel	0,75	0,84	
Précision	0,92	0,97	**0,78**
F-mesure	**0,83**	**0,90**	

FIG. 4.2 – Evaluation de l'analyse des GBP.

n'ayant pas servi lors de la mise au point du système. Globalement, chaque annotateur doit dériver à partir d'un GBP un arbre XML conforme à la notation GEM (un guide d'annotation sommaire était fourni aux annotateurs). Même si toutes les annotations n'ont pas été effectuées par un expert du domaine, toutes ont été au moins validées par un expert (praticien hospitalier du LIM&Bio, laboratoire d'informatique médicale de l'Université Paris 13). Plusieurs guides ont été doublement annotés, chaque annotateur travaillant évidemment indépendamment de son collègue. L'accord entre annotateurs est ensuite calculé en comparant le nombre de recommandations reliées à l'action ou à l'ensemble d'actions dont elle dépend dans les deux ressources annotées [Bouffier, 2008].

L'accord entre annotateurs, calculé à partir d'un ensemble de 162 noeuds, est de 0,96 (157 segments correctement raccordés sur 162). Ces résultats montrent la faisabilité de la tâche et une variation dans l'annotation réduite, au moins sur certains GBP. Des expériences faites en comparant une annotation non-experte avec une annotation experte montrent seulement des différences minimes. Ceci permet de dégager deux conséquences :

1. l'annotation ne repose que légèrement sur des connaissances expertes et ;
2. elle peut être faite manuellement en s'appuyant sur les caractéristiques linguistiques et typo-dispositionnelles, dans la mesure où les connaissances du domaine ne sont que peu sollicitées.

D'autres études (par exemple [Bilhaut, 2006]) ont mis en évidence une plus grande variabilité entre annotateurs humains. Il serait intéressant de prolonger l'expérience pour voir dans quelle mesure ces résultats varient en fonction du GBP analysé, des guides d'annotations fournis et des annotateurs concernés.

Évaluation du découpage en segments et de leur portée

On évalue le découpage en segments élémentaires en calculant les scores de rappel et de précision par rapport à la référence (GBP annotés manuellement). Les résultats sont calculés à partir de l'analyse de cinq GBP (la F-mesure est la moyenne harmonique du rappel et de la précision).

Pour l'analyse de la portée (rattachement des recommandations aux conditions, dernière colonne), seul le nombre d'éléments correctement rattachés par rapport à la référence est pris en compte afin d'obtenir un score de précision (*accuracy*).

Les résultats obtenus sont une F-mesure de 0,83 pour la reconnaissance des recommandations et 0,90 pour la reconnaissance des conditions. Il s'agit de segments souvent bien marqués dans le texte, repérables par des indices linguistiques peu ambigus. La portée des conditions est analysée correctement dans 78 % des cas. Il semble que ce résultat soit suffisant pour être utilisable, pourvu évidemment qu'il soit corrigé et validé manuellement : il permet de réduire le temps d'analyse manuelle et offre une base systématique, dans la mesure où l'analyse automatique est fixe et régulière.

Ces résultats peuvent être comparés favorablement avec d'autres types d'analyse discursive, visant par exemple à déterminer la portée d'adverbiaux temporels ou locatifs, même si une telle comparaison est difficile. Les GBP ont cependant des frontières mieux marquées et un style plus régulier, qui facilite l'analyse.

Dans la plupart des cas, la portée de la condition peut être repérée à partir de l'application des règles par défaut. Cependant, quelques cas importants sont résolus grâce aux règles fondées sur l'analyse des exceptions, qui permettent d'aller en deçà ou au-delà de la portée par défaut. Le système échoue en cas de portée étendue, quand celle-ci est exprimée par des marques de cohésion liées au vocabulaire du domaine (utilisation de synonymes, d'hyponymes ou d'hyperonyme) ou de structures complexes (anaphores nominales, structures syntaxiques complexes). Résoudre ces cas rares reviendrait à introduire de nombreuses connaissances du domaine dans le système, avec tous les risques que cela comporte. Nous avons volontairement écarté cette voie dans la mesure où on souhaite garder un système relativement portable, faisant un usage minimal de connaissances spécifiques.

Il semble naturel de vouloir élargir l'analyse à d'autres types de textes, du domaine médical notamment. Les GBP sont structurés autour de la notion de cas cliniques, mais de nombreux autres documents procéduraux peuvent offrir des informations complémentaires pour des cas similaires, sans être eux-mêmes « estampillés » comme GBP.

4.2 Extension de l'analyse à d'autres types de textes

Afin d'étendre le travail effectué sur les GBP à d'autres genres de textes, il faut d'abord s'interroger sur les caractéristiques de l'objet manipulé, et sur les notions de *genre* et de *type*. Nous pourrons alors voir dans quelle mesure la notion de procéduralité permet de définir un type de textes particulier, et comment le traiter.

4.2.1 Qu'est-ce qu'un texte ?

Beaucoup de tentatives ont été faites pour caractériser la notion de texte, souvent liée voire confondue avec la notion d'écriture. Platon, dans *Phèdre*, souligne le caractère figé du texte : celui-ci a été produit par un individu mais il est livré au lecteur hors contexte. A l'opposé d'une situation de dialogue, nul n'est présent pour réagir, répondre à des questions qui se poseraient par rapport au texte.

Or, ce scénario ne s'applique pas toujours. Un texte est produit dans un cadre et est reçu dans un autre cadre, par des individus qui en prennent connaissance, le lisent et peuvent agir en fonction de son contenu. Ce schéma n'est pas sans rappeler le schéma traditionnel du processus de communication, où un locuteur (*resp.* l'auteur) émet un message (un texte) qui doit être décodé par un interlocuteur (lecteur). Ce qui est souligné, notamment chez Platon, c'est le fait que le texte est livré tel quel, que c'est en quelque sorte un « objet mort » : le passage au support écrit produit une distanciation et, sauf exception, le processus de compréhension n'est pas fondé sur un échange avec l'auteur[62]. Il n'empêche que la connaissance des conditions de production du texte, du type de document et de ses visées (pratiques ou non) sont souvent précieuses, voire indispensable à sa bonne compréhension, surtout s'il s'agit d'un texte de nature technique.

Ces indications sont précieuses car le texte s'inscrit dans des pratiques sociales déterminées. Il a une dimension fortement collective : c'est le contexte et l'existence de pratiques partagées qui permet l'interprétation (on retrouve donc ici la notion d'inter-subjectivité déjà entrevue précédemment, cf. chapitre 1, page 15) : comprendre un texte, c'est savoir activer les règles implicites qui en commandent l'interprétation [Condamines, 2003]. A ce stade, nous pouvons faire nôtre la définition donnée par Rastier [2001, 303] :

> *Un texte est une suite linguistique autonome (orale ou écrite) constituant une unité empirique et produite par un ou plusieurs énonciateurs dans des pratiques sociales attestées. Les textes sont l'objet de la linguistique.*

Le support informatique renforce le caractère dynamique du texte, à la fois au niveau de la production et de l'accès. Il est dorénavant possible d'offrir des parcours de lecture actif, non linéaires et guidés par le but. Enfin, le web permet depuis quelques temps un face-à-face plus direct entre l'auteur et le lecteur (notamment du fait de la généralisation de fonctionnalités interactives autrefois peu employées mais aujourd'hui mises en avant à travers le web 2.0). L'auteur se rapproche de plus en plus du lecteur, les lecteurs réagissent entre eux[63] et les nouveaux supports permettent un rapprochement entre « producteurs » et « consommateurs » d'écrit.

L'écrit s'est considérablement renouvelé ces dernières années, à travers notamment la notion de « document numérique ». La notion de document dépasse le texte : le document délimite le texte et lui attribue une cohérence informationnelle. Les aspects numériques permettent de gérer sur un même support des données possiblement hétérogènes, comme du texte avec des images ou du son ; dans le même temps, il est possible de composer dynamiquement des documents virtuels, au sens où ils sont composés dynamiquement

[62] Dans le *Phèdre*, Socrate explique que « *l'écriture a, tout comme la peinture, un grave inconvénient. Les œuvres picturales paraissent comme vivantes ; mais, si tu les interroges, elles gardent un vénérable silence. Il en est de même des discours écrits. Tu croirais certes qu'ils parlent comme des personnes sensées ; mais, si tu veux leur demander de t'expliquer ce qu'ils disent, ils te répondent toujours la même chose. Une fois écrit, tout discours roule de tous côtés ; il tombe aussi bien chez ceux qui le comprennent que chez ceux pour lesquels il est sans intérêt ; il ne sait point à qui il faut parler, ni avec qui il est bon de se taire. S'il se voit méprisé ou injustement injurié, il a toujours besoin du secours de son père, car il n'est pas par lui-même capable de se défendre ni de se secourir.* » Platon, *Phèdre*, **275**.

[63] Rien de très nouveau à cela, que l'on pense aux salons littéraires autrefois !

à partir d'éléments hétérogènes (dans ce mémoire, nous ne considérons que les aspects textuels). Les rapports entre texte et document ont donc été considérablement renouvelés grâce aux techniques numériques, qui imposent de reconsidérer la relation entre le fond et la forme, cf. les réflexions du collectif RTP-DOC [2005, 2006,2007].

C'est dans ce contexte que nous nous sommes interrogé sur l'extension des travaux présentés dans la première partie de ce chapitre à des textes différents des seuls guides de bonnes pratiques. Cela implique de s'interroger sur la notion de genre textuel : comment identifier automatiquement des textes qui soient proches des GBP ? Peut-on se contenter de considérer le niveau documentaire (sommairement matérialisé par un fichier informatique par exemple) ou faut-il considérer des parties de texte plus fines ?

4.2.2 Traitements automatiques et genres textuels

Le fait qu'un texte puisse être corrélé à une pratique sociale laisse penser que ce texte obéit à des contraintes linguistiques, stylistiques et matérielles. On peut penser que ces contraintes pourraient permettre d'identifier un texte comme appartenant à un genre donné [Rastier, 2001] [Adam, 2005]. Il arrive cependant fréquemment que les auteurs brouillent les pistes et « étiquettent » une œuvre d'une façon qui semble contradictoire avec son contenu. Ce brouillage des genres, courant pour les textes littéraires, est aussi à l'œuvre pour les textes techniques. Ce brouillage n'est pas obligatoirement volontaire mais il est le propre de tout écrit, même fortement stéréotypé. Nous y reviendrons.

La réflexion sur les genres a tout d'abord porté de façon privilégiée sur les textes littéraires et leurs subdivisions en comédie, tragédie, épopée, roman, *etc.* Les travaux en TAL relèvent quant à eux de deux démarches distinctes suivant que l'on considère que l'ensemble des classes est donné *a priori* ou est trouvé *a posteriori*. Dans le domaine de l'apprentissage, une terminologie particulière est définie pour ces deux démarches : on parle de catégorisation quand les classes sont prédéfinies, et de classification (ou de *clustering*) pour l'induction de classes.

La première démarche pose comme hypothèse l'existence préalable d'une classification des textes en genres. L'idée est d'explorer si des genres prédéfinis se différencient par un ensemble de traits linguistiques donné, et d'envisager parallèlement le potentiel classificatoire de sous-ensembles de traits afin de différencier des familles de textes plus fines à l'intérieur d'un genre. Par exemple, D. Malrieu et F. Rastier [2001] traitent un corpus de 2541 ouvrages couvrant quatre discours (littéraire, scientifique, juridique, essayiste) et dix-sept genres. Ils utilisent 25 variables allant du niveau le plus global (le texte ou le paragraphe) jusqu'au niveau le plus fin (temps et personnes des verbes, types d'adjectifs et même ponctuation) et font émerger, grâce à une analyse statistique multivariée, différents facteurs caractéristiques du genre narratif : poids des pronoms personnels, de la première personne du singulier ou du présent par exemple.

La deuxième démarche est inductive, c'est-à-dire qu'elle tente, à partir de textes quelconques fournis, de faire émerger des types de textes à partir d'observations concernant le partage de certains traits linguistiques (ou au contraire l'évitement d'autres traits).

D. Biber [1988] part de 67 traits linguistiques, grammaticaux et lexicaux (marqueurs de temps et d'aspect, adverbes de lieu, de temps, passifs, modaux, négation...) et d'un corpus composé de textes issus d'une quinzaine de genres différents (reportages, conversations, articles de recherche...). Il utilise la statistique multidimensionnelle pour mettre en évidence des corrélations, positives et négatives, entre ces traits et des techniques de classification automatique pour regrouper les documents. L'auteur constate alors que les regroupements effectués ne correspondent pas aux genres distingués au préalable.

