Mémoire d'Habilitation à diriger les Recherches
Spécialité : Informatique

Donner accès au contenu des documents textuels

Acquisition de connaissances et analyse de corpus spécialisés

Adeline Nazarenko
Laboratoire d'Informatique de Paris-Nord
Université Paris-Nord

Présentée le 10 décembre 2004 devant le jury composé de :
Madame Maria Tereza Pazienza (rapporteur)
Monsieur Benoît Habert (président)
Monsieur Daniel Kayser (examinateur)
Monsieur François Lévy (examinateur)
Monsieur Jean-Marie Pierrel (rapporteur)
Monsieur Bernard Victorri (rapporteur)

Remerciements

Je tiens à remercier les membres du jury de l'intérêt qu'ils ont manifesté pour mon travail. Je suis très reconnaissante à Maria Tereza Pazienza, Jean-Marie Pierrel et Bernard Victorri d'avoir accepté d'être rapporteurs. Leurs remarques ainsi que celles de Benoît Habert sont précieuses pour la poursuite de ma réflexion. Daniel Kayser et François Lévy m'ont convaincue de me lancer dans ce travail. Je les remercie sincèrement de leurs conseils et de leur soutien.

***

Rédiger un mémoire d'Habilitation à Diriger les Recherches, c'est toujours faire un bilan des années de recherches écoulées. Je tiens à dire combien les rencontres avec d'autres chercheurs ont enrichi ma réflexion.

Je voudrais en premier lieu exprimer ma gratitude à Thierry Guillotin et Benoît Habert avec qui j'ai travaillé au Centre Scientifique d'IBM France et à l'Ecole Normale Supérieure de Fontenay-St Cloud avant de rejoindre le LIPN. Collaborer avec eux a été tout à la fois enrichissant et stimulant. J'ai beaucoup appris à leurs côtés. Les échanges de vues avec mes collègues de l'équipe "Représentation des Connaissances et Langage Naturel" (RCLN) du LIPN, au cours de ces dernières années, m'ont aidé à prendre du recul dans mon travail : je leur sais gré de leur questionnement exigeant et bienveillant.

Il est difficile de dire la fécondité des discussions des groupes de travail ESPOIR, Terminologie et Intelligence Artificielle, A3CTE et ASSTICCOT auxquels j'ai participé tant les collaborations sur lesquelles ils ont débouché sont diverses en nature et en importance. Je voudrais exprimer ma gratitude à tous ces collègues, notamment à Nathalie Aussenac-Gilles, Anne Condamines, Claire Nédellec, Monique Slodzian, Philippe Bessières, Gilles Bisson, Jacques Bouaud, Didier Bourigault, Jean Charlet, Benoît Habert, Claude de Loupy, François Yvon et Pierre Zweigenbaum.

Ma réflexion scientifique s'est également enrichie des nombreux échanges que j'ai eu avec les jeunes chercheurs que j'ai encadrés : Thierry Hamon et Thierry Poibeau qui sont aujourd'hui des collègues, mais aussi Touria Aït El

Mekki, Danièle Cohen, Goritsa Ninova, Christine Reynaud, Frederik Cailliau, Olivier Hamon, Guillaume Vauvert et Davy Weissenbacher. Je souhaite dire ici le plaisir que j'ai eu ou que j'ai à travailler avec eux.

***

Je voudrais exprimer ma reconnaissance à Daniel Kayser, Jacqueline Vauzeille et Christophe Fouqueré de la confiance qu'ils m'ont accordée. A la tête de l'équipe RCLN ou du laboratoire, ils ont soutenu mes projets tout en me laissant une grande liberté d'action.

Je tiens également à remercier mes collègues Brigitte Biébow, Françoise Gayral, Catherine Récanati, Sylvie Salotti, Sylvie Szulman, Thierry Hamon, Daniel Kayser et François Lévy de la confiance qu'ils m'ont témoignée en me donnant la responsabilité de l'équipe RCLN il y a deux ans.

L'ambiance chaleureuse et respectueuse du LIPN en fait un lieu propice à l'épanouissement scientifique. J'en remercie tous mes collègues.

***

J'ai pour finir une pensée pour Jean Fargues qui m'a accueillie en thèse au Centre Scientifique d'IBM France et qui nous a quittés en 1997.

Résumé

Au-delà de la recherche d'information qui se contente de sélectionner des documents dans une base documentaire, l'un des enjeux majeurs de l'accès à l'information consiste à développer des outils permettant l'exploration du contenu même des documents textuels. Cet objectif pose des défis importants à l'Informatique et notamment au Traitement Automatique des Langues : il s'agit de concevoir des méthodes et des outils capables de "comprendre" les documents.

La difficulté vient de ce qu'il faut avoir des connaissances pour comprendre : les outils d'accès au contenu des documents reposent sur des dictionnaires, ontologies, règles d'interprétation qui varient d'un domaine et d'une application à l'autre. Analyser des corpus va donc de pair avec l'acquisition des connaissances : les connaissances acquises sont exploitées pour accéder au contenu des documents et l'analyse de corpus représentatifs d'un domaine de spécialité sert à acquérir les connaissances liées à ce domaine d'activité.

Nous montrons qu'on peut acquérir des classes sémantiques, des relations terminologiques et plus largement des règles d'extraction à partir de corpus et que ces connaissances sont exploitables pour accéder au contenu des documents textuels. Trois méthodes différentes d'accès au contenu sont explorées, chacune dans un cadre particulier : l'indexation fine de document vue comme la construction d'un réseau hypertextuel devant guider la navigation du lecteur ; l'extraction d'information pour assister les biologistes dans leur travail de fouille de textes ; la recherche d'information spécialisée.

Au-delà des divergences dans les objectifs et l'architecture des systèmes à développer, il apparaît que l'analyse repose toujours sur les mêmes techniques : repérage des termes et des entités nommées, calcul de pertinence, étiquetage morpho-syntaxique et parfois analyse syntaxique. Ces techniques sont adaptées au cas par cas aux spécificités du corpus et de la tâche considérés. Ceci amène à défendre une conception du traitement automatique des langue reposant sur la modularité et l'adaptation : un ensemble de modules élémentaires génériques pouvant être diversement combinés en fonction de la

tâche visée et adaptés par l'intégration de connaissances acquises en fonction de la même tâche.

Table des matières

1 Introduction **9**

2 Comprendre les textes **13**
2.1 Un objectif ambitieux 14
2.1.1 Que signifie « comprendre un texte ? » 14
2.1.2 Qu'est-ce qu'un système de compréhension de texte ? . 14
2.2 Une expérience : le projet KALIPSOS 17
2.2.1 Un projet ambitieux 17
2.2.2 Une analyse classique combinant syntaxe et sémantique 18
2.2.3 Les graphes conceptuels comme formalisme sémantique 18
2.3 Limites 20
2.3.1 Les limites de la sémantique logico-référentielle 20
2.3.2 Une complexité d'analyse difficile à maîtriser 22
2.3.3 Une compréhension impossible à évaluer ? 23
2.4 Bilan 23

3 Acquérir des connaissances **25**
3.1 Pourquoi partir de textes ? 26
3.2 Contexte 27
3.2.1 L'essor de l'analyse de corpus 28
3.2.2 Le renouveau de l'analyse terminologique 30
3.2.3 Le renouveau des ontologies 31
3.2.4 Points critiques 32
3.3 Acquérir des classes sémantiques 33
3.3.1 Démarche 33
3.3.2 Explorer des cartographies lexicales 34
3.3.3 Evaluer l'acquisition 35
3.3.4 Bilan 38
3.4 Acquérir des relations lexicales 39
3.4.1 Problématique 40
3.4.2 Approche structurelle : SynoTerm 41

3.4.3 Approche par règles : les noms de gènes renommés . . . 43
3.4.4 Bilan partiel . . . 46
3.5 Acquérir des règles d'extraction ? . . . 47
3.5.1 Assister l'écriture de règles . . . 48
3.5.2 Le recours aux techniques l'apprentissage . . . 48
3.5.3 Apprendre des règles à partir de corpus . . . 49
3.6 Conclusion . . . 53

4 Donner accès au contenu 57
4.1 Indexation fine de documents . . . 59
4.1.1 Approche générale . . . 61
4.1.2 Méthode d'acquisition . . . 61
4.1.3 Résultats et évaluation . . . 63
4.1.4 Bilan et perspectives . . . 64
4.2 Extraction d'information . . . 66
4.2.1 Approche générale : acquisition et production . . . 67
4.2.2 La mise en place d'une chaîne de traitement . . . 71
4.2.3 Aller plus loin dans la normalisation . . . 77
4.3 Vers la recherche d'information spécialisée ? . . . 78
4.4 Conclusion . . . 81

5 Bilan et perspectives 83

A Présentation du formalisme des Graphes Conceptuels 89

B Liste des contrats 91

Chapitre 1

Introduction

Comprendre mobilise des connaissances mais il faut comprendre pour acquérir des connaissances.

Voilà plus de dix ans que je fais face à ce paradoxe sur les textes : on ne peut comprendre un texte sans connaissances sur la langue et sur le contexte, mais où acquiert-on la maîtrise de la langue et toutes les connaissances nécessaires, si ce n'est d'abord dans les textes ? Dix ans plus tard, j'ai toujours l'ambition de transformer le paradoxe en cercle vertueux. Avec l'idée que dans un processus en spirale, les connaissances acquises puissent servir à mieux comprendre les textes et à en apprendre davantage.

Les mots ont changé : je ne parle plus de « comprendre les textes » mais de « donner accès aux documents ». C'est mettre davantage l'accent sur les applications et le service rendu à l'utilisateur. Donner accès au contenu des documents textuels est en effet un défi important dans notre société de la communication, où le volume de la documentation à consulter ne cesse de croître. Les techniques de Traitement Automatique des Langues (TAL) actuelles, éventuellement combinées à d'autres, permettent de répondre directement aux questions de l'utilisateur en interrogeant le web, de caractériser succinctement les mails de la clientèle, d'analyser et de comparer des textes de brevets.

La difficulté vient de ce que les outils d'analyse de texte génériques voient leur performance s'effondrer quand ils sont testés sur des textes techniques ou scientifiques, qui ont souvent un fonctionnement langagier particulier. Il faut donc mettre au point des outils d'analyse qui puissent être adaptés en fonction du domaine d'application, par l'ajout de connaissances linguistiques et extralinguistiques dédiées à cette application.

On retrouve la problématique initiale, avec deux axes de recherche qui s'enrichissent mutuellement : l'acquisition élabore des connaissances qui sont ensuite utilisées pour analyser les textes, les méthodes d'analyse pouvant

aussi servir l'acquisition. Le défi que j'ai essayé de relever dans mon travail de recherche consiste à déterminer comment combiner au mieux acquisition et analyse.

Point de vue adopté

Ce mémoire présente un parcours de recherche qui s'étale sur une dizaine d'années, depuis la fin de ma thèse de doctorat, soutenue début 1994, jusqu'à aujourd'hui. Il montre comment l'ensemble de mes travaux de recherche s'organisent autour de deux problématiques principales : l'acquisition de connaissances à partir de textes et la conception d'outils d'accès au contenu des documents spécialisés[1]. Je n'entre pas dans la description précise et technique de ces travaux, qui ont été publiés par ailleurs. Je cherche surtout à expliciter le parcours qui a été le mien : les questions qui se sont posées à chaque étape, la justification des choix effectués, leur enchaînement, et le bilan de chaque expérience.

Ce document n'est pas organisé de manière strictement chronologique. Même s'il débute sur le bilan de mon travail de thèse antérieur, il présente séparément deux lignes de travail qui ont été développées en parallèle et souvent de manière entremêlée. Je ne cherche pas à retracer l'historique précis de ces travaux, qui a souvent été en partie dicté par les circonstances. J'explicite en revanche à chaque fois le contexte dans lequel ils se sont déroulés, en précisant les conditions scientifiques, humaines et financières de leur réalisation. Réunir toutes ces conditions a fait partie de mon travail.

Le parcours qui est décrit ici n'est pas isolé. Il s'inscrit au contraire dans le mouvement général qui a remodelé le TAL depuis une quinzaine d'années, sous la pression de différents facteurs : l'émergence des corpus et des données réelles, la modularisation accrue des traitements, leur lexicalisation, l'essor des méthodes statistiques et l'importance prise par la problématique des ontologies.

Ce parcours n'est pas solitaire non plus. La plupart des chantiers présentés dans ce mémoire sont le fruit de collaborations : avec les jeunes chercheurs que j'ai encadrés, avec des partenaires industriels, des collègues et d'autres disciplines. Au final, il est impossible de retracer précisément la contribution de chacun mais les échanges et les collaborations ont été fructueuses sur bien des points. Les travaux présentés ici en sont le résultat.

[1] Ce mémoire passe ainsi sous silence d'autres thématiques de recherche auxquelles je m'intéresse en contrepoint, la réflexion sur la notion de causalité et son expression notamment.

Plan du mémoire

Le premier chapitre fait le bilan de mon travail de doctorat dans un projet de développement d'un système générique de compréhension de textes. Cette expérience a mis en évidence le rôle crucial des ressources (grammaires, lexiques, ontologies) dans un système d'analyse de textes et la nécessité de prendre en compte l'application dans le processus d'analyse. Le bilan de ce projet a déterminé la suite de mon parcours.

Le deuxième chapitre montre l'intérêt de l'acquisition à partir de corpus. Cette approche permet d'exploiter la richesse des informations présentes dans les textes. Elle assure aussi une bonne articulation entre les niveaux conceptuel et linguistique, ce qui est indispensable dès lors que les connaissances doivent aussi servir l'analyse linguistique. Différents types de connaissances peuvent être acquis : des classes sémantiques, la terminologie du domaine considéré, des relations sémantiques structurant l'espace des termes, voire des règles d'extraction. Acquérir des connaissances à partir de textes est cependant un processus complexe qui ne peut être totalement endogène : il faut articuler les connaissances issues du corpus, les connaissances préexistantes et une expertise humaine dans un processus coopératif.