Pour Biber, ce n'est pas une surprise puisque ce contraste correspond aux deux façons complémentaires d'étudier l'espace textuel d'une langue ; d'une part, une étude des « registres » de textes correspond à des classes de textes déterminées par les situations dans lesquelles ces textes sont émis : reportage journalistique, conversation, cours, article scientifique, lettre personnelle[64] ; d'autre part, une étude des types de textes relève, elle, de la seule analyse linguistique [Biber, 1995] (voir Folch *et al.* [2000] et Beaudouin *et al.* [2001] pour des expériences très similaires à partir de sites web marchands et personnels).

La difficulté de ces approches est qu'elles utilisent les procédures statistiques comme des boîtes noires. A ce titre, leurs sorties doivent nécessairement être interprétées : les facteurs (ou groupes de traits ou dimensions) trouvées par l'analyse factorielle ou les classes trouvées par la classification automatique doivent être, en retour, confrontés aux textes et à leurs représentations.

4.2.3 Élargir l'étude à d'autres types de textes procéduraux

Comme on l'a vu, nous avons d'abord étudié un genre de textes particulier : les guides de bonnes pratiques cliniques. Nous souhaitons à présent identifier des types de documents proches, afin de pouvoir appliquer notre analyse au-delà des GBP. Les outils d'accès au contenu gagneront en effet à compléter l'information acquises à partir des GBP par d'autres sources de connaissances, mais cela implique que les techniques d'analyse mises en place gardent un taux de réussite satisfaisant. Or, nous faisons l'hypothèse que les traitements dépendent largement du genre de texte envisagé.

L'intuition pousse à vouloir identifier des genres correspondant à des pratiques données. Comme le souligne Rastier [2001, 228-229] « à chaque type de pratique sociale correspond un domaine sémantique et un discours qui l'articule ». De fait, le genre est « doublement médiateur » dans la mesure où il « assume non seulement le lien entre le texte et le discours, mais aussi entre le texte et la situation, tels qu'ils sont unis dans une pratique » (cf. chapitre 1, page 16).

Le constat de Rastier permet de comprendre pourquoi on ne peut établir une liste de genres prédéfinie, dans la mesure où il n'existe pas de liste de pratiques sociales, exhaustive et finie, posée une fois pour toutes. Les démarches inductives, vers lesquelles il peut sembler logique de se tourner, posent de nombreux autres problèmes : on ne sait

[64] La notion de registre ne recouvre pas tout à fait celle de genre, mais les deux notions ont en commun de se fonder sur des éléments extérieurs au discours. Le registre réfère aux propriétés situationnelles des textes tandis que le genre correspond à une étiquette généralement attribuée par l'auteur ou l'éditeur.

pas quels sont les traits pertinents pour la définition des genres. Même si on disposait d'une telle liste de traits, leur corrélation est problématique, ne serait-ce que parce qu'ils sont de nature hétérogène. Enfin, rien ne dit que le genre est réductible à un ensemble de traits linguistiques, encore moins à des traits suffisamment simples pour être repérés de manière automatique[65].

Il nous a donc semblé nécessaire de « faire un pas de côté ». Comme nous l'avons vu, au lieu de s'engager directement dans une démarche typologique, nous avons préféré partir d'un ensemble de textes fortement contraints, voire stéréotypés, pour en identifier les principales caractéristiques.

La suite de l'étude consiste maintenant à élargir le corpus, en ayant recours à d'autres textes du domaine médical plus variés, ne contenant pas que des recommandations. Nous nous sommes donc penché sur différents textes du domaine médical (à travers notamment le corpus médical EQueR — il s'agit d'un corpus mis au point pour une campagne d'évaluation de systèmes de questions-réponses [Ayache *et al.*, 2006] ; la campagne d'évaluation comportait en fait deux parties : une portant sur la langue générale et une autre sur une langue de spécialité — la médecine. Seule cette deuxième partie est utilisée dans la suite de ce chapitre). Il est alors possible de voir dans quelle mesure les marques procédurales restent les mêmes et dans quelle mesure elles varient.

Soulignons enfin la visée pratique de l'étude : il ne s'agit pas de constituer des regroupements de textes dont la pertinence est jugée par introspection. La modélisation des guides de bonnes pratiques est une tâche intéressant fortement les autorités en matière de santé. L'identification de séquences procédurales au sein d'un corpus plus large comme EQueR est aussi une tâche pertinente, notamment pour les systèmes de questions-réponses. Cette recherche fait partie du projet TEXTCOOP dont le but est de développer un système répondant mieux aux questions en « *comment* ».

Dans cette section, nous avons admis sans discussion que les Guides de Bonnes Pratiques étaient des textes procéduraux. Ceux-ci correspondent bien à la définition commune des textes procéduraux, dans la mesure où ils visent à indiquer au médecin quelle démarche entreprendre en fonction d'une pathologie donnée. Il semble dès lors intéressant et pertinent de compléter l'analyse par d'autres documents ayant la même finalité.

4.3 Repérage de séquences procédurales au-delà des GBP

Le projet ANR RNTL TEXTCOOP (2006–2008) est une tentative pour étendre l'analyse effectuée au-delà des GBP (et même au-delà du domaine médical). Comme on l'a vu en introduction, il existe des besoins avérés en matière de consultation de documentation

[65] L'humain se fonde sur des critères linguistiques et non linguistiques pour identifier les genres : conditions de production, de réception des documents, indications péri-textuelles et métadonnées, *etc.* Il faudrait par ailleurs s'interroger sur la nature du genre du point de vue de la réception des textes. Dans cette perspective, A. Condamines [2003] parle de « genre interprétatif ».

électronique. Les utilisateurs souhaitent avoir accès directement au contenu pertinent, en disposant d'un système capable de répondre à des questions de type « *comment faire... ?* ». L'analyse des procédures et le repérage préalable de documents procéduraux semblent donc très pertinents dans ce cadre.

Il faut donc d'abord mieux cerner la notion de procéduralité. On peut dire de manière intuitive qu'un document procédural (ou « injonctif » ou « à consigne ») a une visée fonctionnelle [Péry-Woodley, 2001], il détaille en général des manières de procéder, des méthodes et des marches à suivre pour obtenir un résultat. Comme le souligne L. Heurley [1997], cette catégorie regroupe, notamment, « les modes d'emploi, les notices explicatives, les manuels et les guides d'utilisation, les consignes de sécurité, les recettes de cuisine, les *do-lists* utilisées en aéronautique, *etc.* ».

Une étude du corpus EQueR, menée principalement par M.-P. Jacques, permet de trouver assez facilement des extraits prototypiquement procéduraux.

Recommandations pratiques de la WGO : Vaccination contre l'hépatite B

Test

Si vous voyez un vaccin HBV congelé, alors il est endommagé. Cependant un vaccin peut avoir été congelé antérieurement puis décongelé. Le test suivant peut être utilisé pour vérifier si un vaccin a été endommagé par une congélation antérieure.

– Comparer le vaccin que vous suspectez avoir été congelé puis décongelé, avec un autre, produit par le même fabriquant mais dont vous êtes sûr qu'il n'a jamais été congelé.
– Secouez les ampoules de vaccin.
– Regardez soigneusement les contenus.
– Garder les vaccins côte à côte pendant 15-30 minutes pour laisser au sédiment le temps de se déposer.
– Ne pas utiliser celui où un sédiment est déposé sous un liquide presque clair.

FIG. 4.3 – Extrait du fichier `tc0010080153bis-cismef-rpc-tei.xml`, corpus EQueR

Le texte de la figure 4.3, parce qu'il comporte une liste d'items, parce que chaque élément de la liste commence par un impératif ou un infinitif (de manière non homogène) et parce que le titre global inclut le terme *recommandation*, peut assez facilement être étiqueté comme « procédural ». Chacun de ces indices n'est pas suffisant en soi, mais leur caractère simultané permet le repérage par le système développé pour l'analyse des GBP.

Mais même dans ce cas relativement clair, il faut constater la description préliminaire (la situation donnée), qui permet de "situer" la procédure. Du coup, on peut s'interroger : qu'est-ce qu'un texte procédural ? Quelles caractéristiques sont nécessaires, suffisantes ou obligatoires pour étiqueter un texte avec un genre donné ?

4.3.1 Genre, type et fonction discursive

Comme le niveau textuel est hétérogène, il est tentant de se pencher, à un niveau plus fin, sur des séquences échappant au moins partiellement à l'hétérogénéité du texte pris

globalement. Ceci nous amène à distinguer, en plus des genres et des types, la notion de ***fonction discursive***. La fonction discursive désigne la visée pragmatique d'un segment textuel. Par exemple, un segment textuel peut décrire la façon d'accomplir une action. On parlera alors d'un fragment dont la visée est « procédurale », le caractère procédural correspondant à une fonction discursive [Jacques *et al.*, 2008].

En « entrant » à l'intérieur du document, on prend mieux en compte sa richesse, les différents points de vue qui s'y expriment (ce qui, chez Bakhtine [1978], correspond partiellement à la notion de polyphonie textuelle).

La caractérisation de la notion de procéduralité que nous avons donnée est uniquement fondée sur la visée fonctionnelle des textes ou des segments de texte[66]. Il importe à présent de vérifier si cette notion est facilement identifiable dans les textes, si des personnes différentes disposent de jugements intuitifs et fiables sur le caractère procédural ou non d'un segment de texte. Ceci nous permettra, en cas d'accord, d'identifier d'éventuels marqueurs linguistiques pertinents.

4.3.2 Étude manuelle du corpus : variations sur la procéduralité

Le projet TEXTCOOP vise, comme on l'a dit, à améliorer la qualité des réponses fournies automatiquement à des questions en « comment ». Vu l'ampleur et la diversité du problème, l'annotation a porté sur différents documents, provenant de différents domaines. Avant d'en venir à l'annotation manuelle du corpus EQueR, jetons un coup d'oeil rapide aux autres domaines.

Une partie de l'annotation a porté sur des textes très stéréotypés qui ne faisaient pas partie du domaine médical. Il est évident que des fiches bricolages, des instructions pour jeux vidéo ou des recettes de cuisine constituent des familles de documents homogènes, qui se prêtent bien à une annotation globale avec un bon taux d'accord entre annotateurs.

Dans le domaine médical, la notion de genre est davantage brouillée. Il existe par exemple des zones correspondant à l'établissement du diagnostic. Tantôt cette zone contient des instructions sur des actions à effectuer ; tantôt il s'agit d'une simple liste de signes à examiner. Cette liste se limite assez souvent à un résumé, sans verbe et sans qu'il y ait une action claire sous entendue. D'un autre côté, on peut toujours assimiler le diagnostic à un ensemble d'action, quand bien même celles-ci se borneraient à observer une série de signe.

D'autres textes sont plus problématiques encore, au premier rang desquels les normes. Celles-ci sont parfois réellement injonctives quant à la manière d'aborder un problème ;

[66] Nous reprenons en ce sens la démarche de M.-P. Péry-Woodley [2001] : « Qu'appelle-t-on discours procédural, texte injonctif, texte à consignes ? Plutôt que de tenter une définition *in abstracto*, je poserai ici la question en référence à un projet particulier : la constitution d'un corpus de textes à partir desquels examiner les réalisations spécifiques des consignes. Impossible de prendre pour critère de sélection des traits linguistiques particuliers, présence d'impératifs, par exemple : cela reviendrait à répondre *a priori* à la question posée et exclurait toute possibilité de découverte de formulations inattendues. Il est clair que pour que l'examen des réalisations linguistiques soit réellement ouvert, les textes doivent être sélectionnés sur des bases autres que des critères linguistiques. »

à d'autres moments il s'agit de textes descriptifs ou prescriptifs mais n'incitant pas à l'action immédiate. Que dire d'un texte qui décrit la structure d'une salle de réveil ? Une partie de ce texte peut inciter à l'action (quand il s'agit d'un manque ponctuel facilement réparable par exemple) mais ce n'est pas toujours le cas. Dans ce cas, la visée fonctionnelle est un peu brouillée.

Afin de cerner les difficultés observées, une annotation de corpus à plusieurs mains a été effectuée. En l'absence de consignes claires, la tâche d'annotation est fortement subjective. Certains cas ne posent pas de problème mais une assez forte minorité de cas dépend très fortement de ce qui est « projeté » par le lecteur, par sa lecture forcément subjective du texte. On retrouve ici des éléments vus à plusieurs reprises (par exemple, au chapitre 2, pour les glissements de sens des entités nommées) : le jugement dépend d'une pratique, il concorde assez bien sur les cas typiques mais achoppe sur les autres. On retrouve ici la notion d'« air de famille » et de prototype (cf. chapitre 1, page 16).

L'annotation à plusieurs mains permet d'identifier les cas problématiques, de donner des consignes afin de parvenir à une annotation plus stable d'un annotateur à l'autre. Même si nous avons souhaité partir sans spécifier la tâche *a priori* au niveau linguistique, comme le suggère M.-P. Péry-Woodley [2001], il est ensuite nécessaire de fournir des indications précises aux annotateurs. Ce travail de guidage est un problème ouvert, mais il a été entamé dans le cadre de TEXTCOOP à travers la rédaction des bases d'un guide d'annotation. La stratégie consiste évidemment à se fonder autant que possible sur des éléments objectifs, marqueurs linguistiques repérables et présents dans les textes[67].