Le troisième chapitre présente les trois méthodes différentes d'accès au contenu des documents textuels sur lesquelles j'ai travaillé, chacune dans un cadre particulier : l'indexation fine de documents vue comme la construction d'un réseau hypertextuel devant guider la navigation du lecteur ; l'extraction d'informations pour assister les biologistes dans leur travail de fouille de textes ; la recherche d'informations spécialisées. Au-delà des divergences dans les objectifs et l'architecture des systèmes à développer, il apparaît que l'analyse repose souvent sur les mêmes techniques fondamentales : repérage des termes et des entités nommées, calcul de pertinence, étiquetage morpho-syntaxique et parfois analyse syntaxique. Ces techniques doivent être adaptées au cas par cas aux spécificités du corpus et de la tâche considérés. Pour le traitement de corpus spécialisés, ceci m'amène à défendre une conception du TAL reposant sur la modularité et l'adaptation : un ensemble de modules élémentaires génériques pouvant être diversement combinés en fonction de la tâche visée et adaptés par l'intégration de connaissances, elles-mêmes acquises en fonction de cette tâche.

Le dernier chapitre conclut le mémoire et synthétise les perspectives qui s'en dégagent.

Chapitre 2

Comprendre les textes automatiquement : bilan d'une expérience

Au cours de ma thèse de doctorat, j'ai été associée au projet KALIPSOS du centre scientifique d'IBM à Paris. Ce projet ambitieux visait à développer un système générique de compréhension automatique de texte et de questions-réponses en langue naturelle [Bérard-Dugourd et al., 1989]. Par son ambition, ses objectifs et ses présupposés, KALIPSOS s'apparentait à d'autres projets développés de manière plus ou moins concomitante au niveau européen[1]. Le système KALIPSOS a lui-même joué un rôle central dans le projet européen Menelas [Zweigenbaum et Consortium MENELAS, 1994].

En 1990, lorsque j'ai rejoint le projet, un premier prototype existait et j'ai participé à l'enrichissement du système initial. En cherchant à mieux rendre compte de l'expression de la causalité, j'ai tenté d'améliorer les performances du systèmes KALIPSOS relativement aux questions causales, ce qui supposait d'intervenir à la frontière entre l'analyse syntaxique, l'analyse sémantique et le module de raisonnement.

Je reviens ici sur le bilan de cette expérience parce qu'il a en grande partie déterminé mes choix scientifiques ultérieurs. La question de la compréhension de texte ouvre ce chapitre : c'était le but du projet KALIPSOS et cela reste un objectif à atteindre, même si on emprunte aujourd'hui des chemins détournés. Les deuxième et troisième sections décrivent à grands traits l'approche retenue pour le projet KALIPSOS et ses limites. La quatrième section présente les conclusions que ce travail m'a permis de tirer et jette les bases

[1] Comme le projet Acord [Bès et Guillotin, 1992] ou le projet Lilog dont le bilan a été dressé dans [Herzog et Rollinger, 1991].

de mon travail ultérieur.

2.1 Un objectif ambitieux : la compréhension de texte

2.1.1 Que signifie « comprendre un texte ? »

Il faut considérer ici le « texte » dans un sens large de « production écrite »[2] : il peut s'agir de tout type de textes, quels qu'en soient le genre, la taille ou la complexité. On s'est ainsi intéressé à comprendre automatiquement de courts messages au format assez stéréotypé, comme des dépêches de presse [MUC, 1995] ou des articles journalistiques ou scientifiques [Bessières et al., 2001].

Que signifie « comprendre un texte » ? On s'accorde à reconnaître que cette faculté de compréhension est essentielle et largement partagée mais la définir est délicat. Ce n'est en tout cas pas une faculté binaire : on peut « comprendre un texte » sans le comprendre « vraiment » et sans le comprendre « entièrement ». Le fait de ne pas connaître certains mots ou de ne pas comprendre certaines phrases n'interdit pas toujours de comprendre le texte globalement.

De manière abstraite, on peut considérer que « comprendre un texte » signifie être capable de modifier sa représentation du monde en fonction des informations véhiculées par le texte. Cela suppose qu'un être humain ou un système intelligent dispose d'un ensemble de connaissances qui constitue sa vision de son environnement physique, intellectuel, social et symbolique. Dans cette perspective, la compréhension se traduit par l'ajout, la suppression ou la correction de connaissances. En pratique, le niveau de compréhension dépend de l'objectif visé et de la nature du texte considéré. On ne lit pas un texte de loi ou une police d'assurance comme un article de presse, un manuel scolaire comme une notice pharmaceutique. En soi, la compréhension n'est pas une tâche. C'est une activité préalable à de nombreuses tâches, comme le résumé, la traduction, l'exécution d'instructions...

2.1.2 Qu'est-ce qu'un système de compréhension de texte ?

Si la compréhension de texte n'est pas une tâche en tant que telle mais une activité nécessaire à l'accomplissement d'autres tâches, il est illusoire de

[2]Par opposition au discours oral qui soulève des difficultés de compréhension supplémentaires spécifiques, sur lesquelles je n'ai pas travaillé.

parler de systèmes de compréhension de texte. Il n'existe que des systèmes utilisant une forme limitée de compréhension nécessaire à la tâche visée, que ce soit dans un cadre de questions-réponses, pour extraire des informations clefs ou même pour rechercher des documents dans une base documentaire. Au plus bas niveau, indexer un vaste ensemble de documents comme le font les systèmes de recherche d'information, c'est déjà avoir une compréhension (certes limitée) du contenu de ces documents. A l'opposé, un système de questions-réponses capable de répondre à des questions complexes sur un texte quelconque suppose une compréhension beaucoup plus précise du sens de ce texte.

Par « système de compréhension de texte », je considère ici tout système qui produit une représentation du sens du texte et qui utilise cette représentation sémantique comme point de départ d'un processus inférentiel. En ce sens, il existe des systèmes de compréhension de textes très divers.

Avant de poursuivre, il est utile de mentionner quelques grands types de systèmes traditionnels qui répondent à cette définition. Je reviendrai à la fin de ce document sur des approches plus émergentes.

Recherche d'information ou recherche documentaire

Les systèmes de recherche d'information[3] s'inscrivent dans une longue pratique documentaire. Ils visent à retrouver dans une base documentaire un sous-ensemble de documents pertinents au regard d'une requête formulée par l'utilisateur. Dans les moteurs de recherche aujourd'hui disponibles sur le web, les systèmes de recherche d'information sont associés à un butineur (*crawler*) qui construit et met à jour une base documentaire à partir des pages accessibles sur le web.

Extraction d'information

La notion d'extraction d'information a été précisément définie à partir de 1987 par le programme des Conférences sur la compréhension de messages de la DARPA (MUC, pour Message Understanding Conference) [MUC, 1995, MUC, 1998, Pazienza, 1997]). L'extraction d'information consiste, dans un domaine retreint, à extraire des éléments d'information précis à partir d'un ensemble de textes homogènes et à remplir des formulaires prédéfinis avec ces éléments d'information.

Entre 1987 et 1998, les systèmes d'extraction d'information ont été testés sur différents types de textes : récits d'attentats (MUC-3, MUC-4), dépêches d'agence de presse concernant des joint-ventures ou le secteur financier (MUC-6), annonces de produits du secteur microélectronique

[3]Bien qu'ils ne sont d'ordinaire pas comptés parmi les systèmes de compréhension de texte, les systèmes de recherche d'information répondent à la définition donnée ci-dessus, un système d'index étant une représentation sémantique fruste.

(MUC-5), etc. La tâche consiste toujours à remplir un formulaire (*template*) fixé à l'avance en fonction du domaine : dans le cas des spectacles, il faut identifier le lieu et la date de l'événement, son auteur, son type, etc. (voir figure 2.1).

Questions-réponses

Les systèmes de questions-réponses visent à retrouver dans un document ou une base documentaire une réponse à la question posée par l'utilisateur, la réponse étant fournie sous une forme textuelle et *a priori* sans dialogue avec l'utilisateur. On trouve de tels systèmes de questions-réponses intégrés aux outils d'aide de certains logiciels grand public. L'idée est ancienne [Simmons, 1965, Lehnert, 1977] mais, comme dans le cas de l'extraction d'information, les recherches autour des systèmes de questions-réponses ont été réactivées par les campagnes d'évaluation « Question/Answering » des Text Retrieval Conferences (TREC) [Voorhees et Tice, 1999].

Les systèmes de questions-réponses sont très proches des précédents dans leurs objectifs et dans leur méthodes mais l'extraction d'information vise à répondre de manière fiable et exhaustive à des questions prédéfinies de l'utilisateur, qui peuvent être complexes, alors que les systèmes de questions-réponses sont généralement conçus pour répondre à des questions tout-venant et donc plus simples (par ex. la question *Dans quelle ville se trouve le Colisée ?* lancée sur le web).

Résumé automatique

La notion de résumé est ancienne elle aussi. Le résumé d'un texte ou d'un ensemble de textes T est un texte secondaire T' qui doit être plus court que T, rendre compte fidèlement de l'ensemble du contenu de T tout en restant cohérent et lisible. Un système de résumé automatique vise à produire ce type de résumé, soit sous la forme textuelle traditionnelle soit dans un autre langage de représentation, pour le ou les textes donnés en entrée et éventuellement en tenant compte de contraintes fixées par l'utilisateur (longueur du texte, par exemple). La recherche dans ce domaine est active depuis longtemps [Minel, 2002].

Malgré leur très grande utilité et leur succès auprès d'un large public, les techniques de recherche d'information ne permettent pas de répondre à des besoins d'information précis dans la mesure où elles n'appréhendent que le niveau du document et ne donnent pas directement accès au contenu informationnel des documents eux-mêmes. L'utilisateur doit encore lire les documents issus de la recherche d'information. Accéder au contenu est aujourd'hui un enjeu important pour tout le domaine de la gestion de l'information et des connaissances : les trois autres types de systèmes vont dans ce sens, chacun

TEXTE
19 mai 2003. Paul McCartney à Rome - Le 10 Mai dernier, Paul McCartney donnait à Rome, au Colisée, un concert exceptionnel.

FORMULAIRE

Spectacle	type	concert
	date	10 mai 2003
	auteur	Paul McCartney
	ville	Rome
	lieu	Colisée

FIG. 2.1 – *Extraction d'information : exemple de formulaire de spectacle rempli par extraction à partir d'un texte de dépêche*

à leur manière.

2.2 Une expérience : le projet KALIPSOS

Cette section décrit les grandes lignes de l'approche retenue dans le projet KALIPSOS, qui est à la fois classique dans la méthode d'analyse et originale par le formalisme utilisé pour la modélisation linguistique.

2.2.1 Un projet ambitieux

Le système KALIPSOS a été conçu par Jean Fargues comme un système générique. Générique d'abord par rapport aux tâches visées. Si le système a d'abord été pensé comme un système de questions-réponses, il n'a pas été conçu exclusivement dans cet objectif. Il s'agissait au contraire de construire un système polyvalent, en s'appuyant sur une analyse sémantique suffisamment solide pour être utilisable dans différents processus inférentiels[4]. Générique également par rapport au domaine d'application, même si l'adaptation d'un domaine à l'autre peut imposer de reconstruire les lexiques et bases de connaissances nécessaires.

L'exhaustivité est la seconde ambition du projet. Il s'agissait à la fois de comprendre tout *le* texte et tout *du* texte. Vouloir comprendre la causalité

[4]Il a ainsi été utilisé pour extraire des éléments de réponse [Bérard-Dugourd et al., 1989], faciliter l'accès et l'exploitation des compte-rendus médicaux dans le cadre du projet Menelas [Zweigenbaum et Consortium MENELAS, 1994], filtrer des dépêches financières [Landau et al., 1993].

dans les textes [Nazarenko, 1994] relevait de la même ambition : c'était s'attaquer à la fois à des phrases complexes (avec souvent des ambiguïtés de portée et mettant en jeu les deux niveaux factuels et argumentatifs du discours) et au raisonnement causal dont les mécanismes cognitifs sont aujourd'hui encore mal formalisés [Dague et al., 2004].

2.2.2 Une analyse classique combinant syntaxe et sémantique

La méthode d'analyse de KALIPSOS est donc guidée par le texte, plutôt que par l'application. L'analyse vise à produire une représentation sémantique du texte analysé.

Les principaux composants sont l'analyseur morpho-syntaxique et l'analyseur sémantique, qui exploitent respectivement un lexique morpho-syntaxique et un lexique sémantique. L'analyse repose sur une approche séquentielle et compositionnelle classique [Gayral, 1998]. L'analyse syntaxique produit un arbre d'analyse pour chaque phrase (plusieurs en cas d'ambiguïté). Le lexique sémantique associe à chaque mot une représentation sémantique (plusieurs en cas de polysémie). Ces définitions associent une catégorie sémantique à chaque mot non vide et un schéma de sous-catégorisation aux prédicats pour spécifier le nombre et le type sémantique de leurs arguments. Ces définitions sémantiques jouent le rôle de briques élémentaires pour l'analyse sémantique qui les combine en structures de plus en plus complexes, l'arbre d'analyse syntaxique spécifiant quels éléments doivent être combinés et dans quel ordre. Une phrase est supposée correctement analysée si une structure sémantique globale a pu lui être associée. Les ambiguïtés sont en principe levées au fil de l'analyse sémantique, les règles de composition ne permettant pas de construire des structures sémantiques pour toutes les analyses syntaxiques.

2.2.3 Les graphes conceptuels comme formalisme sémantique

L'une des originalité du projet KALIPSOS est de reposer sur le formalisme des graphes conceptuels pour modéliser la sémantique des langues naturelles. Ce formalisme a été défini par J. Sowa [Sowa, 1984, Sowa, 1992] dans la lignée des travaux sur les réseaux sémantiques[5].