4.3.3 Discussion et perspectives : vers un repérage automatique ?

Une fois faite l'annotation manuelle de corpus, des techniques d'analyse classiques peuvent être utilisées, afin d'identifier les éléments pertinents (linguistiques, dispositionnels, *etc.*), leur poids respectif en fonction de leur contribution à la tâche et les relations qu'ils entretiennent entre eux. Ce repérage peut se faire entièrement automatiquement par apprentissage, en calculant les traits les plus caractéristiques.

Dans le cadre du projet TEXTCOOP, un groupe d'étudiants Sup'Galilée (formation d'ingénieur en informatique de l'Université Paris 13), sous la direction de Marie-Paule Jacques et Françoise Gayral, a développé un logiciel permettant de reconnaître certaines expressions pertinentes pour la tâche, afin d'identifier des passages procéduraux. L'implémentation de la partie linguistique se fait au moyen de variables, évaluées sur un corpus d'entraînement, seules ou en combinaison. Il s'agit ainsi de déterminer les configurations les plus caractéristiques de la fonction procédurale. Enfin, comme la thèse d'A. Bouffier

[67] Ainsi, en médecine, les opérations de diagnostic se fondent ainsi sur des listes d'indices : doit-on assimiler ces listes à un constat ou à une série d'actions à effectuer ? Un verbe comme *vérifier* peut être présent ou non, rendant l'incitation plus ou moins explicite. Malgré son caractère arbitraire, le choix finalement retenu a été de favoriser au maximum les éléments présents dans le texte. S'il n'y a pas d'action explicite marquée par un verbe ou un nom, alors la séquence ne doit pas être marquée comme procédurale. Ce choix permet de largement limiter les cas de désaccord et la résolution ultérieure des conflits entre annotateurs.

le montre à l'envi, la structure visuelle joue un rôle important [Bouffier, 2008]. L'étude de l'interaction de celle-ci avec d'autres marqueurs linguistiques est également étudiée.

Il faut remarquer que l'on retombe, à ce point, sur le problème identifié au début de ce chapitre : il est très difficile d'évaluer la pertinence et le poids des éléments retrouvés, et ce particulièrement quand on travaille sur des corpus variés. Les expériences menées sur le domaine médical donnent des résultats mitigés : d'un côté on obtient des classes sémantiques, des marqueurs qui semblent pertinents par rapport au corpus de travail, mais la validité des analyses menace toujours d'être remise en cause si l'on a affaire à des documents par trop différents.

On entrevoit ici les mêmes interrogations que celles qui ont été posées auparavant (cf. chapitre 1, 16). Une fois les catégories traditionnelles écartées, une fois que l'on s'est éloigné des éléments les plus caractéristiques, voire les plus caricaturaux, jusqu'à quel point peut-on classer les textes ? Constituer des classes homogènes utiles au traitement ? La perte des catégories traditionnelles et du confort de l'approche ontologisante s'accompagne obligatoirement d'un flottement sur la nature des résultats et les possibilités d'évaluation.

Ce chapitre tout comme les autres met donc en lumière la question du partage des rôles entre catégories pré-établies (liste de genres) et induction à partir de corpus (identification dynamique de types). Dans bien des cas, l'étude de l'apport respectif de ces deux pôles et de leur complémentarité reste à faire : elle est difficile dans la mesure où elle dépend partiellement du contexte et de facteurs extérieurs au texte, difficiles à modéliser. Nous donnons cependant quelques pistes dans la conclusion.

4.4 Synthèse

Cette section a permis d'aborder la question du repérage et de la modélisation de types de textes particuliers. Les GBP, documents médicaux très réguliers, peuvent ainsi être analysés au moyen de l'analyse de structures discursives typiques, s'étendant au-delà de la phrase. Ce chapitre a ensuite montré la difficulté pour étendre ce type d'analyse à d'autres types de documents, dans la mesure où les notions de type et de genre ne sont pas stabilisées et subissent de grandes variations à l'intérieur de frontières mouvantes.

Conclusion

A l'issue de cette étude, je dresse un bilan des expériences menées avant de donner quelques pistes permettant de poursuivre ces travaux.

1 Bilan des réalisations

Ce mémoire a permis de présenter mes recherches autour de la compréhension de textes. Pour reprendre une terminologie employée dans l'introduction et empruntée à Rastier [1994], j'ai détaillé des analyses correspondant aux niveaux suivants :

- le niveau microsémantique avec l'analyse des entités nommées,
- le niveau mésosémantique avec le repérage de schémas prédicatifs,
- enfin le niveau macrosémantique avec l'analyse de séquences procédurales dans les textes.

Les analyses mises en œuvre partagent toutes une même approche : mise au point d'un système à base de règles puis extension avec un processus d'adaptation dynamique. Pour les entités nommées, le repérage de contextes discriminants permet l'analyse de séquences inconnues ou le retypage de séquences mal analysées. Pour les schémas prédicatifs, l'analyse de gros corpus permet d'obtenir une modélisation fine du phénomène traité sous forme de distributions de probabilités. Pour les séquences procédurales, l'analyse par défaut peut être remise en cause en fonction de contextes particuliers.

Ces analyses ont débouché sur des réalisations pratiques, dans un cadre industriel lors de ma thèse puis à travers des projets exploratoires ou pré-compétitifs (souvent en relation avec des industriels), par la suite. On a pu le voir par exemple avec les travaux de thèse d'Amanda Bouffier, qui visent à automatiser la modélisation de documents médicaux précis. Les cadres applicatifs de ce type garantissent l'intérêt pratique de la réalisation (à ce sujet, on pourra se reporter aux analyses de Simondon, autour de la notion d'*objet technique* [Simondon, 2001 (1958)]). Les évaluations et les confrontations avec des experts des domaines traités permettent en outre de valider les approches adoptées, les résultats obtenus et l'utilisabilité des solutions développées.

Mais, au-delà de l'intérêt applicatif, il importe de garder en tête une perspective critique, sur ce que ces applications nous enseignent sur la langue elle-même. J'ai essayé de montrer, à chaque fois, les limites des traitements proposés. L'analyse des entités nommées

s'inscrit dans une sémantique référentielle mais ce cadre pose problème dès que l'on a affaire à des phénomènes de glissements de sens (comme la métonymie ou la métaphore). La modélisation des contraintes de sous-catégorisation par des distributions de probabilités offre un cadre intéressant mais l'acquisition à partir de corpus ne permet pas de relever les constructions rares. Enfin, l'analyse des textes échoue sur la question des typologies, dont l'utilité semble évidente (le type de texte ayant une influence sur son contenu et sa structure) mais la caractérisation difficile.

2 « *Les sortilèges du langage* »

De nombreux auteurs s'intéressant à l'intelligence artificielle ou, plus spécifiquement, aux systèmes de compréhension de textes, ont pointé les limites des applications développées, leur manque de généralisation et ont essayé de saisir « le sens profond des difficultés rencontrées » [Winograd, 1990; Dreyfus, 1965, 1992]. J'ai moi-même essayé de souligner à plusieurs reprises dans ce mémoire les limites des solutions développées, voire l'inadéquation des modèles proposés, au-delà de l'application visée. On peut généraliser ces réflexions à travers les points suivants :

- *la complexité des phénomènes envisagés* : on a vu dès l'introduction que la compréhension demande des connaissances quasi-infinies, mobilisables en contexte et adaptables dynamiquement. Même si les techniques d'apprentissage automatique ont permis des avancées, celles-ci restent largement en deçà des besoins requis dans le processus de compréhension.
- *la méconnaissance des processus cognitifs impliqués* : les problèmes complexes — au premier rang desquels la compréhension — ne sauraient se résoudre par une simple atomisation jusqu'à des primitives (sémantiques ou non) en nombre fini. Les traitements sont encore trop pauvres pour rendre compte du raisonnement par analogie [Yvon, 2006], des phénomènes de catégorisation, et plus généralement de la possibilité de saisir du semblable dans des cadres variées.
- *l'inadéquation des outils de modélisation utilisés* : ce point est largement lié aux deux précédents. La complexité des données et des mécanismes de raisonnement impliqués dépasse largement le cadre de la logique, y compris des logiques non classiques contemporaines.

Ces observations ne sont pas sans rapport avec une partie des arguments entrevus chez Wittgenstein, quand celui-ci étudie le langage, ou plutôt la difficulté de la philosophie à saisir le monde du fait des « *sortilèges du langage* », pour reprendre le mot de Bouveresse [2003]. Wittgenstein ne s'intéresse pas en soi à la complexité du phénomène de compréhension ou aux processus cognitifs impliqués, mais il montre dans ses écrits le fait que les règles du langage ne peuvent être fixées, car elles dépendent de l'usage : elles sont donc en partie productives et inconscientes (cf. chapitre 1).

On comprend bien alors pourquoi la compréhension automatique a fait relativement peu de progrès en cinquante ans. Il est possible de mettre au point des applications relativement efficaces pour des domaines limités mais on continue de buter sur les mêmes

difficultés pour modéliser de nombreux phénomènes linguistiques et les étendre au-delà de domaines de spécialité. La complexité est telle que les analyseurs échouent sur des phénomènes très courants, là où un être humain ne détecte même pas d'ambiguïté, comme on l'a vu dès l'introduction de ce mémoire. Au-delà de faits bien circonscrits, les taux d'erreurs augmentent et l'on voit en filigrane l'inadéquation et l'arbitraire des modélisations adoptées. Elles ont notamment beaucoup de mal à saisir les glissements de sens et le caractère flou de certaines catégories.

3 Perspectives : le linguiste, l'ingénieur et l'alchimiste

Il est relativement aisé de donner des perspectives à court et moyen terme. De nombreuses techniques, déjà entrevues pour certaines d'entre elles au cours de l'exposé, permettraient d'améliorer les solutions existantes ou de proposer de nouveaux développements. Il est en revanche beaucoup plus difficile de se projeter au-delà : il semble en effet qu'un travail de fond doive être entrepris pour définir un nouveau cadre théorique, dépassant les limites énumérées ci-dessus. Ce cadre passe, à mon avis, par une meilleure collaboration avec l'utilisateur et par une meilleure connaissance des mécanismes de la compréhension. On se condamne sinon à travailler en aveugle et sans théorie, à l'image des alchimistes d'antan.

3.1 Des améliorations possibles à court terme

Nous présentons ici quelques principes permettant des avancées à court terme.

Améliorer les techniques d'acquisition de ressources

Les travaux présentés dans le chapitre 3 de ce mémoire sont encore très préliminaires. Au-delà de la sous-catégorisation syntaxique, il importe d'acquérir des informations plus riches sur les verbes, ainsi que des classes de comportement sémantique et des informations sur les arguments.

Ce type de recherche a plusieurs buts. Un but pratique tout d'abord : permettre de produire des ressources à moindre coût, moins précises que des ressources développées manuellement, mais plus facilement adaptables à des corpus variés. Différentes expériences, pour d'autres langues que le français, ont montré l'intérêt de ces techniques d'adaptation pour des tâches et des domaines variés [Schulte im Walde, 2006].

Un but théorique ensuite : explorer l'acquisition de connaissances sur la langue à partir des seules réalisations de surface. Il s'agit d'explorer les limites d'une telle expérience, d'identifier les informations qui résistent à l'acquisition à partir de corpus et de réfléchir en retour sur la façon dont les humains acquièrent ces informations. Plusieurs expériences ont ainsi été faites pour vérifier les hypothèses de Pinker [1989] ou Levin [1993] sur l'interface syntaxe-sémantique, mais il s'agit là d'un terrain encore largement en friche.

Enfin, la multiplicité des expériences dans différentes langues doit permettre la comparaison des résultats obtenus, en termes de classes de comportement, syntaxique mais aussi sémantique. Comment des langues de familles différentes se comportent-elles d'un point de vue syntaxico-sémantique ? C'est une perspective qui nous semble très riche et qui peut faire avancer les recherches sur la typologie et les universaux linguistiques.

Adapter dynamiquement les stratégies d'analyse

L'analyse des entités nommées (chapitre 2) a montré comment un système à base de règles, lisible et facilement adaptable était susceptible d'être complété par un système d'apprentissage, permettant d'étendre l'analyse initiale. L'analyse des séquences conditions-recommandations dans les Guides de Bonnes Pratiques (chapitre 4) a également montré l'apport d'un traitement par défaut, fondé sur une norme, qui peut éventuellement être remis en cause en fonction du contexte.

Il faut étendre ce type d'architectures à d'autres problèmes que ceux évoqués ci-dessus. La notion de norme a été en définitive assez peu exploitée dans ce mémoire mais d'autres ont montré la potentialité de cette notion pour modéliser la compréhension et le raisonnement [Kayser et Nouioua, 2005]. Il est ainsi possible de modéliser des règles par défaut qui peuvent être remises en cause. Ces techniques de modélisation et de programmation permettent d'obtenir des systèmes dynamiques, susceptibles de s'adapter au contexte.

Ces notions poussent également à intégrer des techniques de programmations dynamiques, ou d'apprentissage au sein des systèmes même, et pas simplement pour l'acquisition de connaissances. Les architectures informatiques devront évoluer pour permettre la prise en compte d'esquisses [Sabah, 1996], c'est-à-dire de résultats partiels, affinés progressivement au cours de l'analyse.

Redonner la main à l'utilisateur

Alors que le TAL a cherché a concevoir des systèmes complètement automatiques pendant des années, on assiste ces deniers temps au retour de la prise en compte de l'utilisateur et de l'expert. Cela se manifeste d'abord par la mise au point de nouvelles techniques d'analyse interactive sollicitant l'expert de façon minimale mais pertinente. Ensuite par le développement de campagnes d'évaluation tenant compte de jugements humains, au-delà de ce qui peut être calculé automatiquement par la machine.

On a vu au chapitre 3 que Asium avait été choisi essentiellement en raison de son caractère interactif. De nouvelles techniques d'apprentissage permettent de focaliser encore davantage l'entraînement sur des exemples pertinents. Classiquement l'apprentissage s'intéresse au cas où les exemples sont indépendants identiquement distribués. En revanche, comme le dit J. Mary [2005], dans l'apprentissage actif, l'algorithme d'apprentissage a la possibilité d'influer sur la distribution des exemples plutôt que de « subir » la distribution naturelle. Ce type de techniques peut facilement s'appliquer à la plupart des problèmes décrits dans ce mémoire.

Les méthodes d'évaluation elles-mêmes prennent de plus en plus en compte les facteurs humains. L'organisation récente d'un atelier sur la question lors d'une conférence internationale[68] est révélateur de ce retour en grâce, après des années où le tout automatique a souvent été privilégié. Nombre de ressources sont discutables quant à leur contenu et nombre d'expériences ont pu paraître artificielle faute d'avoir été proposées sous couvert de besoins réels. Mais pour être complet, il faut noter à l'inverse, que des résultats intermédiaires peuvent paraître étranges mais donner *in fine* de fort bons résultats. Remettre l'homme « dans la boucle », pour reprendre une expression de C. Pierce et Cardie [2001], c'est donc une vraie problématique, plus complexe qu'il ne semble au premier abord.

3.2 Une réflexion à mener sur le long terme

Plusieurs auteurs ont remis en cause la possibilité de modéliser les phénomènes liés à la cognition en dehors de modèles situés et embarqués, dans la mesure où les mécanismes d'apprentissage et de compréhension sont étroitement liés à notre expérience du monde extérieur (voir par ex. Lakoff et Johnson [1999]). Deux pistes nous semblent intéressantes pour répondre à ces objections. D'une part, la conception de nouveaux modes d'accès au texte, conçus comme des moyens d'aide à l'interprétation, qui n'excluent pas l'utilisateur mais, au contraire, visent à l'assister. D'autre part, l'étude des mécanismes cognitifs impliqués dans le processus de compréhension, non pas pour les reproduire à l'identique mais pour mieux identifier les contraintes à prendre en compte et à modéliser.

Mieux prendre en compte la dimension interprétative

L'homme aura, pour longtemps sinon pour toujours, une faculté de création, de jugement et d'initiative que la machine n'a pas. On a eu l'occasion de le voir à maintes reprises : face au texte, il manque à la machine la capacité d'interprétation, c'est-à-dire la capacité de projeter un ensemble de significations au-delà des mots. Sauf dans des cadres fort limités, cette capacité n'est pas modélisable en l'état de nos connaissances car elle ne procède pas d'une réduction du texte à un ensemble d'atomes de sens[69] [Dreyfus, 1992].

Une première piste de recherches consiste alors à inscrire l'analyse dans la matière textuelle elle-même. Allant dans ce sens, Rastier a proposé une théorie originale à travers la sémantique différentielle[70], où les nuances de signification s'expriment par oppositions entre unités, du mot au texte [Rastier, 2001]. Ces nuances de signification sont discernables par les contextes variables d'apparition des unités signifiantes, en partie appréhendables par la machine. Le travail d'interprétation reste une tâche dévolue au lecteur ; il s'agit avant tout d'un processus interne au texte lui-même.

[68] Atelier "Human Judgements in Computational Linguistics", lors de COLING 2008 à Manchester (`http://workshops.inf.ed.ac.uk/hjcl/`).

[69] Contrairement à la position défendue au sein de la communauté web sémantique, renommé récemment *web des données*, où le terme de *donnée* vise explicitant des morceaux de connaissance échappant à l'arbitraire interprétatif [Rastier, 2008].

[70] Même si la sémantique de Rastier s'inscrit dans une tradition où Saussure et Hjelmslev, Pottier et Greimas figurent en bonne place.

Au-delà, d'autres chercheurs ont proposé de mieux prendre en compte les intentions de l'auteur et l'interprétation du lecteur [Hirst, 2007]. Si la réalisation de tels systèmes reste encore largement hypothétique, ce sont des objectifs réalisables de manière interactive. J'ai été récemment impliqué dans des projets autour de la numérisation et de la mise à disposition de collections de textes historiques, philosophiques ou littéraires (corpus des *Grammatici Latini* au Laboratoire d'Histoire des Théories Linguistiques coordonné par Alessandro Garcea[71], projet Discovery coordonné par P. d'Iorio autour de l'œuvre de plusieurs philosophes[72], *etc.*). Dans ce cadre, une simple indexation n'a pas de sens si elle ne restitue pas le contexte, les nuances de sens, voire les diverses interprétations. La dimension interprétative ne peut être ignorée, en relation avec les spécialistes de l'œuvre concernée (cf. le projet Scholarweb, dans l'annexe).

Mieux connaître les processus cognitifs impliqués

Sur le long terme, une meilleure connaissance des processus cognitifs en jeu semble également nécessaire. Il ne s'agit pas de revenir aux premiers temps de l'intelligence artificielle en visant à reproduire directement par la machine certains comportements humains. Cette vision a été à juste titre critiquée et elle n'a pas donnée de résultats probants [Sabah, 2006].

Cependant, il semble que les limites des systèmes actuels ne peuvent pas être dépassées si l'on reste dans le paradigme de recherche courant. Au-delà d'améliorations techniques et ponctuelles, il semble clair que les processus d'acquisition de connaissances et de confrontation avec le monde supposent des capacités cognitives que ne peuvent atteindre les machines. Il est important de mieux connaître ces mécanismes pour percevoir les limites du cadre actuel.

Ce programme de recherche est ambitieux et pluridisciplinaire ; il dépasse largement le cadre de ce mémoire. Les recherches autour des sciences cognitives — mêlant linguistique, psychologie, informatique, *etc.* — permettent d'explorer ce terrain encore en friche [Tomasello, 2005]. Ma participation à des projets autour de l'acquisition du langage ou de l'étude des processus cognitifs en jeu s'inscrit dans ce cadre (voir par exemple le projet ANR Léonard autour de l'acquisition du langage chez l'enfant[73], ou encore le projet *Computational Natural Language Processing and the Neuro-Cognition of Language*, à l'Université de Cambridge [Buttery *et al.*, 2007]).

A l'issue de cette étude, on se trouve en quelque sorte dans la situation décrite par Winograd [1977] dans la citation en exergue de ce mémoire. A l'image des alchimistes, les recherches en traitement des langues ont permis de développer des systèmes com-

[71] `http://kaali.linguist.jussieu.fr/CGL`
[72] `http://www.discovery-project.eu/`
[73] `http://anr-leonard.ens-lsh.fr/`

plexes ne reposant sur aucune théorie clairement définie. Est-ce qu'une théorie du sens est susceptible d'émerger de l'ensemble de ces recherches ? La question reste ouverte[74].

[74] La citation de Winograd se poursuit ainsi sur une note plutôt positive : "*it was the practical experience and curiosity of the alchemists which provided the wealth of data from which a scientific theory of chemistry could be developed.*" [Winograd, 1977].

Références

1 Références personnelles

Aurélien Bossard et Thierry Poibeau. Regroupement automatique de documents en classes événementielles. In *Traitement Automatique du Langage Naturel (TALN 2008)*, Avignon, 2008.

Amanda Bouffier et Thierry Poibeau. Automatic structuring of Practice Guidelines using the GEM DTD. In *Biology and Natural Language Processing (BioNLP 07)*, Prague, 2007a.

Amanda Bouffier et Thierry Poibeau. Restructuring Documents : A Discourse-based Approach. In *Recent Advances In Natural Language Processing (RANLP 07)*, Borovets, 2007b.

Thierry Delbecque, Pierre Zweigenbaum, Jean-François Berroyer, et Thierry Poibeau. Le système STIM/LIPN à EQueR 2004, tâche médicale. In *Atelier EQueR (conférence TALN 2005)*, Dourdan, 2005.

David Faure et Thierry Poibeau. First Experiments of Using Semantic Knowledge Learned by ASIUM for Information Extraction Task using INTEX. In *Proceedings of the ECAI workshop on Ontology Learning (ECAI 2000)*, pages 7–12, Berlin, 2000.

Thierry Hamon, Adeline Nazarenko, Thierry Poibeau, Sophie Aubin, et Julien Derivière. A Robust Linguistic Platform for Efficient and Domain specific Web Content Analysis. In *RIAO 2007 - 8th Conference on Large-Scale Semantic Access to Content (Text, Image, Video and Sound)*, Pittsburgh, 2005.

Marie-Paule Jacques, Thierry Poibeau, et Françoise Gayral. *How Procedural Question-Answering on the Web could Benefit from Genre Identication ? A Corpus Study.* Manuscrit non publié, 2008.

Cédric Messiant, Anna Korhonen, et Thierry Poibeau. LexSchem : A Large Subcategorization Lexicon for French Verbs. In *Language Resource and Evaluation Conference (LREC 2008)*, Marrakech, 2008.

Thierry Poibeau. *Extraction automatique d'information, du texte brut au web sémantique.* Hermès, Paris, 2003.

Thierry Poibeau. The Multilingual Named Entity Recognition Framework. In *European Association for Computational Linguistics Conference (EACL 2003)*, pages 155–158, Budapest, 2003.

Thierry Poibeau. Dealing with Metonymic Readings of Named Entities. In *The 28th Annual Conference of the Cognitive Science Society (COGSCI 2006)*, pages 1962–1968, Vancouver, 2006.

Thierry Poibeau. Knowledge Poor Methods (Sometimes) Perform Poorly. In *ACL 2007 Semeval workshop (a workshop on semantic evaluation)*, Prague, 2007.

Thierry Poibeau et Dominique Dutoit. Automatic Extraction of Paraphrastic Phrases from Small Size Corpora. *Linguisticae Investigationes*, 32(1), 2008.

Thierry Poibeau et Leila Kosseim. Proper Name Extraction from Non-Journalistic Texts. In *Proceedings of the Eleventh Meeting of Computational Linguistics in the Netherlands (CLIN)*, pages 144–157, Twente, 2001.

Thierry Poibeau et Cédric Messiant. Do we Still Need Gold Standard for Evaluation ? In *Language Resource and Evaluation Conference (LREC 2008)*, Marrakech, 2008.

2 Références générales

Anne Abeillé. *Les nouvelles syntaxes : grammaires d'unification et analyse du français.* Armand Colin, Paris, 1993.

Anne Abeillé, Lionel Clément, et François Toussenel. Building a Treebank for French. In Anne Abeillé, éd., *Treebanks : Building and Using Parsed Corpora*, pages 165–188, Dordrecht, 2003. Kluwer Academic Publishers.

John Aberdeen, John Burger, David Day, Lynette Hirschman, Patricia Robinson, et Marc Vilain. MITRE : Description of the Alembic System used for MUC-6. In *Proceedings of the 6th Message Understanding Conference*, pages 141–152. Morgan Kauffmann, 1995.

Jean-Michel Adam. *La linguistique textuelle – Introduction à l'analyse textuelle des discours.* Armand Colin, Paris, 2005.

Douglas Appelt et David Martin. Named Entity Extraction from Speech : Approach and Results Using the TextPro System. In *Proceedings of the DARPA Broadcast News Workshop*, pages 51–54, 1999.

Sylvain Auroux, éd. *Histoire des idées linguistiques (tome 1) : La naissance des métalangages en orient et en occident.* Mardaga, Bruxelles, 1989.

Sylvain Auroux, éd. *Histoire des idées linguistiques (tome 2) : Le Développement de la grammaire occidentale.* Mardaga, Bruxelles, 1995.

John L. Austin. *Quand dire, c'est faire.* Seuil, Paris, 1962. Traduction (1970) : G. Lune.