Les graphes conceptuels constituent un formalisme (pas un modèle séman-

[5]Une présentation rapide de ce formalisme est donné en annexe. Voir [Chein et Mugnier, 1991] pour une présentation d'ensemble

tique[6]) mais s'il est « ontologiquement neutre » comme le précise J. Sowa, le formalisme choisi guide le travail de modélisation et induit certains choix épistémologiques. Il est donc intéressant de s'interroger sur les implications de ce choix sur la modélisation sémantique.

Une sémantique logico-référentielle

Le choix des graphes conceptuels comme formalisme de représentation sémantique traduit le parti pris d'une sémantique que j'appelle logico-référentielle pour souligner à la fois son fondement logique et l'approche référentielle qui la sous-tend.

J. Sowa définit une sémantique logique pour les graphes simples : représenter une phrase par un graphe conceptuel revient ainsi à la traduire en une formule logique des prédicats du 1^{er} ordre qui s'interprète en termes vériconditionnels.

Cette démarche s'inscrit dans une approche référentielle courante en sémantique lexicale [Lyons, 1980] : dans le lexique sémantique de KALIPSOS, à chaque entrée lexicale est associée un ou plusieurs (en cas de polysémie) concepts ou graphes conceptuels. Ce parti pris référentiel se retrouve dans la sémantique grammaticale qui combine les éléments de sens précédents pour associer une représentation sémantique à la phrase, elle aussi référentielle : le sens d'une phrase repose sur l'existence de référents, qui dénotent des entités concrètes ou abstraites du monde réel, et sur l'énoncé des propriétés (classes et relations) de ces référents.

Connaissances ontologiques *vs.* connaissances assertionnelles

De manière intéressante, le formalisme des graphes conceptuel a permis de poser la question de la distinction entre connaissances du domaine et connaissances assertionnelles véhiculées par le texte, qui correspond à la dichotomie entre les niveaux terminologique et assertionnel des logiques de description [Brachman, 1979].

Un langage de graphe se définit par un *support* qui introduit le « vocabulaire » du domaine. Il introduit essentiellement la hiérarchie des types de concepts ainsi que les signatures des types de concepts et des relations. Il exprime les connaissances générales du domaine. Dans la mesure où ces connaissances du domaine sont formellement organisées autour d'une hiérarchie de types, on peut parler ici de « connaissances ontologiques »[7]. Les

[6] A la différence de la théorie des représentations discursives [Kamp, 1981], par exemple.

[7] Le terme de « connaissances terminologiques », classique en logique de description, paraît peu adapté ici dans la mesure où les connaissances exprimées dans le support peuvent

connaissances exprimées par le texte analysé sont elles aussi traduites en graphes conceptuels, de taille et de complexité variables.

L'opposition entre support et graphes d'analyse ne recouvre cependant pas l'opposition entre connaissances ontologiques et connaissances particulières ou événementielles. Les graphes d'analyse, comme les phrases, peuvent aussi traduire des connaissances ontologiques. Si un support est nécessaire pour analyser un texte, les connaissances modélisées à partir du texte peuvent aussi venir enrichir l'ontologie. Ce point est intéressant à noter parce qu'il justifie l'acquisition de connaissances à partir de corpus textuels.

2.3 Limites

L'approche défendue dans le projet KALIPSOS a des limites qu'il est intéressant d'analyser : elles concernent à la fois l'approche logico-référentielle, la méthode d'analyse et l'objectif en tant quel.

2.3.1 Les limites de la sémantique logico-référentielle

L'approche référentielle de la sémantique a été critiquée tant sur le plan lexical que sur le plan grammatical [Kayser, 1987, Rastier, 1991] [Condamines, 2003, pp. 16-20]. Je ne fais que souligner ici les points qui m'ont paru les plus critiques dans le projet KALIPSOS.

Une notion de référent problématique

Si la sémantique lexicale traditionnelle tend à considérer que ce sont surtout les noms qui fonctionnent comme expressions référentielles, dans le formalisme des graphes conceptuels, tout concept est associé à un référent, y compris des concepts événementiels associés le plus souvent à des verbes. Ceci soulève une vaste question sur la nature de ces référents événementiels. En pratique, il est relativement aisé d'associer un référent à un nom entièrement déterminé (qui renvoie alors à une entité du monde clairement identifiée, c'est le cas des noms propres mais aussi de certaines déterminations contextuelles) ou à un verbe ponctuel, accompli et passé, avec un agent unique, mais tous les autres cas (référents virtuels, indéterminés, événements répétitifs, etc. [Gayral et al., 1994]) sont problématiques.

être complexes, alors qu'on a souvent une vue réductrice de ce qu'est une terminologie (listes de termes associés aux concepts auxquels ils renvoient). Par ailleurs, « terminologie » est employé dans une perspective différente dans la suite de ce document.

Nous verrons dans la suite que les méthodes actuelles d'accès au contenu des documents s'inscrivent aussi dans la perspective d'une interprétation référentielle des textes, mais en se focalisant sur les entités nommées, elles évincent les aspects problématiques évoqués ci-dessus[8].

Des modalités difficiles à prendre en compte

Les modalités font partie des phénomènes linguistiques les plus difficiles à prendre en compte. Je désigne par le terme de *modalité* les « attitudes propositionnelles » au sens de [Barwise et Perry, 1983, ch. 9], c'est-à-dire non seulement les verbes modaux (*may*, *should*, *can*, etc. et leurs équivalents en français) mais plus largement toutes les tournures de phrases qui traduisent le point de vue ou la distance du locuteur par rapport à ce qu'il affirme.

Pour représenter les modalités, J. Sowa a introduit la notion de contexte dans les graphes conceptuels. Cela permet d'exprimer le fait qu'une modalité porte sur toute une portion de graphe [Sowa, 1992] mais deux problèmes restent entiers : celui de l'ambiguïté, car la portée de la modalité est souvent difficile à cerner avec précision et celui de la valeur de ces modalités, c'est-à-dire de leur interprétation, notamment quand elles sont combinées, ce qui est fréquent.

La question de la représentation de la causalité m'a amenée à travailler sur ces deux points. La question de l'ambiguïté ne peut pas être résolue sur les seules bases linguistiques et il s'agit plus souvent de sous-spécification que d'ambiguïté. J'ai apporté une réponse fruste à la question de la valeur des modalités dans le cadre du projet KALIPSOS. En m'inspirant des travaux de la linguistique de l'énonciation [Récanati, 1979], j'ai proposé de distinguer simplement les modalités qui construisent des contextes transparents et celles qui construisent des contextes opaques. Les premières, à la différence des secondes, ne modifient pas la valeur de vérité et plus généralement l'interprétation des propositions sur lesquelles elles portent [Nazarenko, 1996].

Les méthodes d'accès au contenu des documents occultent largement cette question des modalités, aussi importante soit-elle, et les propositions que je peux faire moi-même (cf. 4.2.3) ne vont guère au-delà de ce que j'avais mis en place dans le système KALIPSOS !

Le texte, grand absent

On l'aura compris, l'approche sémantique mise en oeuvre dans le système KALIPSOS est centrée sur la proposition et la phrase. La dimension

[8] D'où le relatif échec des approches proposées pour remplir par extraction des formulaires plus complexes prenant en compte des séquences d'événements ou de scénarios.

textuelle n'est pas prise en compte et c'est à mon sens l'une des plus grandes faiblesses de cette approche. On verra dans les chapitres suivants, que les méthodes d'accès au contenu donnent une plus grande place au texte lui-même considéré dans sa globalité et sa structure.

2.3.2 Une complexité d'analyse difficile à maîtriser

Indépendamment de l'approche sémantique retenue, la méthode d'analyse du système KALIPSOS a elle aussi montré ses limites. Je ne fais que souligner les plus importantes au regard des travaux actuels : il s'agit des problèmes d'ambiguïté, tant structurelles que lexicales.

Les ambiguïtés de rattachement

Le phénomène est connu, une proportion importante de l'ambiguïté est liée au fait qu'un constituant syntaxique peut se rattacher à différentes têtes. L'ambiguïté de rattachement des compléments prépositionnels est d'autant plus pénalisante qu'on s'intéresse aux textes scientifiques ou techniques où les groupes nominaux sont souvent complexes. L'analyse sémantique compositionnelle de KALIPSOS permet d'éliminer les alternatives qui violent des contraintes de sélection mais c'est insuffisant. C'est le rapprochement avec d'autres passages du même texte qui permet souvent de réduire ces ambiguïtés. Je montrerai plus loin l'intérêt de l'analyse terminologique sur ce point (voir figure 4.7, p. 76, [Bourigault, 1994]).

Les distinctions de sens

Comme souvent, l'approche référentielle du lexique conduit à une représentation discrète de la signification des items lexicaux. Les items lexicaux susceptibles de renvoyer à différents concepts sont considérés comme polysémiques et on leur associe plusieurs définitions différentes. C'est l'approche retenue dans le lexique de KALIPSOS. Décrire un mot par une liste de sens occulte la gradualité et les aspects continus de la sémantique lexicale [Victorri et Fuchs, 1996] ainsi que les phénomènes de déformation du sens, notamment en contexte [Kayser et Abir, 1991, Rastier et al., 1994]. Dans KALIPSOS, cette approche a été adoptée comme une simplification opératoire, mais c'était sous-estimer la difficulté du travail consistant à lister les différents sens d'un mot. Comment être sûr qu'on a pris en compte tous les sens d'un mot (question de complétude) ? Jusqu'où aller dans le nombre et la précision des distinctions de sens (question de granularité) ? Comment organiser les différents sens d'un mot si l'on veut autoriser différents niveaux d'analyse

(question de structuration des entrées lexicales) ? D'un côté, la complexité de l'analyse croît avec le nombre de distinctions de sens tandis que, de l'autre côté, l'absence d'un sens peut bloquer l'analyse d'une phrase.

Cette approche a été justement critiquée sans pour autant qu'on ait pu, me semble-t-il, apporter des solutions satisfaisantes et exploitables à grande échelle. Il faut trouver un compromis entre la taille du lexique et la complexité de l'analyse mais sur quels critères établir ce compromis *a priori*, si ce n'est à l'aune d'une application et à partir d'un corpus ? C'est la piste qu'explore la terminologie textuelle (voir p. 31).

2.3.3 Une compréhension impossible à évaluer ?

Certains projets ont cherché à définir des protocoles d'évaluation des outils de compréhension de texte et de la qualité de la représentation sémantique qu'ils construisent : au niveau européen (voir le projet FraCaS[9]) comme au niveau national, avec l'Action de Recherche Concertée « Compréhension de textes »[10] [Blache et al., 2000], à laquelle j'ai participé.

Les résultats de ces projets ont été mitigés [Nazarenko et Poibeau, 2004]. Le projet FraCas a permis d'élaborer un jeu de tests destiné à évaluer la capacité des analyseurs sémantiques, mais l'approche proposée est centrée sur les théories et méthodes de sémantique computationnelle. Les phénomènes traités sont les phénomènes sémantiques connus pour être « difficiles » et ils sont considérés isolément. Le fragment des phénomènes sémantiques considéré par FraCaS est-il représentatif des problèmes que posent les applications réelles ? Quelle part prennent-ils dans la bonne compréhension globale du texte ?

L'échec de ces tentatives d'évaluation pose question. Si on ne sait pas évaluer ce qu'est une « bonne » représentation sémantique, c'est sans doute qu'il n'existe pas de bonne représentation sémantique en soi. La compréhension n'est pas une tâche en tant que telle et ce n'est qu'en regard d'une tâche particulière (recherche d'information, résumé automatique, extraction d'information, question/réponse) que la compréhension peut être évaluée (voir [Chaudiron, 2004]).

2.4 Bilan

Au terme de cette expérience, deux voies de recherche paraissent prioritaires : définir une méthode d'acquisition de connaissances et mettre l'application au centre de l'analyse.

[9] « A Framework for Computional Semantics », 1994.

[10] Action de Recherche Concertée de AUPELF-UREF & CNRS (ARC4, 1995-97).

Un nouveau défi : acquérir des connaissances

L'utilisation du formalisme des graphes conceptuels a permis d'opposer formellement les connaissances ontologiques et les connaissances assertionnelles : les premières sont exploitées pour le calcul sémantique alors que les secondes sont construites par l'analyse sémantique des textes. En pratique, construire une base de connaissances ontologiques pour une application (et le lexique sémantique en fait partie) est apparu à la fois difficile et coûteux. En parallèle, l'expérience du projet Menelas a montré l'inadéquation des lexiques antérieurs de KALIPSOS et la nécessité de reconstruire des lexiques adaptés au domaine médical à traiter. Il est ainsi apparu critique de définir des méthodes facilitant la construction de lexiques sémantiques et plus globalement de bases de connaissances ontologiques adaptées pour chaque nouveau domaine. Je me suis alors intéressée à cette question, notamment en essayant d'exploiter les connaissances ontologiques véhiculées par le texte.

Un impératif : mettre l'application au centre de l'analyse

L'expérience de KALIPSOS a montré qu'on ne peut penser l'analyse qu'en référence à une application. Par « application », j'entends à la fois une tâche et un domaine. J'ai souligné ci-dessus l'importance de la tâche. Au-delà des grandes types évoqués plus haut (extraction, questions-réponses, etc.), la tâche demande à être définie avec précision (pour quel type d'utilisateur ? dans quel contexte ? quel est le degré de fiabilité requis ?...). Circonscrire le domaine n'est pas moins important.

Pourquoi s'intéresser à des domaines « spécialisés » ? Par choix et par nécessité tout à la fois. Par choix d'une recherche tournée vers les applications car les besoins documentaires sont pour beaucoup spécialisés. Par nécessité aussi, parce que les domaines spécialisés considérés comme restreints et plus stables sont plus faciles à analyser et à modéliser.