Christelle Ayache, Brigitte Grau, et Anne Vilnat. EQueR : the French Evaluation campaign of Question Answering system EQueR/EVALDA. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 1157–1160, Gênes, 2006.

Collin Baker, Charles J. Fillmore, et John Lowe. The Berkeley FrameNet Project. In *Proceedings of the Meeting of the Association for Computational Linguistics*, pages 86–90, Montreal, 1998.

Mikhaïl Bakhtine. *Esthétique et théorie du roman.* Gallimard, Paris, 1978.

Valérie Beaudouin, Serge Fleury, Benoît Habert, Gabriel Illouz, Christian Licoppe, et Marie Pasquier. TyPWeb : décrire la Toile pour mieux comprendre les parcours. In *CIUST'01 (Colloque International sur les Usages et les Services des Télécommunications)*, pages 492–503, Paris, 2001.

Tim Berners-Lee et Mark Fischetti. *Weaving the Web : Origins and Future of the World Wide Web.* Harper, San Francisco, 1999.

Tim Berners-Lee, James Hendler, et Ora Lassila. The Semantic Web. *Scientific American*, 284(5) : 34–43, 2001.

Douglas Biber. *Variation Across Speech and Writing.* Cambridge University Press, Cambridge, 1988.

Douglas Biber. A typology of English texts. *Linguistics*, 27 : 3–43, 1989.

Douglas Biber. *Dimensions of Register Variation : A Cross-linguistic Comparison.* Cambridge University Press, Cambridge, 1995.

Frédéric Bilhaut. Introducteurs d'univers intra-prédicatifs et leur détection automatique. In *Actes du Colloque International Discours et Document*, pages 41–50, Caen, 2006.

Pascal Boldini. Des catégories aux types : un itinéraire en mathématiques appliquées. In *Séminaire Mathématiques, musique et relations avec d'autres disciplines (MAMUX)*, Paris, 2006.

Amanda Bouffier. Segmentation automatique des textes procéduraux. In *International Symposium : Discourse and Document (ISDD'06)*, 2006.

Amanda Bouffier. From Texts to Structured Documents : The Case of Health Practice Guidelines. In *International Semantic Web Conference, Doctoral Consortium (ISWC 08)*, Busan, 2007.

Amanda Bouffier. *Analyse discursive automatique de textes : application à la modélisation de textes incitatifs.* Thèse de Doctorat, Université Paris 13, 2008.

Simon Bouquet. Sur la sémantique Saussurienne. *Cahiers Ferdinand de Saussure*, 53 : 135–139, 2000.

Didier Bourigault. *Un analyseur syntaxique opérationnel : SYNTEX.* Habilitation à Diriger des Recherches, Université Toulouse-Le Mirail, 2007.

Didier Bourigault, Marie-Paule Jacques, Cécile Fabre, Cécile Frérot, et Sylwia Ozdowska. Syntex, analyseur syntaxique de corpus. In *Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles*, Dourdan, 2005.

Jacques Bouveresse. *Philosophie, mythologie et pseudo-science : Wittgenstein lecteur de Freud.* éditions de l'éclat, Paris, 1992.

Jacques Bouveresse. *Essais 3 : Wittgenstein et les sortilèges du langage.* Agone, Marseille, 2003.

Aziz A. Boxwala, Mor Peleg, Samson Tu, Omolola Ogunyemi, Qing T. Zeng, Dongwen Wang, Vimla L. Patel, Robert A. Greenes, et Edward H. Shortliffe. GLIF3 : a Representation Format for Sharable Computer-Interpretable Clinical Practice Guidelines. *Journal of Biomedical Informatics*, 37(3) : 147–161, 2004.

Michael R. Brent. From Grammar to Lexicon : Unsupervised Learning of Lexical Syntax. *Computational Linguistics*, 19 : 203–222, 1993.

Ted Briscoe et John Carroll. Automatic Extraction of Subcategorization from Corpora. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, pages 356–363, Washington, DC., 1997.

Caroline Brun, Maud Ehrmann, et Guillaume Jacquet. A Hybrid System for Named Entity Metonymy Resolution. In *ACL-SemEval 2007, 4th International Workshop on Semantic Evaluations*, Prague, 2007.

Paula Buttery, Aline Villavicencio, et Anna Korhonen, éds. *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition.* Association for Computational Linguistics, Prague, Czech Republic, June 2007.

Joan Bybee. *Frequency of Use and the Organization of Language.* Oxford University Press., Oxford, 2006.

Pierre Cadiot et Yves-Marie Visetti. *Pour une théorie des formes sémantiques.* Presses Universitaires de France, Paris, 2001.

John Carroll, Guido Minnen, et Ted Briscoe. Can Subcategorisation Probabilities Help a Statistical Parser ? In *Proceedings of the 6th ACL/SIGDAT Workshop on Very Large Corpora*, Montreal (Canada), 1998.

John Chadwick. *Le déchiffrement du linéaire B. Aux origines de la langue grecque.* Gallimard – Bibliothèques des Histoires, Paris, 1972.

Michel Chambreuil. *Grammaire de Montague : langage, traduction, interprétation.* Adosa, Clermont Ferrand, 1991.

Michel Charolles. L'encadrement du discours - Univers, champs, domaines et espaces. *Cahiers de recherche linguistique*, 6 : 1–73, 1997.

Michel Charolles, Anne le Draoulec, Marie-Paule Péry-Woodley, et Laure Sarda. Temporal and Spatial Dimensions of Discourse Organisation. *Journal of French Language Studies*, 15(2) : 203–218, 2005.

Michel Charolles et Marie-Paule Péry-Woodley. Introduction, numéro thématique les adverbiaux cadratifs. *Langue Française*, 148 : 3–8, 2005.

Christiane Chauviré. *Voir le visible : La Seconde Philosophie de Wittgenstein.* Presses Universitaires de France, Paris, 2003.

Paula Chesley et Susanne Salmon-Alt. Automatic Extraction of Subcategorization Frames for French. In *Language Resources and Evaluation Conference (LREC)*, Gênes, 2006.

Noam Chomsky. *Syntactic structures.* Mouton, La Haye, 1957.

Noam Chomsky. *Aspects of the Theory of Syntax.* MIT Press, Cambridge, MA, 1965.

Noam Chomsky. *La linguistique cartésienne.* Le Seuil, Paris, 1966. Traduction (1969) : E. Delannoe et D. Sperber.

Kenneth W. Church et Patrick Hanks. Word Association Norms, Mutual Information, and Lexicography. In *Proceedings of the 27th. Annual Meeting of the Association for Computational Linguistics*, pages 76–83, Vancouver, B.C., 1989.

Anne Condamines. *Sémantique et Corpus spécialisés : Constitution de bases de connaissances terminologiques.* Habilitation à Diriger des Recherches, Université Toulouse-le-Mirail, 2003.

Marcel Cori et Jacqueline Léon. La constitution du TAL. Étude historique des dénominations et des concepts. *Traitement Automatique des Langues*, 43(3) : 21–55, 2002.

Antoine Culioli et Claudine Normand. *Onze rencontres sur le langage et les langues.* Ophrys, Paris, 2005.

Roland Dachelet. *Sur la notion de sous-langage.* Thèse de Doctorat, Université Paris 8, 1994.

Philippe De Lara. *L'expérience de langage, Wittgenstein philosophe de la subjectivité.* Ellipses, Paris, 2005.

David Denison, Evelien Keizer, et Gergana Popova, éds. *Fuzzy Grammar.* Oxford University Press, New York, 2006.

Michael Devitt. *Ignorance of Language.* Oxford University Press, Oxford, 2006.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, et Ralph Weischedel. Automatic Content Extraction (ACE) Program - Task Definitions and Performance Measures. In *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC)*, pages 837–840, Lisbonne, 2004.

Hubert L. Dreyfus. *Alchemy and AI.* The Rand Corporation, 1965.

Hubert L. Dreyfus. *What Computers Still Can't Do : A Critique of Artificial Reason.* MIT Press, Cambridge, USA, 1992.

Dominique Dutoit. *Quelques opérations texte → sens et sens → texte utilisant une sémantique linguistique universaliste apriorique.* Thèse de Doctorat, Université de Caen, 2000.

Umberto Eco. *La recherche de la langue parfaite dans la culture européenne.* Le Seuil, Paris, 1994.

Maud Ehrmann. *Les Entités Nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation.* Thèse de doctorat, Université Paris 7 – Denis Diderot, 2008.

Cécile Fabre et Didier Bourigault. Exploiter des corpus annotés syntaxiquement pour observer le continuum entre arguments et circonstants. *Journal of French Language Studies*, 18(1) : 87–102, 2008.

Richárd Farkas, Eszter Simon, György Szarvas, et Dániel Varga. GYDER : Maxent Metonymy Resolution. In *ACL-SemEval 2007, 4th International Workshop on Semantic Evaluations*, Prague, 2007.

Gilles Fauconnier. *Les espaces mentaux.* Minuit, Paris, 1984.

David Faure. *Conception de méthode d'apprentissage symbolique et automatique pour l'acquisition de cadres de sous-catégorisation de verbes et de connaissances sémantiques à partir de textes : le système ASIUM.* Thèse de Doctorat, Université Paris 11–Orsay, 2000.

David Faure et Claire Nedellec. ASIUM : Learning Subcategorization Frames and Restrictions of Selection. In *Proceedings of the Text Mining workshop, 10th European Conference on Machine Learning (ECML 98)*, Chemnitz, 1998.

Chritiane Fellbaum, éd. *WordNet : An Electronic Lexical Database.* MIT Press, Cambridge, MA, 1998.

Charles J. Fillmore. The Case for Case. In Bach et Harms, éds., *Universals in Linguistic Theory*, pages 1–88, New York, 1968. Holt, Rinehart, and Winston.

Charles J. Fillmore. Frame Semantics. In *Linguistics in the Morning Calm*, pages 111–137, Seoul, 1982. Hanshin Publishing Co.

Jenny Rose Finkel, Trond Grenager, et Christopher Manning. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings, 43th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 363–370, Univ. of Michigan – Ann Arbour, 2005.

John R. Firth. Modes of Meaning. In *Papers in linguistics (1934-1951)*, 1957a.

John R. Firth. *Papers in Linguistics (1934-1951)*. Oxford University Press, Oxford, 1957b.

John R. Firth. The technique of semantics. In *Papers in linguistics (1934-1951)*, Oxford, 1957c. Oxford University Press.

John R. Firth. Descriptive Linguistics and the Study of English. In *Selected Papers of John R. Firth*, 1968.

Jerry Fodor. *The Language of Thought*. Harvard University Press, Harvard, 1975.

Helka Folch, Serge Heiden, Benoît Habert, Serge Fleury, Gabriel Illouz, Pierre Lafon, Julien Nioche, et Sophie Prévost. TyPTex : Inductive Typological Text Classification Analysis for NLP Systems Tuning/Evaluation. In *Proceedings of 2nd International Conference on Language Resources and Evaluation (LREC)*, pages 141–148, Athènes, 2000.

Catherine Fuchs. *Paraphrase et énonciation*. Ophrys, Paris, 1994.

Catherine Fuchs. *La linguistique cognitive*. Maison des Sciences de l'Homme / Ophrys, Paris, 2004.

William Gale, Kenneth W. Church, et David Yarowsky. One Sense Per Discourse. In *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, pages 233–237, New York, 1992.

Michel Galmiche. *Sémantique linguistique et logique, un exemple : la théorie de R. Montague*. Presses Universitaires de France, Paris, 1991.

Claire Gardent, Bruno Guillaume, Guy Perrier, et Ingrid Falk. Maurice Gross' Grammar Lexicon and Natural Language Processing. In *Proceedings of the 2nd Language and Technology Conference*, Poznan, 2005.

Marie-Noëlle Gary-Prieur. *Grammaire du nom propre*. Presses Universitaires de France, Paris, 1994.

Françoise Gayral, Daniel Kayser, et Nathalie Pernelle. In Search of the Semantic Value(s) of an Occurrence : an Example and a Framework. *Computing Meaning*, 77 : 53–69, 2001.

Peter T. Geach. *Mental Acts*. Routledge, Londres, 1957.

Dedre Gentner et Susan Goldin-Meadow, éds. *Language in Mind : Advances in the Study of Language and Thought*. MIT Press, Cambridge, MA, 2003.

Danièle Godard et Jacques Jayez. Towards a Proper Treatment of Coercion Phenomena. In *Proceedings of the Sixth European Association for Computational Linguistics Conference (EACL)*, pages 168–177, Utrecht, 1993.

Adele Goldberg. *Constructions : A Construction Grammar Approach to Argument Structure*. University of Chicago Press, Chicago, 1995.

Adele Goldberg. *Constructions at Work : The Nature of Generalization in Language*. Oxford University Press, Oxford, 2006.