Chapitre 3

Acquérir des connaissances à partir de textes

L'analyse sémantique des textes exploite des ressources de différents types : connaissances morphologiques, connaissances syntaxiques (catégories et règles syntaxiques), lexique sémantique associant une signification aux items lexicaux et, plus largement, les connaissances ontologiques, ensemble de règles d'interprétation ou de connaissances implicites liées au domaine considéré et non rattachées à un item lexical particulier.

Je m'intéresse à l'acquisition des connaissances dans la perspective particulière de l'analyse de corpus textuels : les connaissances à acquérir sont destinées à servir de ressources aux outils de TAL.

Ces connaissances ont *a priori* des degrés de généralité variables, mais cette opposition entre connaissances générales et spécialisées est difficile à formaliser. Mon point de vue est très pragmatique ici. Il s'agit d'une opposition relative : il n'y pas de ressources générales ou spécialisées en soi mais des ressources plus ou moins générales selon le nombre et la diversité des applications auxquelles elles peuvent servir. En pratique, cette notion de « réutilisabilité » (éventail des applications dans lesquelles une ressource peut être utilisée) est souvent davantage présumée que prouvée, même si je pense possible aujourd'hui de se donner des critères explicites d'appréciation[1].

Mon travail de recherche a principalement porté sur les ressources sémantiques (même si elles ne peuvent être pensées indépendamment des niveaux morphologique et syntaxique), qui sont *a priori* les plus complexes à élaborer et les plus coûteuses à construire. Définir des méthodes permettant d'acquérir automatiquement tout ou partie des connaissances constituant ces

[1] Cette question a fait l'objet du travail effectué en commun avec Goritsa Ninova dans le cadre du stage de DEA de cette dernière.

ressources, ou d'en faciliter l'acquisition est donc essentiel et une part importante de mon travail de recherche a porté sur l'acquisition de connaissances à partir de textes.

Ce chapitre se subdivise en cinq sections. Après avoir justifié le fait de partir de textes pour construire de telles ressources, il situe ce travail dans son contexte, à la croisée de trois mouvements scientifiques importants auxquels j'ai participé à des degrés divers : le renouveau de l'analyse terminologique, l'essor de l'analyse de corpus en France et l'émergence de la question des ontologies. Les trois dernières sections présentent trois axes de mon travail sur l'acquisition à partir de textes : l'acquisition de classes sémantiques, de relations sémantiques et de règles d'extraction.

3.1 Pourquoi partir de textes ?

Chercher à acquérir des connaissances à partir de textes – plutôt que par recueil ou introspection comme le ferait un ingénieur de la connaissance ou un linguiste – se justifie d'un point de vue opératoire et pragmatique.

Exploiter l'existant

Dans les méthodes traditionnelles de l'ingénierie des connaissances[2], la mise au jour des connaissances repose souvent sur des entretiens avec les experts. Or, cela a été souvent souligné, les experts sont rares, rarement disponibles et parfois peu à même d'expliciter leurs propres connaissances. Le recours au texte s'est donc imposé comme une alternative intéressante, si on peut trouver des textes représentatifs de l'expertise qu'on cherche à modéliser.

C'est l'objectif qui a animé les travaux du groupe Terminologie et Intelligence Artificielle (TIA)[3]. Ce groupe de travail pluridisciplinaire tend à identifier les convergences et les complémentarités entre les méthodes et les objectifs de la terminologie et ceux de l'intelligence artificielle. En ingénierie

[2] J'emploie le terme d'*ingénierie des connaissances* pour désigner la sous-communauté de l'Intelligence Artificielle s'intéressant à l'*acquisition des connaissances*. Je réserve ce dernier terme à la tâche d'acquisition proprement dite.

[3] Ce groupe de travail pluridisciplinaire rassemblant des chercheurs en linguistique, en intelligence artificielle et en traitement automatique des langues a été créé par Anne Condamines et Didier Bourigault, en 1993. Même si je n'ai formellement rejoint ce groupe de travail qu'en 1998, les échanges ont été en réalité antérieurs, à la fois parce que j'étais arrivée à l'issue de mon travail de thèse à des conclusions très proches des hypothèses de travail du groupe TIA (nécessité de construire des ressources spécialisées, projet d'acquisition de connaissances à partir de corpus) et parce j'ai collaboré étroitement entre 1994 et 1996 avec Benoît Habert, qui était lui-même membre de ce groupe.

des connaissances, créer le modèle conceptuel d'un domaine consiste à rechercher les objets du domaine, à les définir et à les structurer en une ontologie. Le groupe TIA défend l'idée que ces objets sont identifiables à travers l'analyse terminologique du domaine. Concrètement, le groupe TIA vise à élaborer et tester des méthodes et outils de travail sur corpus pour produire de manière systématique des ressources terminologiques riches en informations sémantiques et exploitables dans diverses applications [Aussenac-Gilles et al., 2004].

Il s'agit d'induire des connaissances à partir du matériau textuel. Cette démarche soulève des questions spécifiques, que l'on retrouvera plus loin : sur quels textes convient-il de s'appuyer ? comment passer du texte à une base de connaissances exploitables ? quel biais la dimension linguistique du matériau initial introduit-elle dans le travail de modélisation ?

Articuler les niveaux ontologique et linguistique

Quand il s'agit de construire des ressources exploitables pour des outils de Traitement Automatique des Langues, partir de textes présente un intérêt particulier. Les bases de connaissances produites sont destinées à guider l'interprétation de nouveaux textes, il est donc essentiel de penser l'articulation des deux niveaux ontologique et linguistique. Quelle que soit la richesse d'un réseau de concepts, s'il n'établit pas le lien entre les mots de la langue et ces concepts, il n'est d'aucune utilité pour l'analyse des textes.

Cette démarche d'acquisition des connaissances rejoint de ce fait celle de la linguistique qui a une tradition plus ancienne de construction de dictionnaires[4] ou de grammaires à partir de corpus[5], le rôle de ces derniers étant variable d'une expérience à l'autre.

3.2 Contexte

Le projet d'acquérir des connaissances à partir de textes s'inscrit en réalité à la conjonction de trois courants de recherche. Il me paraît important de le replacer dans ce contexte. Ils sont introduits ici par ordre d'importance relativement à mon travail.

[4]Par exemple le « Collins Cobuild English Language Dictionary » publié en 1987 en anglais ou le Trésor de la Langue Française.

[5]Avec une antériorité des travaux anglo-saxons, comme nous l'avons souligné dans [Habert et al., 1997].

3.2.1 L'essor de l'analyse de corpus

L'intérêt accru porté aux corpus réels dans le TAL est un phénomène majeur des années 1990 en France [Habert et al., 1997]. Le TAL s'intéresse aux corpus à la fois comme source de connaissances (ils servent alors à mettre au point et à tester des outils) et comme objet d'étude (quand l'objectif est d'explorer les corpus). Mon travail s'inscrit dans cette double perspective : il vise à *acquérir des connaissances* à partir de corpus, connaissances qui sont ensuite utilisées pour *analyser des corpus*.

Les mutations du TAL

La prise en compte des corpus a induit des changements dans les objectifs que se fixe le TAL et donc dans les outils qu'il produit. On n'a plus aujourd'hui les ambitions de généricité et d'exhaustivité d'un projet comme KALIPSOS. Le traitement de corpus favorise l'efficacité et la couverture des traitements aux dépens de la précision et de la complétude de l'analyse. On privilégie également la modularité : cela permet que chaque tâche puisse être réalisée par une méthode spécifique et cela facilite le réordonnancement des traitements, les approches séquentielles rigides ayant montré leurs limites. En ce qui concerne la généricité enfin (comprise comme la capacité à traiter un grand nombre de textes), l'ambition d'y parvenir demeure le plus souvent sans qu'on se donne réellement les moyens de la mesurer ni de la réaliser, ce qui supposerait des méthodes d'adaptation, me semble-t-il.

Ce déplacement de perspective s'accompagne naturellement d'un changement dans les méthodes employées. La statistique est convoquée au premier chef et dans tous les domaines, depuis l'étiquetage morpho-syntaxique jusqu'à l'analyse syntaxique et l'étiquetage sémantique [Lebart et Salem, 1994, Klavans et Resnik, 1996, Manning et Schütze, 1999]. Le recours aux méthodes d'apprentissage relève de la même logique, même s'il est plus récent et moins répandu.

Sur le plan sémantique, l'inspiration harrissienne est également intéressante, bien que de moindre portée[6]. Je retiens quatre aspects du travail de Harris [Harris et al., 1989, Harris, 1991] (voir [Habert et Zweigenbaum, 2002] pour une discussion plus approfondie) :

1. La notion de *sous-langage* exprime le fait que, dans certains domaines de spécialité, le fonctionnement linguistique est particulier, avec des classes de mots particulières, des contraintes spécifiques de combinaison et une moindre variation notamment.

[6] B. Habert et moi avons animé une journée d'étude de l'ATALA sur l'« approche distributionnelle de l'analyse sémantique » en janvier 1999.

2. Cette propriété des sous-langages permet de fonder, pour Harris, la méthode de l'*analyse distributionnelle* qui vise à mettre au jour des « grammaires sémantiques »
3. La mise en oeuvre de cette méthode distributionnelle suppose une *normalisation* préalable des phrases.
4. L'objectif, pour Harris, est de mettre au jour des schémas ou *formules informationnelles* [Harris et al., 1989] qui sont censés être stables à travers différentes langues pour des textes du même domaine et du même registre.

Sans nécessairement souscrire au projet informationnel de Harris (point 4 ci-dessus), on peut s'inspirer de la démarche harrissiennne pour acquérir des connaissances à partir de textes. Ce principe a été repris sous différentes formes, avec ou sans référence explicite à Harris. Pour ma part, je n'aborde pas directement la question du sous-langage ou de son statut, la notion restant difficile à cerner au niveau de la langue [Condamines, 1997] [Cabré, 1998, p. 61]. Je préfére parler de discours spécialisé, et pour ce qui me concerne de « corpus spécialisé », sans nier que mes méthodes exploitent la régularité des sous-langages que Harris souligne. L'idée de normalisation a été davantage négligée dans les travaux de TAL. Elle me paraît néanmoins importante, et je tente de l'exploiter (voir section 3.5.3, p. 51).

La notion de corpus et de corpus spécialisé

Qu'entend-on par « corpus » ? Comment construire un corpus ? Ces questions ont reçu des réponses variées, étant donné la diversité des travaux et des disciplines qui prétendent reposer sur des corpus. En linguistique, la réflexion a été initiée par la lexicographie et par la communauté anglo-saxonne qui a eu la première à coeur de constituer des corpus enrichis [Habert et al., 1997].

La définition de J. Sinclair [Sinclair, 1996] est souvent citée, mais je retiens plutôt celle de S. Atkins (1992), citée par [Pearson, 1998, p. 42] : « a subset of [Electronic Text Library] built according to explicit design criteria for a specific purpose [...] ». Elle souligne les points suivants :

- un corpus est un construit et non une donnée ;
- un corpus étant un échantillon d'un ensemble plus vaste, il ne contient jamais la totalité des phénomènes dont on cherche à rendre compte : l'acquisition des connaissances à partir de textes repose donc sur une démarche inductive ;
- il est important d'expliciter les critères (linguistiques ou autres) de construction du corpus, pour apprécier les observations qui peuvent y être faites ou les connaissances qui peuvent en être extraites.

– la construction d'un corpus est toujours subordonnée à un objectif précis, ce que j'ai appelé la visée applicative, liée à la fois à la tâche et au domaine. En ce sens, un corpus est toujours « orienté », me semble-t-il.

La constitution de corpus soulève plusieurs questions [Péry-Woodley, 1995] [Habert et al., 1997, ch. 7]. Je n'en reprends que deux ici.

La question de la *représentativité* est classique. Si le corpus est un échantillon construit, il faudrait qu'il soit représentatif de l'ensemble dont il est issu et que l'on cherche à décrire. Dans les applications de TAL qui nous occupent ici, le corpus est construit de manière très empirique. On fait souvent « avec ce qu'on a » ou « avec ce qu'on nous donne », en se limitant aux textes disponibles sur support électronique, en excluant les données confidentielles, etc. En pratique il faut surtout veiller à ce que tous les aspects de l'application soient reflétés dans le corpus : les différents registres (ou genres de texte) utilisés dans le domaine, la diachronie, les différents modes de production, tout l'éventail des acteurs concernés.

L'idée de *clôture* présuppose pour [Condamines, 2003, pp. 70-71] que le corpus fonctionne comme un « système autonome » et laisse penser qu'on peut s'appuyer « sur le seul corpus sans faire appel aux connaissances extérieures ». Je ne reprends pas à mon compte cette idée de clôture. Dans les faits, on ne dispose jamais de corpus clos. Cela implique qu'on ne peut pas se passer de connaissances extérieures et qu'il faut au contraire penser l'interaction entre corpus et connaissances extérieures, ce que j'ai tenté de faire dans mon travail.

3.2.2 Le renouveau de l'analyse terminologique

Il était naturel, cherchant à construire des bases de connaissances et notamment des lexiques sémantiques pour des applications particulières, donc dans des domaines spécifiques, que je m'intéresse aux propositions de la terminologie. En lien avec l'essor des analyses de corpus, les années 1990 voient un renouvellement de l'analyse terminologique.

Le rapprochement entre les communautés de la terminologie et de l'Intelligence Artificielle a amené à considérer la terminologie sur des bases nouvelles [Meyer et al., 1992]. En France, la volonté de donner corps à ce rapprochement entre la réflexion terminologique et l'Intelligence Artificielle a conduit à la création du groupe TIA (voir 3.1). La réflexion de ce groupe a tout d'abord permis de relativiser la vision traditionnelle de la terminologie, telle qu'elle est représentée par les travaux d'E. Wüster et les thèses du Cercle de Vienne (travaux cités d'après [Pearson, 1998, Bourigault et Slodzian, 2000]). Ces derniers défendent l'idée d'un savoir scientifique stable dont la terminologie permet de fixer le vocabulaire. Le terme est vu comme le représentant

unique et univoque d'une notion qui existe *a priori*, le travail de l'expert consistant simplement à dénommer les notions du domaine qu'il considère.

Cette approche fixiste de la terminologie a progressivement laissé la place à une « terminologie textuelle » qui repose sur des bases radicalement différentes : le terme n'est pas donné, il doit être construit. Ce n'est pas une simple étiquette mais une unité textuelle dont il faut prendre en compte la variation en corpus. Il n'existe pas une terminologie unique représentant l'ensemble des connaissances d'un domaine mais différentes terminologies à construire pour différentes applications.

En marge de cette réflexion théorique sur le statut de la terminologie, les travaux de TIA ont permis de faire des propositions concrètes pour outiller le travail d'exploration de corpus et de construction de terminologies ou d'ontologies [Bourigault et Jacquemin, 2000, Aussenac-Gilles et al., 2004].

3.2.3 Le renouveau des ontologies

Un autre phénomène important des années 1990 a été l'émergence du terme « ontologie » dans le domaine de l'Intelligence Artificielle, un nombre important de travaux portant sur cette question, avec en ligne de mire la possibilité de construire un web sémantique.

Même si l'emploi même du terme « ontologie » s'est beaucoup généralisé dans les dernières années, l'idée consistant à expliciter la manière dont on choisit de conceptualiser un domaine[7] ne l'est pas. Elaborer la couche terminologique des logiques de description ou le support d'un modèle à base de graphes conceptuels revient déjà à construire l'ontologie des domaines qu'on modélise.

L'intérêt pour les ontologies s'organise autour de différentes directions de travail portant sur :

- Divers projets de construction d'ontologie, depuis les projets très ambitieux d'ontologie générale comme CYC [Lenat et Guha, 1990], jusqu'aux ontologies régionales [Benjamins et Fensel, 1998] ;
- Une réflexion autour des propriétés formelles des ontologies [Brachman, 1979, Guarino, 1995], ce qui soulève la question de l'écart entre terminologies, thesaurus et bases de connaissances lexicales et les ontologies [Gangemi et al., 2001] ;
- Une réflexion autour de la place de ces ontologies dans le développement du web sémantique ;
- La définition de modèles formels permettant d'encoder des ontologies

[7] « An ontology is an explicit specification of a conceptualization » [Gruber, 1993].

(notamment OIL[8] et OWL[9]). L'objectif du web sémantique étant de se servir des ontologies pour annoter les documents et services, il faut se donner les moyens de faire des inférences à partir de ces annotations.

La question de l'élaboration des ontologies reste un point délicat. Une ontologie est liée à la tâche pour laquelle elle a été élaborée : on peut douter de l'utilisabilité des ontologies générales et privilégier des ontologies régionales mais réutiliser des ontologies préexistantes est délicat [Charlet et al., 1996]. D'où l'importance de définir des méthodes permettant sinon de construire automatiquement des ontologies, du moins d'en guider la construction en tirant profit de l'analyse de corpus textuels.

3.2.4 Points critiques

Ces trois nouveaux mouvements de recherche laissent cependant des questions en suspens.

Les résultats des outils terminologiques sont souvent assez bruités et ne peuvent pas être exploités directement (outils d'acquisition de terminologie, par exemple), ce qui pose la *question de la validation humaine*. La nécessité de valider les résultats limite l'utilisation de ce type de méthode. D'où l'importance des interfaces de validation dans les travaux présentés ici (voir section 4.1, p. 59).

Les premiers travaux en terminologie textuelle ont mis l'accent sur le terme et les méthodes d'extraction de termes. On s'est également intéressé aux relations mais en focalisant sur l'hyperonymie aux dépens des autres types de relations [Hearst, 1992]. En France, la *question de la mise en réseau* de ces termes a relativement peu retenu l'attention jusqu'à [Borillo, 1996]. Les mettre en relation facilite cependant leur interprétation et leur validation. D'où l'importance des méthodes de « structuration de terminologie »[10]. C'est ce qui a motivé mon intérêt pour l'acquisition de classes et de relations.

Comme souvent la pratique a devancé la théorie et on a exploité des corpus avant de s'interroger sur la nature d'une sémantique reposant sur les corpus. Il me semble important pourtant de s'interroger sur le *fondement sémantique des méthodes d'analyse* [Nazarenko, 2005].

L'évaluation des méthodes d'acquisition de connaissances, qu'elles soient terminologiques ou ontologiques, reste très en-deçà de ce que l'on peut faire pour d'autres tâches du TAL, je rejoins en cela la conclusion de

[8] http ://www.ontoknowledge.org/oil/

[9] http ://www.w3.org/TR/owl-features/

[10] T. Hamon et moi avons organisé une journée d'étude de l'ATALA sur ce thème en mars 2001 [Nazarenko et Hamon, 2002].

[Habert et Zweigenbaum, 2002]. La difficulté vient de ce que toutes les méthodes d'acquisition ne peuvent pas être évaluées de la même manière, l'évaluation devant elle aussi dépendre de l'application visée. Je reviens sur cette question de l'évaluation au fur et à mesure de ce mémoire.

3.3 Acquérir des classes sémantiques

Je me suis d'abord intéressée à l'acquisition de classes sémantiques. Celle-ci présente un intérêt à la fois terminologique et ontologique. Avec le développement des extracteurs de termes, se pose la question de la validation des listes de candidat-termes produites. Regrouper les candidats sémantiquement proches pour aider le terminologue à choisir les termes canoniques constitue une première aide à la validation et assure une meilleure cohérence dans le travail de validation lui-même. Les classes de mots ou de termes servent aussi au repérage conceptuel d'un domaine, les classes sémantiques étant alors considérées comme des noyaux conceptuels à partir desquels l'ontologie du domaine peut être élaborée.

Ce travail sur l'acquisition de classes sémantiques a été effectué en collaboration avec B. Habert dans les années 1994-96[11]. Nous avons travaillé sur le corpus Menelas. Nos premiers résultats ayant intéressé B. Bachimont qui avait contribué à l'élaboration de l'ontologie du projet Menelas, nous avons essayé d'estimer plus précisément l'intérêt de notre approche pour la modélisation ontologique. En 1996-98, la collaboration s'est ainsi poursuivie, dans le cadre du groupe de travail ESPOIR (avec P. Zweigenbaum et J. Bouaud) [12] et dans une perspective d'évaluation. En 1996, j'ai été recruté à l'Université Paris 13 et j'ai cessé de collaborer aussi étroitement avec B. Habert, qui a poursuivi de son côté [Habert et Fabre, 1999].

Cette section présente notre démarche d'acquisition, qui a été concrétisée par le développement de l'outil Zellig, analyse les résultats obtenus (cartographies lexicales plutôt que classes sémantiques à proprement parler) et montre comment nous avons cherché à évaluer ces résultats.

3.3.1 Démarche

Notre démarche d'acquisition repose, comme beaucoup d'autres travaux, sur l'hypothèse qu'une sémantique distributionnelle est possible, c'est-à-dire

[11] J'étais alors ATER à l'ENS de Fontenay/Saint-Cloud.

[12] http ://www.biomath.jussieu.fr/ESPOIR/. B. Bachimont, P. Zweigenbaum et J. Bouaud étaient à l'époque membres du DIAM – Service d'Informatique Médicale/DSI/AP-HP et avaient tous participé au projet Menelas.

qu'on peut repérer des propriétés sémantiques à partir des distributions des mots en corpus, la *distribution* d'un mot étant l'ensemble des contextes dans lesquels il figure, la notion de contexte pouvant être défini de différentes manières.

Comme nous l'avons montré dans [Habert et al., 1997, p. 178], les trois ordres d'affinité entre les mots que G. Grefenstette met en évidence définissent une méthode d'acquisition en trois étapes [Grefenstette, 1994a].

La première étape consiste à identifier des *relations de cooccurrence* entre les mots : dans notre cas, il s'agit de relations de dépendances syntaxiques, choix que nous avons justifié dans le cas de corpus de taille modeste comme celui de Menelas (85 000 mots) [Habert et Nazarenko, 1996].

Lors de la seconde étape, on calcule des similarités entre les mots sur la base des cooccurrences dans lesquelles ils entrent. Différentes mesures de distance ont été proposées dans la littérature depuis [Hindle, 1990]. Dans Zellig, nous avons considéré que deux mots sont similaires s'ils partagent un nombre minimal x de contextes, une mesure simple qui s'est avérée mieux adaptée à la tâche que la mesure de Hindle, fondée sur l'information mutuelle. Cette dernière privilégie les événements rares, ce qui est souhaitable quand on cherche à mettre en évidence des compositions lexicales incongrues mais qui l'est moins quand on veut mettre au jour les relations conceptuelles « normales » qui structurent un domaine de connaissance.

La troisième étape est la plus problématique. Au-delà des axes sémantiques proposés par [Grefenstette, 1994b, p. 126], ce dont on a besoin, c'est d'une *relation d'équivalence* pour obtenir de véritables classes sémantiques. A partir d'une distance de similarité, les techniques de classification permettent de construire des classes [Bensch et Savitch, 1995, Mahon et Smith, 1996, Dagan et al., 1999], mais les classes obtenues ne peuvent pas être utilisées telle que telles : elles doivent être retouchées et étiquetées. Le travail d'interprétation des résultats est indispendable. Il peut prendre différentes formes : noyaux sémantiques introduits au début de l'analyse distributionnelle [Riloff et Shepherd, 1997], coopération au fur et à mesure du processus de classification [Faure et Nédellec, 1999]. Convaincus qu'il n'est pas possible d'obtenir automatiquement des classes bien formées, nous avons pris le parti de dissocier dans ZELLIG traitement automatique et interprétation, et de construire des cartographies des similarités du corpus pour épauler au mieux le travail d'interprétation.

3.3.2 Explorer des cartographies lexicales

Les résultats de ZELLIG se présentent donc sous la forme de graphes de similarité qui donnent à l'utilisateur une vue globale du fonctionnement lexical

du corpus analysé.

La liste des couples de mots considérés comme similaires pour une valeur donnée de x étant impossible à exploiter manuellement, il fallait des outils d'exploration. Nous avons proposé de représenter l'ensemble de ces relations de similarité sous la forme de graphes.

Un graphe de similarité est un graphe non orienté et étiqueté [Habert et Nazarenko, 1996]. Les noeuds sont formés par les unités lexicales considérées. Deux noeuds sont reliés par un arc si et seulement si les deux unités lexicales correspondantes partagent plus de x contextes, *i.e.* si elles ont plus de x relations de cooccurrence en commun. Les arcs sont étiquetés avec la liste de ces contextes partagés. La figure 3.1 montre le sous-graphe des adjectifs obtenu à partir du corpus Menelas. Ce graphe a été construit de manière entièrement automatique, seule sa disposition a été faite manuellement.

En quoi ces graphes peuvent-ils aider le travail de classification ? Que donnent-ils à voir ? Ils présentent une vue différentielle du vocabulaire du corpus en montrant quelles unités sont les plus proches. Le graphe de la figure 3.1 montre des rapprochements intéressants (voir [Habert et al., 1996, Zweigenbaum et al., 1997] pour une analyse détaillée) mais toute la question est de savoir comment exploiter cette vue.

L'expérience acquise sur le corpus Menelas a permis de jeter les bases d'une telle méthode d'analyse des résultats. Nous avons mis l'accent sur les sous-graphes caractéristiques du graphe global : les composantes connexes et les cliques. Contrairement à notre attente, les cliques ne constituent que rarement des embryons de classes. Il convient au contraire de s'appuyer sur les composantes connexes. On observe en effet que les cliques sont souvent formés de mots fréquents et polysémiques qui se trouvent donc en relation avec beaucoup d'autres mots : elles reflètent des glissements de sens.

3.3.3 Evaluer l'acquisition

Les observations que nous avons pu faire sur les graphes de Menelas, aussi prometteuses soient-elles, n'ayant aucune valeur de preuve, il fallait se donner les moyens de mesurer l'intérêt des graphes de similarité dans une perspective d'acquisition de connaissances. Les graphes devant servir à guider l'interprétation, il faut mesurer l'aide apportée à l'utilisateur en charge du travail de modélisation : est-ce plus facile, plus économique, plus fiable de construire une ontologie à partir des graphes de similarités que sans ? Le groupe ESPOIR a cherché à répondre à la question de cohérence de ces graphes par rapport à des références existantes dans le domaine.