Guillaume Gravier, Jean-François Bonastre, Edouard Geoffrois, Sylvain Galliano, et Khalid Choukri. ESTER, une campagne d'évaluation des systèmes d'indexation automatique d'émissions radiophoniques en français. In *Journées d'Etude de la Parole*, Fès, 2004.

Ralph Grishman. The NYU system for MUC6 or where's the syntax ? In *Proceedings of the 6th Message Understanding Conference*. Morgan Kaufmann, 1995.

Ralph Grishman et Beth Sundheim. Message understanding conference 6 – A brief history. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 466–471, 1996.

Gaston Gross. Degré de figement des noms composés. *Langages*, 90 : 57–72, 1988.

Maurice Gross. *Méthodes en syntaxe*. Hermann, Paris, 1975.

Benoît Habert. *Des mots complexes possibles aux mots complexes existants : l'apport des corpus*. Habilitation à Diriger des Recherches, Université Lille 3, 1998.

Benoît Habert, Adeline Nazarenko, et André Salem. *Les linguistiques de corpus*. Coll. U Linguistique. Armand Colin/Masson, Paris, 1997.

M.A.K. Halliday et Ruqaiya Hasan. *Cohesion in English*. Longman, Londres, 1976.

M.A.K. Halliday et Christian M. I. M. Matthiessen. *An Introduction to Functional Grammar*. Arnold, Londres, 2004. 3e éd.

Randy H. Harris. *The Linguistics Wars*. Oxford University Press, New York, 1993.

Zellig Harris. *Methods in Structural Linguistics*. University of Chicago Press, Chicago, 1951.

Zellig Harris. *Language and Information*. Columbia University Press, New York, 1988.

Zellig Harris, Michael Gottfried, Thomas Ryckman, Paul Mattick Jr., Anne Daladier, Tzvee N. Harris, et Suzanna Harris. *The Form of Information in Science : Analysis of an immunology sublanguage*. Boston Studies in the Philosophy of Science, n°104. Kluwer Academic Publishers, Dordrecht, 1989.

Laurent Heurley. Vers une définition du concept de texte procédural : le point de vue de la psycholinguistique. *Les Cahiers du Français Contemporain*, 4 : 109–133, 1997.

Graeme Hirst. Views of Text-Meaning in Computational Linguistics : Past, Present, and Future. In Gordana Dodig-Crnkovic et Susan Stuart, éds., *Computation, Information, Cognition – The Nexus and the Liminal*, pages 270–279, Newcastle-upon-Tyne, 2007. Cambridge Scholars Publishing.

John Hutchins, éd. *Early Years in Machine Translation : Memoirs and Biographies of Pioneers*. John Benjamins, Amsterdam, 2000.

Lidjia Iordanskaja et Igor Mel'cuk. The Notion of Surface-Syntactic Relation Revisited (Valence-Controlled Surface-Syntactic Relations in French). In *Slovo v tekste i v slovare. Sbornik statej k semidesjatiletiju akademika Ju. D. Apresjana)*, pages 391–433, Moscou, 2000. Jazyki russkoj kul'tury.

Marie-Paule Jacques et Josette Rebeyrolle. Titres et structuration des documents. In *International Symposium : Discourse and Document (ISDD)*, Caen, 2006.

Guillaume Jacquet et Maud Ehrmann. Vers une double annotation des entités nommées. *Traitement Automatique des Langues*, 47(1), 2006.

Kerstin Jonasson. *Le Nom propre. Constructions et interprétations*. Duculot, Louvain-la-Neuve, 1994.

Daniel Kayser. Une sémantique qui n'a pas de sens. *Langages*, 87 : 33–45, 1987.

Daniel Kayser. What Kind of Thing is a Concept ? *Computational Intelligence*, 4(2) : 158–165, 1988.

Daniel Kayser. Réponse à Kleiber et Riegel. *Linguisticae Investigationes*, 13 : 419–422, 1989.

Daniel Kayser et Farid Nouioua. About Norms and Causes. *International Journal of Artificial Intelligence Tools*, 14(1-2) : 7–24, 2005.

Karin Kipper-Schuler. *VerbNet : a Broad Coverage, Comprehensive, Verb Lexicon*. Thèse (PhD), University of Pennsylvania, 2003.

Georges Kleiber. *Problèmes de référence : descriptions définies et noms propres*. Klincksieck, Paris, 1981.

Georges Kleiber. *La sémantique du prototype, Catégories et sens lexical*. PUF, Paris, 1990.

Georges Kleiber. *Nominales : essais de sémantique référentielle*. Armand Colin, Paris, 1994.

Georges Kleiber et Martin Riegel. Une sémantique qui n'a pas de sens n'a vraiment pas de sens. *Linguisticae Investigationes*, 13 : 405–417, 1989.

Georges Kleiber et Martin Riegel. Sens lexical et interprétations référentielles. Un écho à la réponse de D. Kayser. *Linguisticae Investigationes*, 15 : 181–201, 1991.

Anna Korhonen et Ted Briscoe. Extended Lexical-Semantic Classification of English Verbs. In Dan Moldovan et Roxana Girju, éds., *HLT-NAACL 2004 : Workshop on Computational Lexical Semantics*, pages 38–45, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.

Anna Korhonen, Genevieve Gorrell, et Diana McCarthy. Statistical Filtering and Subcategorization Frame Acquisition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, Hong Kong, 2000.

Anna Korhonen, Yuval Krymolowski, et Ted Briscoe. A Large Subcategorization Lexicon for Natural Language Processing Applications. In *Proceedings of the 5th international conference on Language Resources and Evaluation*, Genova, Italy, 2006a.

Anna Korhonen, Yuval Krymolowski, et Nigel Collier. Automatic Classification of Verbs in Biomedical Texts. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL 2006)*, pages 345–352, Sydney, 2006b.

Anna Korhonen, Yuval Krymolowski, et Zvika Marx. Clustering Polysemic Subcategorization Frame Distributions Semantically. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, pages 64–71, Sapporo, 2003.

Saul Kripke. *La logique des noms propres.* Minuit, Paris, 1982 (1972).

Francis Kubala, Richard Schwartz, Rebecca Stone, et Ralph Weischedel. Named Entity Extraction from Speech. In *Proceedings of the DARPA Broadcast Transcription and Understanding Workshop*, 1999.

Anna Kupść. Extraction automatique de cadres de sous-catégorisation verbale pour le français à partir d'un corpus arboré. In *Traitement Automatique du Langage Naturel (TALN 2007)*, Toulouse, June 2007.

John Lafferty, Andrew McCallum, et Fernando Pereira. Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the International Conference on Machine Learning (ICML-2001)*, Williamstown, USA, 2001.

George Lakoff. *Women, Fire, and Dangerous Things : What Categories Reveal about the Mind.* University of Chicago Press, Chicago, 1987.

George Lakoff et Mark Johnson, éds. *Philosophy in the Flesh : The Embodied Mind and Its Challenge to Western Thought.* Basic Books, New York, 1999.

Eric Laporte. In Memoriam Maurice Gross. *Archives of Control Sciences (Actes Language and Technology Conference)*, 15(3) : 257–278, 2005.

Gilbert Lazard. *L'actance*. Presses Universitaires de France, Paris, 1994.

Anne Le Draoulec et Marie-Paule Péry-Woodley. Encadrement temporel et relations de discours. *Langue Française*, 148 : 45–60, 2005.

Michel Le Du. *La nature sociale de l'esprit : Wittgenstein, la psychologie et les sciences humaines*. Vrin, Paris, 2005.

Céline Le Meur, Sylvain Galliano, et Édouard Geoffrois. Conventions d'annotations en Entités Nommées - ESTER. Rapport technique de la campagne Ester, 2004.

Michelle Lecolle. Personnifications et métonymies dans la presse écrite : comment les différencier ? *Semen*, 15, 1999.

Michelle Lecolle. Polyvalence des toponymes et interprétation en contexte. *Pratiques*, 129/130 :107–122, 2006.

Dominique Legallois. *Les arguments du discours contre ceux du verbe : construction, colligation, coercition*. Manuscrit en cours de soumission (communication personnelle), 2008.

Dominique Legallois et Jacques François, éds. *Autour des grammaires de construction et de patterns*, Cahier du CRISCO, 21, 2006.

Gottfried Wilhelm Leibniz. *L'harmonie des langues*. Garnier-Flammarion, Paris, 2000. Trois essais de 1679 et 1710.

Alain Lemaréchal. Actants ou arguments ? In F. Madray-Lesigne et J. Richard-Zapella, éds., *Lucien Tesnière aujourd'hui*, pages 165–175, Louvain, 1995. Peeters.

Sarah Leroy. *De l'identification à la categorisation : l'antonomase du nom propre en francais*. Peeters, Louvain–Paris, 2004a.

Sarah Leroy. *Le nom propre en français*. Ophrys, Paris, 2004b.

Beth Levin. *English Verb Classes and Alternations : a preliminary investigation*. University of Chicago Press, Chicago et Londres, 1993.

Beth Levin et Malka Rappaport Hovav. *Argument Realization*. Cambridge University Press, Cambridge, 2005.

Jacqueline Léon. From universal languages to intermdediary languages in machine translation : the work of the Cambridge Language Research Unit (1955-1970). In *Proceedings of the 9th International Conference on the History of Language Sciences (ICHoLS)*, Sao Polo, 2002a.

Jacqueline Léon. Le CNRS et les débuts de la traduction automatique en France. *La Revue pour l'histoire du CNRS*, 6 : 6–24, 2002b.

Jacqueline Léon. Empiricism versus Rationalism revisited. Current Corpus Linguistics and Chomsky's Arguments against Corpus, Statistics and Probabilities in the 1950-1960s. In S. Matthaios et P. Schmitter, éds., *Linguistische und Epistemologische Konzepte – Diachron*, pages 157–176, Münster, 2007a. Nodus Publikationen.

Jacqueline Léon. Meaning by Collocation : the Firthian Filiation of Corpus Linguistics. In D. Kibbee, éd., *Proceedings of the 10th International Conference on the History of Language Sciences (ICHoLS)*, Amsterdam, 2007b. John Benjamins.

Christophe Luc et Jacques Virbel. Le modèle d'architecture textuelle : fondements et expérimentation. *Verbum*, 23(1) : 103–123, 2001.

John Lyons. *Linguistic Semantics, an Introduction.* Cambridge University Press, Cambridge, 1995.

Dominique Maingueneau. *Les termes clés de l'analyse du discours.* Seuil, Paris, 1996.

Bronislaw Malinowski. The Problem of Meaning in Primitive Languages. In *Supplement to C. K. Ogden and I. A. Richards (eds.) The Meaning of Meaning : A Study of the Influence of Language upon Thought and the Science of Symbolism*, pages 451–510, Londres, 1923. Routledge & Kegan Paul.

Denise Malrieu et François Rastier. Genres et variations morphosyntaxiques. *Traitement Automatique des Langues*, 42(2) : 548–577, 2001.

William C. Mann et Sandra. A. Thompson. Rhetorical Structure Theory : Toward a Functional Theory of Text Organization. *Text*, 8(3) : 243–281, 1988.

Christopher D. Manning. Automatic Acquisition of a Large Subcategorization Dictionary from Corpora. In *Meeting of the Association for Computational Linguistics*, pages 235–242, 1993.

Christopher D. Manning. Probabilistic Syntax. In Rens Bod, Jennifer Hay, et Stefanie Jannedy, éds., *Probabilistic Linguistics*, pages 289–341, Cambridge, MA, 2003. MIT Press.

Daniel Marcu. A Decision-based Approach to Rhetorical Parsing. In *Proceedings of the 37th Annual Meeting of the Associa-tion for Computational Linguistics (ACL'99)*, pages 365–372, Maryland, 1999.

Katja Markert et Malvina Nissim. Task08 : Metonymy Resolution at Semeval 2007. In *Proceedings of Semeval*, Prague, 2007.

Jérémie Mary. *Etude de l'apprentissage actif : Application à la conduite d'expériences.* Thèse de doctorat, Université Paris 11 – Orsay, 2005.

Margaret Masterman. *Language, Cohesion and Form.* Cambridge University Press, Cambridge, 2005. Editeur : Yorick Wilks.

Tony McEnery et Andrew Wilson. *Corpus Linguistics : An Introduction.* Edinburgh University Press, Edimbourg, 2001.

Igor Mel'cuk. Actants in Semantics and Syntax. *Linguistics*, 42(1–2) : 1–66 ; 247–291, 2004.

John Stuart Mill. *Système de logique : déductive et inductive.* Mardaga, Bruxelles, 1995 (1843).

Emmanuel Morin et Christian Jacquemin. Projecting Corpus-Based Semantic Links on a Thesaurus. In *Proceedings, 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 389–396, Univ. of Maryland, 1999.

MUC7. Proceedings of the 7th Message Understanding Conference, 1998.