Comme il n'existait pas d'outils similaires à ZELLIG, nous avons cherché à

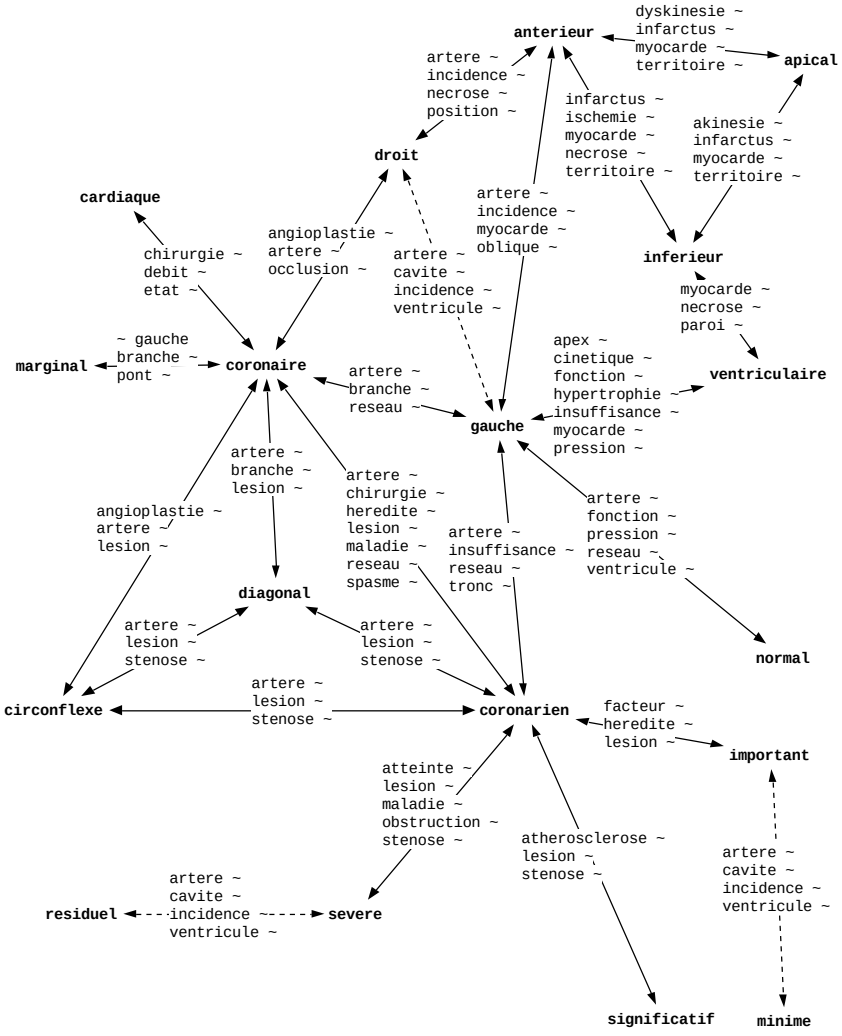

FIG. 3.1 – Exemple de graphe des similarités

confronter les graphes résultats avec des ontologies de référence et à apprécier l'utilité des graphes pour une tâche d'étiquetage sémantique.

Comparaison avec des ontologies existantes

Nous avons comparé les graphes de ZELLIG avec deux modélisations existantes, assez différentes l'une de l'autre : une catégorisation « à gros grain », la SNOMED Internationale [Côté et al., 1993][13] et l'ontologie « fine » créé pour le projet Menelas [Bouaud et al., 1995]. L'expérience est rapportée dans [Bouaud et al., 2000]. Ces comparaisons confirment l'impression intuitive de « choses intéressantes » mais il est difficile d'en tirer directement des conclusions. Indépendamment de l'évaluation en tant que telle, la comparaison de l'ontologie de Menelas et des graphes de ZELLIG met en évidence la différence entre modélisation conceptuelle et modélisation linguistique ainsi que les limites de cette dernière. Avec les phénomènes comme la métonymie ou l'élision, la langue donne une vue déformée du domaine qu'elle est censée refléter[14].

Utilisation des graphes pour l'étiquetage sémantique

La comparaison précise des graphes avec les grandes catégories SNOMED étant impossible à faire directement du fait de la différence de grain dans les descriptions, nous les avons comparés indirectement à travers une tâche d'étiquetage. Pour mesurer la cohérence des graphes de Zellig par rapport aux onze grandes catégories sémantiques de SNOMED, nous avons regardé si les graphes permettent d'étiqueter correctement des mots inconnus, par propagation des étiquettes le long des arcs du graphe, le mot inconnu recevant l'étiquette sémantique de la majorité de ses voisins dans le graphe.

Ces expériences sont décrites dans [Nazarenko et al., 2001]. Si on compare notre méthode d'étiquetage aux résultats qu'on obtiendrait en attribuant l'étiquette la plus fréquente, on constate qu'ils sont meilleurs, ce qui prouve la cohérence sémantique des relations de similarité[15].

[13] Il s'agit d'une nomenclature qui décrit les concepts médicaux et vétérinaire selon différents axes, chaque axe étant lui-même hiérarchisé. Au total, les termes se répartissent entre 11 grandes catégories [Zweigenbaum, 1999].

[14] Ainsi, dans le corpus Menelas, on parle de « sténose de l'artère X », sans expliciter le fait qu'une sténose est en réalité localisée sur un segment d'artére, ce que fait l'ontologie de Menelas. On voit ainsi se dessiner les limites de l'acquisition de connaissances à partir de corpus.

[15] Si l'objectif était réellement de procéder à un étiquetage sémantique, ces résultats seraient insuffisants.

Au-delà de la confrontation des graphes de ZELLIG avec la nomenclature SNOMED, cette tâche d'étiquetage a donné un cadre pour tester différents paramètres de Zellig de manière systématique, notamment concernant la mesure de similarité et le seuil x utilisés. [Nazarenko et al., 2001].

3.3.4 Bilan

Les expériences rapportées ici présentent un certain nombre de limitations liées aux conditions de travail : seuls les mots isolés sont considérés alors qu'il aurait fallu prendre en compte des termes polylexicaux ; les contextes verbaux ne sont pas exploités, alors qu'ils semblent pertinents dans d'autres expériences de classification [Faure et Nédellec, 1999] ; nous avons travaillé sur trop peu de corpus.

Avec ces réserves en tête, on peut néanmoins tirer quelques conclusions de ce travail :

- La méthode d'analyse distributionnelle reste encore mal maîtrisée. En dépit du nombre des travaux effectués, on manque de recul sur les distances à utiliser et leurs domaines d'application. Même s'il est clair qu'on ne peut pas construire des classes de manière entièrement automatique, le projet fait encore illusion. Le fait qu'une part d'interprétation soit nécessaire plaide en faveur de méthodes d'apprentissage supervisé [Riloff et Shepherd, 1997].
- Une part d'interprétation étant incontournable, il faut la prendre en compte dans le processus d'analyse. Cela suppose généralement de concevoir des interfaces adaptées à la validation et à l'exploitation des résultats adaptées. Dans ZELLIG, nous ne sommes pas allés au bout de la démarche harrissienne qui aurait nécessité de mettre au point une interface de modélisation, permettant de naviguer entre les deux graphes duaux des opérateurs et des arguments (dans notre cas des noms et des adjectifs) et de propager dynamiquement les regroupements effectués dans l'ensemble du graphe.
- Cela soulève la question centrale de l'exploitation des résultats de l'analyse distributionnelle. D'un côté on observe que, même bruités, les résulats présentent de l'intérêt, ce que confirment les éléments d'évaluation. D'un autre côté, on se sait pas bien ce qui est utilisable et comment. L'analyse distributionnelle n'est certainement pas la panacée en matière d'acquisition, mais, bien contrôlée et sur des points précis, elle reste prometteuse. Il devrait ainsi être possible d'intégrer un module d'analyse distributionnelle dans un outil de modélisation comme TERMINAE en complément des autres méthodes d'acquisition [Szulman et al., 2002] . De nouvelles expériences en ce sens sont en

cours (voir [Bourigault, 2002] et le projet Extraplodocs, section 4.2). Espérons qu'elles feront progresser la réflexion sur cette question

3.4 Acquérir des relations lexicales : la synonymie

L'une des limites de l'approche distributionnelle pour construire un réseau de termes vient de qu'elle ne permet pas de typer la relation que deux termes repérés comme proches entretiennent entre eux.

Par comparaison avec la construction de classes sémantiques, le repérage et le typage de relation ont reçu une moindre attention. Leur intérêt terminologique est pourtant clair : les relations entre termes attestés peuvent être directement intégrées dans un réseau de termes ; les relations entre candidats termes structurent leur espace sémantique et guident leur validation. Elles ont aussi un intérêt ontologique. Dans le cadre du travail d'acquisition, qui vise à inférer une structure ontologique à partir de l'organisation lexicale décelable dans les textes analysés, on cherche à interpréter les relations lexicales entre termes comme des relations entre les concepts associés à ces derniers. Même s'il n'y pas bijection entre les deux niveaux terminologique et ontologique, le premier constitue une ébauche du deuxième.

L'hyperonymie est la relation la plus étudiée dans cette perspective, à partir des travaux de [Hearst, 1992, Yarowsky, 1992] notamment, et ceux de [Borillo, 1996] puis [Morin, 1999], en France. C'est la relation d'héritage qui structure traditionnellement les thesaurus et les ontologies. L'acquisition des relations de méreonymie [Berland et Charniak, 1999, Séguéla, 1999], d'antonymie [Hatzivassiloglou et McKeown, 1993], de synonymie ou des relations plus spécialisées [Rousselot et al., 1996] a été étudiée de manière plus marginale.

Je présente ci-après un travail visant à faciliter l'acquisition des relations de synonymie. Il s'organise en deux temps :

- La première étape est liée au travail de DEA et de thèse (1997-2000) de Thierry Hamon[16]. Les expériences ont porté sur un corpus technique d'EDF et sur le corpus Menelas.
- Le deuxième travail porte sur le domaine de la biologie. Il est centré sur les entités nommées plus que sur les termes à proprement parler. Il a été effectué dans le cadre du DEA de Davy Weissenbacher et du

[16]Le début de ce travail a été financé par un contrat de collaboration bipartite entre le LIPN et la Direction des Etudes et Recherche de EDF (1997-98, 1 an). La thèse a ensuite été financée par une bourse MENRT.

projet ExtraPloDocs (voir section 4.2, p. 66).

Cette section présente ces deux directions de travail, après avoir avoir posé la problématique générale de l'acquisition de relations de synonymie.

3.4.1 Problématique

Les objectifs de ce travail sur l'acquisition de synonymie sont divers :

- La structuration de terminologie est vue comme une première étape utile pour la construction d'ontologie. Déceler que deux termes sont des synonymes facilite le travail de normalisation consistant à choisir un représentant canonique parmi un ensemble de candidats plus ou moins équivalents.
- Il m'intéressait aussi d'aborder dans ce cadre la question de la contribution respective des connaissances extérieures (qui sont souvent des connaissances « générales » au regard de l'application visée) et des données du corpus. J'ai déjà mentionné le fait que cette question me paraît essentielle pour fonder une sémantique de corpus. En 1997, au sein du groupe TIA, cette question avait été négligée, les connaissances générales étant considérées comme peu pertinentes pour la construction de terminologie particulière[17] et l'idée de l'adaptation lexicale n'ayant pas encore réellement percé [Basili et al., 1997, Basili et al., 1998].
- Ces travaux ont également été motivés par des besoins applicatifs : faciliter la navigation dans les documents (voir section 4.1, p. 59) et normaliser les documents dans une perspective d'extraction d'information (voir section 4.2, p. 66).

La synonymie

La question de la définition de la synonymie est classique en sémantique lexicale. On sait qu'il s'agit d'une notion graduelle continue plutôt qu'absolue [Cruse, 1986, p. 265] [Ploux et Victorri, 1998]. Je considère pour ma part une notion de *synonymie contextuelle* : deux mots sont synonymes dans un contexte s'ils peuvent être substitués l'un à l'autre dans le contexte considéré, avec préservation du sens global de l'énoncé. Cela revient à définir un contexte dans lequel la dimension selon laquelle ils s'opposent (diachronie, registre de langue, etc.) est considérée comme non pertinente. Cette définition soulève néanmoins des questions (qu'est-ce qu'un contexte ? que signifie « préserver le sens global de l'énoncé ? ») auxquelles on ne peut répondre qu'en lien avec un contexte énonciatif ou, pour les tâches visées ici, une application.

[17] Je fais notamment référence ici à des conversations personnelles avec D. Bourigault.

La terminologie traditionnelle fait l'hypothèse que dans un domaine stable où la terminologie est fixée, les phénomènes de synonymie ont été éradiqués. Lorsqu'il en demeure, le contrôle terminologique invite alors à préférer un terme t à d'autres termes plus ou moins synonymes t_1, t_2... t_n dont l'usage est découragé. En pratique, les cas de synonymie sont fréquents. La terminologie textuelle l'a montrée, la conceptualisation des domaines étudiés reste souvent partielle et la variation terminologique est importante. Ces cas de synonymie, et, au sens large, d'équivalence sémantique, sont de différents types : variation typographique et morpho-syntaxique, abréviation ou acronymie, renommage, différence de registre de langue, etc. On comprend aisément que ces phénomènes de synonymie entravent l'accès au contenu des documents.

Méthodes de structuration de terminologie

La structuration permet de transformer une liste de termes en un réseau où les termes sont reliés entre eux par des liens syntaxiques et sémantiques [Nazarenko et Hamon, 2002]. Différentes méthodes ont été proposées : les *approches structurelles* reposent sur l'analyse de la structure interne des termes[18] ; les *approches contextuelles* s'appuient sur les contextes d'emploi pour rapprocher deux termes. Parmi ces dernières, on peut à nouveau distinguer les approches à base de règles (ou de patrons)[19] et les méthodes distributionnelles[20].

Ces sont les deux approches structurelle et par règles qui sont explorées ici, dans le cas de la synonymie.

3.4.2 Approche structurelle : SynoTerm

La première étape de ce travail sur l'acquisition de relations de synonymie a permis la réalisation du prototype SynoTerm [Hamon, 2000]. Je présente ici rapidement la méthode et les résultats qu'elle permet d'obtenir.

Méthode

Il s'agit de structurer une liste de termes ou de candidats-termes complexes[21] en regroupant les termes synonymes. La méthode proposée repose

[18] Voir par exemple les travaux suivants : [Jacquemin, 1997, Assadi, 1997, Cerbah, 2000, Hamon, 2000].

[19] Voir par exemple les travaux suivants : [Hearst, 1992, Kavanagh, 1995, Berland et Charniak, 1999, Morin et Jacquemin, 1999, Séguéla, 1999].

[20] Avec cette réserve cependant qu'au sein des regroupements de termes obtenus, les relations ne sont pas clairement identifiées.

[21] C'est-à-dire composés de plusieurs unités lexicales.

sur une liste de couples de synonymes préexistante (issue d'un dictionnaire, par exemple), et une liste de (candidats-)termes attestés en corpus. Cette liste est considérée comme un réseau dont les noeuds sont les termes et leurs constituants, et les arcs les liens de composition. La méthode repose sur la propagation de la synonymie dans ce réseau de termes : les relations initiales sont propagées le long des liens de composition des termes.

Un lien initial, issu de la ressource extérieure (par exemple, le lien entre *essor* et *développement*), n'est pas intégré tel quel dans le réseau des termes car il n'est pas valide dans tous les contextes (par exemple, dans *développement informatique*, *développement* n'a pas toujours le même sens qu'*essor*). Ce lien n'est conservé que si les deux termes entrent dans des constructions identiques, l'existence de ces constructions parallèles étant un indice de la pertinence du lien initial pour le corpus considéré. On ne propose d'interpréter *développement informatique* comme synonyme de *essor informatique* que pour les corpus où les deux termes sont attestés. On a ici le premier exemple d'application d'un principe qu'il me paraît important de respecter pour l'exploitation des connaissances extérieures : celles-ci ne doivent être exploitées que corroborées en corpus, leur utilisation aveugle engendrant beaucoup de bruit dans les résultats.

Il faut également souligner que les relations inférées ne sont pas directement intégrées dans le réseau terminologique, même corroborées en corpus, elle doivent être validées par un terminologue. Le prototype SynoTerm intègre une interface de validation qui permet de valider, rejeter ou retyper chacune des relations proposées. On voit ici le deuxième principe que je tente de respecter : considérer le processus d'acquisition dans son ensemble et outiller au mieux le travail manuel, à défaut de pouvoir l'automatiser.

Résultats

Dans le cadre de sa thèse, T. Hamon a fait deux séries d'expériences sur deux corpus de taille et de domaine différents[22] en exploitant les connaissances sémantiques issues du dictionnaire Le Robert fourni par l'INaLF où elles figurent sous la forme de liens *voir-aussi*.

L'analyse des résultats montre que SynoTerm met en évidence des liens de synonymie (*bon calibre* et *belle qualité* sont synonymes quand il s'agit d'artère, par exemple) mais aussi des propositions douteuses. Ces résultats sont présentés avec précision dans [Hamon et Nazarenko, 2001]. Il ressort que l'apport du dictionnaire est modéré mais réel :

[22]Un corpus d'EDF de 200 000 mots et le corpus Menelas [Zweigenbaum et Consortium Menelas, 1994] qui en comporte 85 000.

- Peu de relations de synonymie sont proposées au regard de la taille du corpus et du nombre de candidats-termes ;
- Un tiers seulement des liens sont conservés lors de la validation ;
- En revanche, la méthode proposée permet de détecter des relations inédites, difficiles à repérer par une autre méthode ou à introduire spontanément pour un terminologue travaillant sans assistance[23].

Pour rendre un outil comme SynoTerm exploitable, il faut minimiser le travail de validation. Nous avons montré qu'une structuration appropriée des relations facilite cette validation [Hamon et Nazarenko, 2001] : le bruit dans les résultats provient de la projection d'un lien de synonymie qui s'avère non pertinent pour le corpus considéré. En regroupant les résultats autour du lien initial qui a permis de les inférer, on peut éliminer d'un seul clic toutes ces erreurs. Sans augmenter la précision des résultats, on peut ainsi réduire le coût d'élimination du bruit. Ceci constitue à mon avis l'une des originalité de l'interface de validation de SynoTerm.

Le coût de validation doit par ailleurs être relativisé du fait que SynoTerm n'a pas vocation à rester un outil isolé. En général, c'est un produit terminologique global qui doit être validé : l'ensemble du réseau des candidats-termes et de leurs relations dédié à une application. De même que la structuration facilite la validation des candidats-termes, les relations sémantiques gagnent à être validées en contexte. Ce point sera illustré dans la section 4.1.

3.4.3 Approche par règles : les noms de gènes renommés

La seconde expérience portant sur la synonymie est complémentaire de la première : elle repose sur une méthode contextuelle et non plus structurelle. Elle est de moindre envergure (elle correspond au stage de DEA de D. Weissenbacher) et a été menée dans un contexte différent :

- elle s'intègre dans un projet global d'extraction d'information dans les textes de biologie rédigés en anglais ;
- elle concerne un type particulier de synonymie : le renommage des noms de gènes ou de protéines.

Une remarque s'impose à propos du statut de ces derniers. Comme l'ensemble des noms propres, les noms de gènes et de protéines sont d'ordinaire considérés comme des *entités nommées* plus que comme des termes, mais cela concerne surtout leur rôle dans l'analyse des corpus. Même si cela se justifie du point de vue ontologique[24], la distinction me paraît peu pertinente, du

[23] C'est l'avis émis par H. Boccon-Gibod, le terminologue d'EDF qui a été associé à cette étude.

[24] D'un point de vue ontologique, comme il est nécessaire de distinguer le concept de ses instances, les entités biologiques figurent donc dans l'ontologie mais avec le statut

point de vue lexical. Les noms d'entités biologiques font l'objet d'un questionnement terminologique à proprement parler : phénomènes de renommage, règles de création des nouvelles dénominations, recensement des noms et de leurs variantes dans des nomenclatures consultées par les biologistes.

Une attention particulière est portée au recensement des noms de gènes et de protéines en biologie dans les bases de connaissances[25]. De nombreux travaux ont porté sur la reconnaissances de ces entités nommées (voir entre autres [Fukuda et al., 1998, Proux et al., 1998]). Ceux-ci se heurtent cependant aux problèmes de la variation des noms de gènes ou de protéines qui est à la fois massive (environ 40% des noms de gènes référencés dans Flybase admettent un synonyme, souvent plusieurs), hétérogène (variation typographique, renommage, abréviation) et difficilement prédictible[26]. Se donner les moyens d'identifier les relations de synonymie entre gènes est donc important. Les cas de variation typographique et d'abréviation [Chang et al., 2002] ont été les premiers étudiés du fait de leur importance numérique mais aussi parce qu'ils peuvent être captés sur la seule base d'une distance entre chaînes. Les autres cas de synonymie, les renommages notamment, sont plus complexes à repérer et une approche contextuelle est nécessaire. Je mets l'accent sur ces derniers.

L'objectif de ce travail est de repérer en corpus des relations de synonymie entre noms de gènes, en supposant les noms de gènes préalablement identifiés.

Démarche

La méthode proposée est classique. Elle s'inspire des pratiques terminologiques traditionnelles [Pearson, 1998, pp. 174-189] et repose sur le repérage de marqueurs du renommage comme *also known as* ou *termed as* : il s'agit d'exploiter les connaissances lexicales explicitées dans les corpus pour enrichir une base lexicale existante. Cela suppose de connaître les marqueurs exploitables et leur degré de fiabilité. Pour des textes de biologie, [Hishiki et al., 1998] propose *termed*, *designated as* et les parenthèses, mais ces dernières sont en fait très polysémiques[27]. Ces marqueurs étant par ailleurs susceptibles de varier d'un corpus à l'autre [Condamines, 2003, pp. 149-50]), il faut les acquérir en corpus.

---

particulier d'instances de concept (problématique du *peuplement de l'ontologie* (*ontology population*). Un nom de gène particulier (*spoIIID*) est ainsi représenté comme une instance du concept générique de GÈNE ou de GÈNE DE L'ESPÈCE BACILLUS SUBTILIS.

[25]Comme Flybase par exemple (flybase.bio.indiana.edu).

[26]Par exemple, le gène *Hcph* peut être mentionné sous les noms suivants : *PTP1C, SHP, SH-PTP1, PTPN6.*

[27]Elles peuvent aussi introduire un nom d'espèce, une référence bibliographique, etc.

La démarche d'acquisition proposée va dans ce sens. Dans le cadre du DEA de D. Weissenbacher, elle a été testée manuellement. Selon un schéma classique [Morin et Martienne, 1999], elle comporte elle-même une phase d'acquisition et une phase d'exploitation. Pour acquérir les règles d'extraction de relations de renommage, on projette sur le corpus quelques couples de noms de gènes connus pour être synonymes ; on retient les fragments de phrases dans lesquelles figurent au moins deux synonymes et c'est l'analyse de ces fragments qui permet de définir des règles d'extraction (phase d'acquisition). Ces règles sont ensuite appliquées sur l'ensemble du corpus (phase d'exploitation) pour reconnaître de nouveaux couples de synonymes.

Résultat et évaluation

L'étude a porté sur des articles de la base bibliographique de Medline[28]. Elle a d'abord permis de mieux comprendre le fonctionnement de la synonymie des noms de gènes. Elle est fortement *redondante* (plusieurs articles expriment de manières différentes la même synonymie), ce qui invite à privilégier la précision de la méthode d'acquisition par rapport à son rappel. Son repérage se prête bien aux méthodes à base de règles, car l'expression de la synonymie est *locale* et *relativement stable*. Ce repérage suppose néanmoins une analyse précise du fragment de phrase concerné, car il est à la fois *complexe* (elle repose sur différents types d'indices : marqueurs classiques de la synonymie mais aussi ponctuations, conjonctions, acronymes, etc.) et *sensible* aux variations (la présence ou l'absence d'un indice en apparence mineur comme une ponctuation peut modifier l'interprétation du fragment).

Plutôt que de modéliser les tournures synonymiques comme des séquences d'indices comme c'est souvent le cas dans les approches à base d'apprentissage [Yu et Agichtein, 2003], D. Weissenbacher a proposé de les modéliser sous la forme de structures arborescentes, qui permettent de rendre compte de l'ordre des indices, de l'optionalité de certains éléments, de la composition de structures complexes à partir de structures élémentaires et de la hiérarchie des indices. Il met ainsi au jour 3 structures différentes de l'expression de la synonymie. [Weissenbacher, 2004] montre que ce qui « marque » la synonymie, ce n'est pas un marqueur (mot ou expression) considéré isolément mais ce marqueur pris dans un jeu de contraintes contextuelles.

Par comparaison avec les travaux similaires de Yu [Yu et al., 2002], la mise au point des règles est plus complexe mais la précision des résultats obtenus semble nettement supérieure : sur un petit corpus de test, la précision est de 97%. Ce résultat demande cependant à être confirmé :

[28] www.ncbi.nlm.nih.gov

- D'un point de vue technologique, il faut mettre cette méthode en oeuvre sur un plus grand corpus.
- Du point de vue de l'application, et donc sur le plan biologique, la perspective est tout autre. Il faut déterminer non pas la qualité intrinsèque de ces relations mais leur « utilité » : dans le cas présent, cela suppose de mesurer le nombre de relations inconnues découvertes (relations non recensées dans les bases de connaissances existantes).

3.4.4 Bilan partiel

En dépit des différences entre les deux approches, la confrontation des deux expériences est intéressante. Elle pose clairement le problème de la structuration de terminologie : il n'existe pas une méthode adaptée à un type de relation, mais plusieurs méthodes qui contribuent, chacune, à la reconnaissance de différentes relations. Ceci ouvre différentes pistes de recherche complémentaires.

La méthode de propagation sémantique de SynoTerm a fait la preuve de son intérêt dans le cas de la synonymie, mais qu'en est-il pour d'autres types de relations ? Un mécanisme de propagation similaire pourrait être exploité dans le cas de l'hyperonymie, voire de la méréonymie. Les règles et les contraintes de propagation seraient évidemment à redéfinir. Ce point n'a pas encore été approfondi.

La question de l'intégration des différentes approches pour l'acquisition d'une relation se pose. Différentes méthodes (distributionnelle, structurelle ou à base de règles) peuvent concourir à identifier des liens de synonymie mais elles ont été étudiées séparément et les modalités de leur coopération restent à définir : suffit-il de les appliquer en parallèle et de fusionner les résultats ? A première vue, il semble que les connaissances préexistantes et la méthode à base de règles devraient plutôt servir à construire un début de réseau que la propagation pourrait alors saturer. Les distances distributionnelles serviraient finalement à conforter ou disqualifier les relations obtenues et à préparer le travail de validation. Ces propositions n'ont pas été mises en oeuvre et ce schéma a besoin d'être éprouvé.

En pratique, toutes les méthodes d'acquisition n'ont pas forcément la même pertinence pour tous les corpus et toutes les applications. Il est même probable que non : différents auteurs[29] ont mis en évidence une variation entre les genres de textes qui a certainement des conséquences sur l'efficacité des méthodes d'acquisition. Il faudrait donc caractériser le domaine d'application

[29] Voir la synthèse des travaux de l'Action Spécifique STIC, Corpus et Terminologie (ASSTICCOT www.irit.fr/ASSTICCOT/).

de chaque méthode : cela semble difficile à réaliser à ce jour, mais l'enjeu me paraît important.

Ces trois questions relèvent toutes de la problématique globale de l'intégration des différentes méthodes de structuration de terminologie. Le travail de [Jacquemin, 2003] a montré que les règles de propagation de SynoTerm peuvent être facilement intégrées à FASTR : c'est une première étape. D'autres expériences sont en cours : je reviendrais sur celle d'IndDoc en section 4.1 ; S. Szulman a commencé récemment à intégrer SynoTerm dans Terminae [Szulman et al., 2002], ce qui devrait permettre de travailler sur la propagation des relations dans le réseau terminologique ; des questions similaires se posent à D. Bourigault pour Syntex[30].

3.5 Acquérir des règles d'extraction ?

Après les classes sémantiques et les relations terminologiques, les règles d'extraction constituent un troisième type de connaissances à acquérir. Par *règle d'extraction*, je désigne une règle qui explicite l'interprétation d'un fragment de texte en précisant l'ensemble des contraintes contextuelles de cette interprétation. [Hearst, 1992] pose par exemple que les groupes nominaux se trouvant dans la configuration de phase suivante :

$$\textit{Such}\ \mathrm{NP}_0\ \textit{as}\ \mathrm{NP}_1,^*\ \textit{or}|\textit{and}\ \mathrm{NP}_n$$

entretiennent des liens d'hyperonymie, les groupe nominaux correspondant aux NP_i étant les hyponymes de celui qui instancie NP_0. De la même manière, [Weissenbacher, 2004] propose d'interpréter comme synonymes deux noms de gènes qui entrent dans le schéma de phrase suivant :

$$\mathrm{GeneName}_1\ \textit{(formerly known as}\ \mathrm{GeneName}_2\textit{)}$$

Même si ces contraintes sont souvent exprimées sous la forme de patrons ou schémas de phrase comme ci-dessus, je préfère parler de « règles » plutôt que de « patrons d'extraction » (*extraction pattern*), les contraintes pouvant être moins locales.