Mark A. Musen, Samson Tu, Amar K. Das, et Yuval Shahar. EON : a Component-Based Approach to Automation of Protocol-Directed Therapy. *Journal of the American Medical Informatics Association*, 3 : 367–388, 1996.

Adeline Nazarenko. *Donner accès au contenu des documents textuels. Acquisition de connaissances et analyse de corpus spécialisés.* Habilitation à Diriger des Recherches, Université Paris 13, 2004.

Bruce E. Nevin et Stephen M. Johnson, éds. *The Legacy of Zellig Harris : Language and Information into the 21st Century : Computability of Language and Computer Applications.* John Benjamins, Amsterdam, 2002.

Malvina Nissim et Katja Markert. Syntactic Features and Word Similarity for supervised Metonymy Resolution. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, Sapporo, 2003.

Kristóf J. C. Nyíri. Wittgenstein as a Philosopher of Secondary Orality. *Grazer Philosophische Studien*, 52 : 45–57, 1996.

Ruth O'Donovan, Michael Burke, Aoife Cahill, Josef van Genabith, et Andy Way. Large-Scale Induction and Evaluation of Lexical Resources from the Penn-II and Penn-III Treebanks. *Computational Linguistics*, 31(3) : 329–366, 2005.

Charles K. Ogden. *Basic English : A General Introduction with Rules and Grammar.* Paul Treber & Co., Londres, 1930.

Charles K. Ogden et Ivor A. Richards. *The Meaning of Meaning : A Study of the Influence of Language upon Thought and the Science of Symbolism.* Routledge & Kegan Paul, Londres, 1923.

Franck R. Palmer. *Selected papers of John R. Firth.* Longmans, Londres, 1968.

Elsa Pascual. *Représentation de l'architecture textuelle et génération de texte.* Thèse de Doctorat, Université Paul Sabatier, Toulouse, 1991.

Elsa Pascual et Marie-Paule Péry-Woodley. La définition dans le texte. In *Textes de type consigne - Perception, action, cognition*, Toulouse, 1995. Atelier « Texte et Communication » PRESCOT.

Roger T. Pédauque (collectif RTP-DOC). *Le texte en jeu. Permanence et transformations du document.* Document web, 2005. URL `http://archivesic.ccsd.cnrs.fr/sic_00001401.html`.

Roger T. Pédauque (collectif RTP-DOC). *Le document à la lumière du numérique.* C&F éditions, Caen, 2006.

Roger T. Pédauque (collectif RTP-DOC). *La redocumentarisation du monde.* Editions Cepadues, Toulouse, 2007.

David Pears. *Paradox and Platitude in Wittgenstein's Philosophy.* Oxford University Press, Oxford, 2007.

David Pierce et Claire Cardie. User-Oriented Machine Learning Strategies for Information Extraction : Putting the Human Back in the Loop. In *Proceedings of the IJCAI Workshop on Adaptive Text Extraction and Mining (ATEM 2001)*, Seattle, 2001.

Jean-Marie Pierrel, éd. *Ingénierie des Langues.* Editions Hermès, Paris, 2000.

Steven Pinker. *Learnability and Cognition : The acquisition of argument structure.* MIT Press, Cambridge, MA, 1989.

Bernard Pottier. *Sémantique générale.* Presses Universitaires de France, Paris, 1992.

Judita Preiss, Ted Briscoe, et Anna Korhonen. A System for Large-Scale Acquisition of Verbal, Nominal and Adjectival Subcategorization Frames from Corpora. In *Meeting of the Association for Computational Linguistics*, pages 912–918, Prague, 2007.

Marie-Paule Péry-Woodley. Modes d'organisation et de signalisation dans des textes procéduraux. *Langages*, 141 : 28–46, 2001.

James Pustejovsky. *The Generative Lexicon.* MIT Press, Cambridge, MA, 1995.

François Rastier. Prédication, actance et zones anthropiques. In M. Forsgren, K. Jonasson et H. Kronning, éd., *Prédication, Assertion, Information*, pages 443–461, Stockholm, 1998.

François Rastier. Dalla significazione al senso : per una semiotica senza ontologia (De la signification au sens — pour une sémiotique sans ontologie). In Pierluigi Basso et Lucia Corrain, éds., *Eloquio del senso*, pages 213–240, Milan, 1999. Costa e Nolan.

François Rastier. *Arts et sciences du texte.* Presses Universitaires de France, Paris, 2001.

François Rastier. De l'origine du langage à l'émergence du milieu sémiotique. *Marges linguistiques*, 11, 2006.

François Rastier. Que cachent les « données textuelles » ? In *Journées d'Analyse des Données Textuelles (JADT 2008)*, Lyon, 2008. Lexicometrica.

François Rastier, Marc Cavazza, et Anne Abeillé. *Sémantique pour l'analyse*. Masson, Paris, 1994.

François Recanati. *Le sens littéral : Langage, contexte, contenu*. L'éclat, Paris, 2007.

Paul Ricœur. *Sur la traduction*. Bayard, Paris, 2006.

E. Roche et Y. Schabes. *Finite-state language processing*. MIT Press, Cambridge, MA, 1997.

Eleanor Rosch. Natural Categories. *Cognitive Psychology*, 4 : 328–350, 1973.

Gérard Sabah, éd. *L'intelligence artificielle et le langage, représentations des connaissances*. Hermes, Paris, 1988.

Gérard Sabah, éd. *L'intelligence artificielle et le langage, processus de compréhension*. Hermes, Paris, 1989.

Gérard Sabah. Le "carnet d'esquisses" : une mémoire interprétative dynamique. In *Actes de la conférences Reconnaissance des Formes et Intelligence Artificielle (RFIA)*, pages 1096–1105, Rennes, 1996.

Gérard Sabah, éd. *Compréhension des langues et interaction*. Hermes Lavoisier, Paris, 2006.

Naomi Sager, Carol Friedman, et Margaret S. Lyman, éds. *Medical Language Processing : Computer Management of Narrative Data*. Addison-Wesley, Reading, 1987.

Benoît Sagot, Lionel Clément, Eric Villemonte de la Clergerie, et Pierre Boullier. The Lefff 2 Syntactic Lexicon for French : Architecture, Acquisition, Use. In *Language Resource and Evaluation Conference (LREC 2006)*, Gênes, 2006.

Benoît Sagot et Eric Villemonte de la Clergerie. Error mining in parsing results. In *Conference of the Association for Computational Linguistics*, pages 329–336, Syndey, Australie, 2006. Association for Computational Linguistics.

Patrick Saint-Dizier. *Predicative Forms in Natural Language and lexical Knowledge bases*. Kluwer Academic, Dordrecht, 1999.

Patrick Saint-Dizier et Evelyne Viegas, éds. *Computational Lexical Semantics*. Cambridge University Press, Cambridge, 1995.

Geoffrey Sampson. *Empirical Linguistics*. Continuum, Londres, 2001.

Sabine Schulte im Walde. A Subcategorisation Lexicon for German Verbs induced from a Lexicalised PCFG. In *Proceedings of the 3rd Conference on Language Resources and Evaluation*, volume IV, pages 1351–1357, Las Palmas, 2002.

Sabine Schulte im Walde. Experiments on the Automatic Induction of German Semantic Verb Classes. *Computational Linguistics*, 32(2) : 159–194, 2006.

Satoshi Sekine, Kiyoshi Sudo, et Chikashi Nobata. Extended Named Entity Hierarchy. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*, pages 1818–1824, Las Palmas, 2002.

Richard N. Shiffman, Bryant T. Karras, Abha Agrawal, Roland Chen, Luis Marenco, et Sujai Nath. GEM : A Proposal for a More Comprehensive Guideline Document Model Using XML. *Journal of the American Medical Informatics Association*, 7 : 488–498, 2000.

Gilbert Simondon. *Du mode d'existence des objets techniques*. Aubier, Paris, 2001 (1958).

John Sinclair. Beginning the Study of Lexis. In *In memory of J. R. Firth*, 1966.

John Sinclair. *Trust the Text : Language, Corpus and Discourse*. Routledge, Londres, 2004.

John Sinclair. *Essential Corpus Linguistics*. Routledge, Londres, 2007.

Karen Spärck Jones. *Synonymy and Semantic Classification*. Edinburgh University Press, Edimbourg, 1964. Thèse, publiée en 1986.

Karen Spärck Jones. R.H. Richens : Translation in the NUDE. In *Early Years in Machine Translation*, pages 263–278, Amsterdam, 2000. John Benjamins.

Karen Spärck Jones. What's New about the Semantic Web ? Some Questions. *ACM SIGIR Forum*, 38(2) :18–23, 2004.

Georges Steiner. *Après Babel*. Albin Michel, Paris, 1998.

Lucien Tesnière. *Éléments de syntaxe structurale*. Klincksieck, Paris, 1959.

Michael Tomasello. *Constructing A Language : A Usage-Based Theory Of Language Acquisition*. Harvard University Press, Cambridge, MA, 2005.

Karel van Den Eynde et Claire Blanche-Benveniste. Syntaxe et mécanismes descriptifs : présentation de l'approche pronominale. *Cahiers de Lexicologie*, 32 : 3–27, 1978.

Karel van Den Eynde et Piet Mertens. *Le dictionnaire de valence Dicovalence : manuel d'utilisation*. Manuscrit, Leuven, 2006.

C.J. Keith van Rijsbergen. *The Geometry of Information Retrieval*. Cambridge University Press, Cambridge, 2004.

Tony Veale, Pablo Gervas, et Alison Pease. Understanding Creativity : A Computational Perspective. Introduction to the Special Issue of Computational Creativity. *New Generation Computing*, 24 : 203–207, 2006.

Bernard Victorri. La construction dynamique du sens. In *11e congrès de reconnaissance des Formes et d'Intelligence Artificielle (RFIA)*, pages 15–29, Clermont Ferrand, 1998.

Jacques Virbel. Langage et métalangage dans le texte du point de vue de l'édition en informatique textuelle. *Cahiers de Grammaire*, 10 : 5–72, 1985.

Ellen M. Voorhees. Overview of the TREC-2004 Question Answering Track. In *Proceedings 13th Text REtrieval Conference (TREC)*, Gaithersburg, USA, 2004.

Ellen M. Voorhees et Lori P. Buckland, éds. *Proceedings of the Fifteenth Text REtrieval Conference*, Gaithersburg, USA, 2006.

Piek Vossen. Introduction to EuroWordNet. In *EuroWordNet : a Multilingual Database with Lexical Semantic Networks*, pages 1–17, Dordrecht, 2002. Kluwer Academic Publishers.

Joseph Weizenbaum. ELIZA – A Computer Program for the Study of Natural Language Communication between Man and Machine. *Communications of the Association for Computing Machinery*, 9 : 36–45, 1966.

Anna Wierzbicka. *Lingua Mentalis : the Semantics of Natural Language*. Academic Press, Sydney, 1980.

Yorick Wilks. A Preferential Pattern-Seeking Semantics for Natural Language Inference. *Artificial Intelligence*, 6 : 53–75, 1975.

Yorick Wilks. What would a Wittgensteinian Linguistics be like ? In *Proceedings of 9th International Pragmatics Conference*, Riva del Garda, 2005.

Yorick Wilks, Xiuming Huang, et Dan Fass. Syntax, Preference and Right Attachment. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 779–784, Los Angeles, 1985.

Yorick Wilks, Brian Slator, et Louise Guthrie. *Electric Words : Dictionaries, Computers, and Meanings*. MIT Press, Cambridge, MA, 1996.

Geoffrey Williams. La linguistique et le corpus : une affaire prépositionnelle. In *Actes du colloque « Corpus en Lettres et Sciences sociales : des documents numériques à l'interprétation »*, Albi, 2006. Texto ! Sous la dir. de F. Rastier et M. Ballabriga.

Terry Winograd. On some Contested Suppositions of Generative Linguistics about the Scientific Study of Language. *Cognition*, 5 : 151–179, 1977.

Terry Winograd. The Boundaries of Humanity : Humans, Animals, Machines. In D. Partridge et Y. Wilks, éds., *The Foundations of Artificial Intelligence*, pages 167–189, Cambridge, 1990. Cambridge University Press.

Ludwig Wittgenstein. *Tractatus Logico Philosophicus*. Bibliothèque des Idées, Gallimard, Paris, 1919. Traduction (1961) : P. Klossowski.

Ludwig Wittgenstein. *Investigations philosophiques.* Bibliothèque des Idées, Gallimard, Paris, 1953. Traduction (1961) : P. Klossowski.

David Yarowsky. Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 454–460, Nantes, 1992.

François Yvon. *Des apprentis pour le Traitement Automatique des Langues.* Habilitation à Diriger des Recherches, Ecole Nationale Supérieure des Télécommunications, 2006.

Annexe
Projets et campagnes d'évaluation

Je ne mentionne dans cette section que les projets et les campagnes auxquels j'ai participé depuis que je suis au LIPN, soit depuis la fin 2003. Les projets auxquels j'ai participé quand je travaillais chez Thales ne sont pas détaillés.

1 Contrats de recherches

Projet ISCC Scholarweb (Emergence de communautés savantes sur le web) — 2008-2010.

Le projet vise à faire un bilan des initiatives, des besoins et des développements récents en matière d'édition scientifique de textes littéraires, philosophiques ou techniques. Il s'appuie sur l'analyse de l'existant en matière de développement de plates-formes d'édition d'une part, et de traitement des langues d'autre part. Les partenaires sont l'ITEM (Institut des textes et manuscrits Modernes) et le LIPN.

Projet Alliance TAACL (Technologies multilingues pour l'Acquisition Automatique de Connaissances Lexicales) — collaboration bilatérale avec le Laboratoire d'Informatique de l'Université de Cambridge — 2008-2009.

Le projet vise à étudier des méthodes d'analyse automatique de gros corpus pour en extraire des informations lexicales, sur la base de régularités de comportement. Les informations obtenues concernent essentiellement la sous-catégorisation, ainsi que le comportement syntaxique et sémantique des verbes.

Projet ANR CROTAL (Conditional Random Fields pour le TAL) — `http://www.grappa.univ-lille3.fr/~tellier/crotal`, 2008-2009.

Le projet vise à explorer des nouvelles techniques d'apprentissage, et plus spécialement les *Conditional Random Fields* (CRF, [Lafferty *et al.*, 2001]), pour différentes tâches de TAL. Les partenaires du projet sont l'Université de Lille 3, le LIMSI et le LIPN.

Projet INFOMAGIC (Projet phare du Pôle de Compétitivité Cap Digital — Image, Multimédia et Vie Numérique) — `http://www.capdigital.com/`, 2006-2009

Le projet vise à explorer de nouvelles techniques d'ingénierie des connaissances, essentiellement sur des données multimédia (texte, son, image et vidéo). Le Pôle de compétitivité regroupe un ensemble d'acteurs de premier plan autour de ces thématiques — grands groupes, PME, universités — et vise à faire émerger de nouvelles collaborations (à travers ce projet, le LIPN a ainsi pu nouer des collaborations avec des entreprises comme Arisem, Xerox ou Temis).

Projet ANR TEXTCOOP (Rendre l'accès au Texte plus Coopératif) — `http://www.textcoop.org/`, 2006-2008

Le projet porte sur l'amélioration des systèmes de questions-réponses pour les questions non factuelles, en offrant une pré-analyse du contenu des documents. dans ce cadre, le LIPN doit développer des techniques permettant le typage des documents suivants des types pré-établis. Les partenaires sont Sinequa, l'IRIT et le LIPN.

Projet européen STREP ALVIS (*Superpeer Semantic Search Engine*) — `http://www.alvis.info/alvis/`), 2004-2006

Le projet ALVIS visait à développer des systèmes de recherche d'information répartis et spécialisés par sujet (moissonneur, en anglais *crawler*). Le LIPN était plus spécifiquement en charge du développement d'une chaîne d'annotation linguistique efficace et adaptable, capable de traiter les documents envoyés par les moissonneurs spécialisés.

Projet national RNTL ExtraPloDocs (ETRAction de Connaissances pour l'exPLOitation de la DOCumentation Scientifique) — `http://www-lipn.univ-paris13.fr/~poibeau/Extra`, 2002-2005

Le projet ExtraPloDocs concernait l'analyse automatique de textes de biologie. Le but du projet était d'extraire des information en génomique à partir de grandes bases textuelles comme Medline.

2 Campagnes d'évaluation

TAC (*Text Analysis Conference*) — `http://www.nist.gov/tac/about/index.html` — 2008 (en cours)

La campagne TAC porte sur le résumé automatique de documents. Le LIPN participe à deux tâches : *Update* (résumé d'un ensemble de textes puis d'un nouvel ensemble apportant des informations nouvelles par rapport au premier groupe de texte) à partir des travaux d'Aurélien Bossard et *Opinion Pilot* (résumé d'opinions exprimées dans des blogs), à partir des travaux de Michel Généreux.

SEMEVAL (*Semantic Evaluation*) — `http://nlp.cs.swarthmore.edu/semeval/` — 2007

La conférence SEMEVAL concerne l'évaluation de différentes tâches d'analyse sémantique. J'ai participé en 2007 à la tâche métonymie, qui visait l'analyse d'emplois métonymiques d'entités nommées au sein de corpus de presse en anglais. Le système n'a pas obtenu de

bonnes performances, dans la mesure où l'on ne disposait pas d'analyseur syntaxique et que l'on n'a pas fait appel à d'autres données que celles fournies par les utilisateurs pour l'apprentissage (cf. chapitre 2).

EQueR (Evaluation des systèmes de Questions-Réponses) — `www.technolangue.net` — 2004

La campagne EQueR, organisée dans le cadre du programme Technolangue, portait sur les systèmes de questions-réponses. J'y ai participé en collaboration avec Pierre Zweigenbaum, alors aux Hôpitaux de Paris/INSERM (nous avons participé uniquement à la tâche portant sur le corpus spécialisé sur le domaine médical). Le travail du LIPN a principalement porté sur la reconnaissance des entités nommées et le typage des questions. Le système s'est classé 3^{e} sur 7 participants (à part le meilleur participant, les autres ont obtenus des performances assez proches et relativement médiocres).

ESTER (Evaluation des systèmes de Transcription Enrichie de la Parole) — `www.technolangue.net` — 2004

La campagne ESTER, organisée dans le cadre du programme Technolangue, portait avant tout sur la transcription de la parole. J'ai participé à une sous-tâche spécialisée, afin de reconnaître les entités nommées das le corpus transcrit à la main. Le système développé s'est classé 2^{e} sur 3, les trois participants obtenant des résultats très proches (à noter toutefois que le système du LIPN n'était pas spécialement adapté à l'analyse de l'oral). Ce corpus a ensuite été utilisé pour l'analyse de la métonymie, avec de bons résultats (cf. chapitre 2).

Glossaire

Certaines définitions données ici reprennent directement des éléments du mémoire ; ce qui est indiqué dans le glossaire n'est alors qu'une définition grossière destinée à aider le lecteur à se repérer mais les notions sont discutées plus avant dans le mémoire.

Annotation sémantique : Tâche qui vise à mettre en évidence, à même le texte (en général par un jeu de couleurs approprié), des éléments d'information pertinents (mots clés, noms propres, *etc.*).

Cooccurrence : Apparition de deux mots dans un même contexte. On admet généralement que l'apparition fréquente de deux mots dans un même contexte est pertinent d'un point de vue sémantique (on dit alors que les deux mots *cooccurrent* ensemble). Voir *collocation*.

Colligation : Phénomène de comportement grammatical préférentiel (par exemple, *parfois* a tendance à apparaître souvent en début de phrase) ; la colligation est au niveau syntaxique ce que la collocation est au niveau lexical et permet de rendre compte des expressions idiomatiques (*prendre le bus*), qui vont au-delà de la simple cooccurrence.

Collocation : En linguistique, une collocation est une cooccurrence privilégiée, un rapprochement de termes qui n'est pas fortuit, sans être fixe, comme : *voix suave*, *entraîner des conséquences*. On parle aussi d'expression ou de locution figée. Plusieurs modélisation mathématique de la notion de collocation ont été proposées (définition inspirée de Wikipedia).

Entités nommées : Les entités nommées désignent les noms de personnes, de lieux, d'organisations mais aussi les dates ou les unités monétaires. Une discussion sur la nature des entités nommées figure dans le texte, au chapitre 2.

Extraction d'information : Tâche qui vise à extraire des informations structurées pour remplir une base de données (concernant par exemple des rachats d'entreprises, des réseaux d'interactions géniques, *etc.*). L'extraction exige une analyse précise du contenu textuel, par opposition à la recherche d'information, qui peut traiter le document comme un « sac de mots » (c'est-à-dire sans tenir compte de l'ordre linéaire ou de la syntaxe).

Figement : Une unité polylexicale est figée si les possibilités de variations morpho-syntaxiques de l'unité considérée sont limitées. Le figement est rarement total, aussi a-t-on

pu parler de degré de figement [Gross, 1988] (mais il serait sans doute plus juste de parler de continuum). Le figement est souvent lié à la non-compositionnalité ou à une compositionnalité partielle au niveau sémantique : le sens de l'unité ne peut pas être complètement déduit du sens des unités simples qui la composent.

Fonction discursive : La fonction discursive désigne la visée pragmatique d'un segment textuel. Par exemple, un segment textuel peut décrire la façon d'accomplir une action. On parlera alors d'un fragment dont la visée est « procédurale », le caractère procédural correspondant à une fonction discursive.

Genre textuel : Catégorie de textes fondée sur une pratique sociale établie, définie *a priori*. La catégorie est reconnue et validée par le fait qu'elle peut se dénommer. Biber [1989, p. 5] ajoute que les genres correspondent à des catégories de textes distinguées spontanément par les locuteurs confirmés (*mature*) d'une langue. On distingue ainsi le roman, la poésie, le formulaire administratif, l'éditorial, la thèse, *etc.*

Macrosémantique : Sémantique des textes (voir aussi *Microsémantique*).

Mésosémantique : Sémantique des unités linguistique à l'intérieur de la phrase (voir aussi *Microsémantique*).

Microsémantique : Chez Rastier, sémantique du palier inférieur du texte ; elle prend pour limite supérieure la sémie (unité lexicale signifiante). Nous employons le terme dans un sens différent de celui de Rastier dans la mesure où nous ne focalisons pas l'analyse sur la notion de sème. Nous reprenons en revanche à Rastier l'idée d'un continuum et d'une similarité de traitements au sein d'une sémantique applicative qui s'étend du mot au texte. L'analyse fine des contextes permet une analyse différentielle des unités afin d'analyser les nuances et les glissements de sens.

Prédicat (schéma prédicatif) : En linguistique, un prédicat est un conglomérat lexical constitué d'une tête (exprimant une relation) et de compléments (arguments de la relation). On parle aussi de cadre prédicatif pour désigner l'ensemble de la séquence. Dans l'exemple suivant : « *la vente de l'entreprise au concurrent* », le prédicat (la tête) est le mot *vente* qui supporte la relation exprimée ; la structure comprend deux arguments : *l'entreprise* et *le concurrent*.

Questions-réponses : Tâche qui vise à répondre de manière précise à des questions posées en langage naturel, en cherchant des réponses au sein d'un ensemble de textes. La réponse est généralement un élément précis répondant à une question factuelle. Récemment se sont développés des systèmes visant à répondre à différents types de questions (explications, procédures, *etc.*).

Recherche d'information : Tâche qui vise à extraire un ensemble de documents à partir d'un fonds documentaire, en fonction d'une requête utilisateur. Contrairement à ce que le nom peut laisser penser, la recherche d'information porte sur des documents entiers et non sur des éléments précis d'information (à l'inverse de l'extraction d'information).

Résumé automatique : Tâche visant à produire un texte cible résumant un texte source. Il existe aussi du résumé multi-documents, il s'agit alors de produire un texte synthétisant plusieurs documents sur un thème donné. Il existe de nombreux raffinements par rapport

à ce schéma de base (produire un résumé orienté par rapport à une requête, un résumé synthétisant des opinions, *etc.*).

Sous-catégorisation (schéma de sous-catégorisation) : Dans un schéma prédicatif, la sous-catégorisation désigne le nombre, la nature syntaxique et le rôle sémantique des arguments (rôle thématique, rôle sémantique, restriction de sélections).

Texte : On reprend ici la définition de Rastier [2001] : « Un texte est une suite linguistique autonome (orale ou écrite) constituant une unité empirique et produite par un ou plusieurs énonciateurs dans des pratiques sociales attestées. Les textes sont l'objet de la linguistique ».

Type de textes : Catégorie de textes fondée sur l'existence de traits linguistiques communs ou d'un critère pertinent au regard d'un objectif (applicatif ou autre). Les types ne correspondent pas obligatoirement à des pratiques sociales définies. Ils peuvent émerger *a posteriori*, par l'analyse d'un corpus de textes. Une question ouverte consiste à savoir si on peut identifier des genres à partir de traits linguistiques (autrement dit, s'il est possible d'« unifier » les genres et les types). Les résultats semblent plutôt négatifs pour l'instant.

Valence (schéma de valence) : Dans un schéma prédicatif, la valence désigne le nombre et le rôle des arguments. Un schéma de valence n'inclut pas obligatoirement d'information syntaxique sur les arguments, à l'inverse du schéma de sous-catégorisation.

Index des notions

Les entrées en majuscules réfèrent à des projets ou des campagnes d'évaluation liés à mon travail récent.

Liste des auteurs cités

Seule la référence au nom du premier auteur est indexée.