Ces règles sont à la base des systèmes d'extraction d'information. L'écriture manuelle étant coûteuse et difficile à maîtriser, on a cherché les moyens d'automatiser ou d'assister ce travail d'acquisition, pour rendre les systèmes d'extraction des laboratoires opérationnels [Grishman, 1997] [Poibeau et Nazarenko, 1999]. C'était l'objectif de la thèse de T. Poibeau (nov. 1998-mars 2002) réalisée en contexte industriel. Les règles d'extraction

[30] Conversation personnelle, juillet 2003.

servent aussi à acquérir des connaissances de nature terminologique (voir section précédente) ou ontologique [Nédellec et Nazarenko, 2004]. Dans ce cas, c'est le manque de stabilité des règles d'un corpus et d'une application à l'autre qui oblige à automatiser l'écriture des règles. Les règles les plus génériques n'étant pas productives sur tous les corpus, il faut en écrire de plus spécifiques.

Cette section justifie le recours aux techniques d'apprentissage et décrit rapidement les travaux menés dans cette direction. Cette recherche se poursuit aujourd'hui dans le cadre des projets ExtraPloDocs et Alvis qui sont présentés dans le chapitre suivant.

3.5.1 Assister l'écriture de règles

Dans le cadre de son travail de thèse sur l'extraction d'information, Thierry Poibeau a conçu un module d'aide à l'écriture de règles. L'utilisateur de ce module cherche à acquérir des règles d'extraction. Il doit d'abord sélectionner quelques exemples de phrases ou de structures prédicatives illustrant ce qu'il veut extraire. Le système exploite alors un réseau sémantique et un corpus représentatif pour calculer des paraphrases syntaxico-lexicales qui sont finalement proposées à l'utilisateur pour validation [Poibeau, 2003, ch. 8].

On retrouve ici deux aspects importants de la problématique générale de l'acquisition de ressources :

- la coopération : l'utilisateur qui met au point le système intervient non pas pour une validation *a posteriori* mais *a priori* pour donner des exemples qui servent à amorcer l'acquisition et pour contrôler en cours de route le processus de généralisation ;
- l'exploitation de ressources générales dont l'exploitation doit être contrôlée par un corpus d'acquistion.

Cette approche repose sur l'hypothèse que l'utilisateur a la compétence nécessaire pour mettre au point des règles. Une approche alternative est possible : apprendre automatiquement les règles à partir de corpus.

3.5.2 Le recours aux techniques l'apprentissage

La question de l'apprentissage des règles d'extraction a fait l'objet de nombreux travaux dans la communauté s'intéressant à l'extraction dans les années 1990[31]. En France, en revanche, l'exploitation des techniques d'apprentissage pour le TAL était comparativement peu développée, les expé-

[31] A partir des travaux de E. Riloff et S. Soderland notamment [Riloff, 1993, Freitag, 1998, Riloff et Schmelzenbach, 1998, Riloff et Jones, 1999, Soderland, 1999], voir [Nédellec, 2000] pour une synthèse.

riences étant relativement isolées. C'est la raison pour laquelle Claire Nédellec et moi avons lancé un groupe de travail sur ce thème. Nos travaux parallèles sur l'acquisition de classes sémantiques nous avaient rapprochées et avaient mis en évidence les bénéfices de collaborations pluridisciplinaires entre les deux communautés du TAL et de l'apprentissage automatique. Le groupe de travail A3CTE (Applications, Acquisition et Apprentissage de Connaissances à partir de Textes Electroniques) a été créé en septembre 1998[32] visant à réunir les communautés du TAL, de l'apprentissage et de l'ingénierie des connaissances pour travailler sur des applications concrètes. Ce groupe a été actif pendant 4 ans[33]. Il a permis de recenser les besoins et de développer une culture partagée entre les participants.

Concernant l'acquisition de connaissances, je cherchais pour ma part à évaluer l'apport des méthodes d'apprentissage et à définir les modalités de sa mise en oeuvre, pour l'acquisition de règles d'extraction, notamment. Cette réflexion a pris corps dans plusieurs projets portant sur l'acquisition de ressources pour l'extraction d'information (voir section 4.2, p. 66).

Dans ce qui suit, l'accent est mis sur les règles permettant d'extraire des relations, l'extraction de simples éléments ayant été très étudiée (pour la reconnaissance d'entités nommées, notamment).

3.5.3 Apprendre des règles à partir de corpus

Le travail décrit ici s'intègre dans le projet de bio-informatique Caderige (2000-2003). Ce projet exploratoire a permis de jeter les bases d'une méthode d'apprentissage de règles d'extraction pour les interactions entre gènes. Je décris dans cette section les principes de la méthode d'acquisition. Sa mise en oeuvre, qui se poursuit dans le cadre du projet RNTL ExtraPloDocs, est présentée dans le chapitre suivant (section 4.2).

Objectif

Le projet Caderige [Bessières et al., 2001, Alphonse et al., 2004] vise à concevoir une méthode d'extraction de connaissances portant sur les interactions entre gènes dans la littérature scientifique en biologie. Il s'agit de remplir un formulaire comme celui de la figure 3.2 à partir de l'analyse d'un

[32] Il a été rapidement reconnu par le GDR I3, l'AFIA et le Réseau régional de chercheurs en sciences cognitives d'Ile de France (Rescif).

[33] Il s'est réuni environ 4 fois par an pour des séances d'une journée entière combinant tutoriel et atelier de recherche. Différentes manifestations organisées par le groupe ont attesté de sa vitalité.

fragment de résumé d'article issu de la base bibliographique Medline[34].

FRAGMENT DE TEXTE
Previously, it was shown that the GerE protein inhibits transcription in vitro of the sigK gene encoding sigmaK.
FORMULAIRE

Interaction	*type*	negative		
	agent	GerE protein		
	target	expression	*source*	sigK gene
			product	sigmaK

FIG. 3.2 – Remplissage d'un formulaire sur les interactions entre gènes.

Choix de l'apprentissage relationnel

Les méthodes d'extraction automatique appliquées jusque-là en biologie reposent sur des comptages de co-occurrences de mots-clefs [Stapley et Benoit, 2000, Pillet, 2000] qui ne disent rien de la nature des relations sous-jacentes, ou sur des règles ou automates d'extraction définis manuellement [Blashke et al., 1999, Thomas et al., 2000]. Les résultats obtenus présentent soit une précision très faible soit une couverture limitée. L'extraction automatique de connaissances pertinentes dans les documents sélectionnés fait donc appel à des méthodes d'extraction d'information plus complexes qui s'appuient sur des règles comme celle de la figure 3.3) combinant des contraintes lexicales, syntaxiques et sémantiques (pour augmenter la précision) et qui puissent être acquises automatiquement ou semi-automatiquement à partir de corpus textuels (pour assurer une bonne couverture).

SI le sujet X d'un verbe Y d'interaction est un nom de protéine
ET SI l'objet direct Z est un nom de gène ou l'expression d'un gène
ALORS Il y a une interaction dont X est l'agent et Z est la cible.

FIG. 3.3 – Exemple de règle d'extraction, avec des contraintes sur les catégories syntaxiques (verbe, nom, etc.), sur des dépendances syntaxiques (sujet, objet direct) et sur les types sémantiques (verbe d'interaction, nom de gène).

[34] www.ncbi.nlm.nih.gov

La complexité de ces règles nécessite de faire appel à une méthode d'apprentissage relationnel capable de d'apprendre à partir de descriptions exprimées en termes de relations plutôt que par des couples attributs-valeurs [Alphonse et al., 2004]. Dans le projet ExtraPloDocs, ce travail est pris en charge par l'unité MIG de l'INRIA. Je mets l'accent ici sur la préparation des données d'apprentissage.

Elaboration d'un corpus d'apprentissage

Ces méthodes reposent sur un corpus d'apprentissage comportant des exemples positifs et négatifs[35] des relations à apprendre. Le projet Caderige a permis de créer un tel jeu d'exemples, sous la forme d'un ensemble de phrases sélectionnées à partir des résumés de Medline.

Le texte brut des exemples ne suffit pas cependant : il faut qu'ils soient interprétés. Pour ce faire, nous avons conçu, dans Caderige, un outil d'annotation permettant à un biologiste d'expliciter la lecture qu'il fait des textes[36]

Une fois ce travail d'interprétation effectué, les autres traitements sont entièrement automatisés.

The <agent type=protein>*GerE*</agent>*protein* <interaction type=negative>*inhibits* </interaction> <target type=expression>*transcription of* <source type=gene>*the sigK gene*</source> *encoding* <product>*sigmaK*</product></target>

FIG. 3.4 – Fragment annoté : les balises XML traduisent l'interprétation biologique qui est faite du fragment par l'annotateur.

Normalisation des textes

Un point critique pour l'apprentissage concerne la représentation des exemples. Plus la représentation est riche et complexe, plus l'apprentissage est difficile ; à l'inverse, plus la représentation est simplifiée, plus on risque d'éliminer des éléments importants. Dans le cas de l'apprentissage à partir de textes, ces questions sont cruciales. La diversité des formulations linguistiques est un défi pour les techniques d'apprentissage. Comme on ne peut pas augmenter la taille de la base d'exemples pour compenser son hétérogénéité

[35] Les exemples négatifs sont obtenus par complémentarité des exemples positifs : les couples de gènes et protéines pour lesquels aucune interaction n'est explicitée.

[36] L'annotateur CADIXE a été développé sous la direction de Gilles Bisson (LEIBNIZ-IMAG). Il repose sur une technologie XML.

(le coût de l'annotation des exemples est trop important), il faut réduire cette dernière.

L'une des originalité de la méthode d'acquisition que nous avons proposée dans Caderige repose sur une étape de normalisation préalable des fragments textuels. Cela revient à réduire les paraphrases, dans une démarche qui s'inspire de la méthode harrissienne [Dachelet, 1994]. Dans notre cas, ce qui fonde la notion de paraphrase et la possibilité d'une normalisation, c'est la référence que constitue l'interprétation biologique associée aux exemples. Pour prendre un exemple simpliste, si deux verbes comme *inhibits* et *blocks* portent la même étiquette <interaction type=negative>, c'est qu'ils sont considérés comme synonymes dans ce contexte.

Ce travail sur la normalisation est décrit à la section 3.5.3. Il devrait comporter différents volets, qui sont classiques chez Harris et dans les travaux sur la paraphrase : normalisation des synonymes, des nominalisations et du passif, typage sémantique, résolution des anaphores, réduction des modalités. La thèse de Davy Weissenbacher (débutée en 2003) s'inscrit dans ce cadre.

Le schéma de la figure 3.5 décrit le principe global de cette méthode d'acquisition et le rôle de la normalisation. Il met en évidence la collaboration des méthodes d'apprentissage et de TAL. Nous revenons plus en détail sur le module d'analyse (partie exploitation) des textes dans le chapitre suivant.

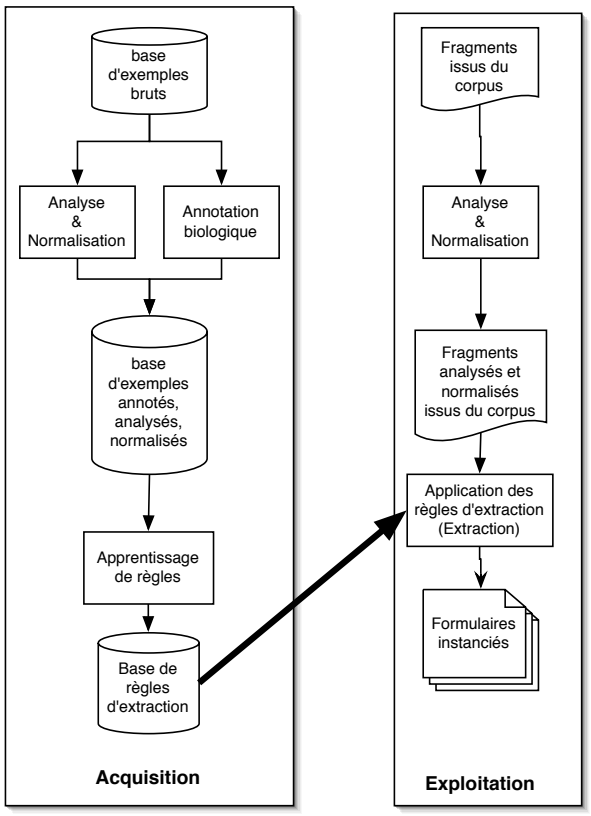

FIG. 3.5 – Méthode d'acquisition de règles

3.6 Conclusion

J'ai décrit dans ce chapitre différentes expériences sur l'acquisition de connaissances destinées à servir de ressources pour des systèmes d'accès au contenu des documents. Avant d'aborder le chapitre suivant qui mettra mieux en lumière leur rôle dans ces systèmes, je reprends ici les conclusions que l'on peut tirer de ces travaux pour l'acquisition de connaissances à partir de corpus.

Au-delà de la seule construction de classes sémantiques, la mise au jour des relations entre les différents types d'unités textuelles est un axe de recherche important. Malgré les travaux qui ont été réalisés dans cette voie depuis la fin des années 1990, l'intégration des différentes méthodes (distributionnelles, structurelles, à base de règles et par propagation) reste problématique : elle suppose de définir les modalités de la coopération avec l'utilisateur et une véritable démarche d'expérimentation.

La coopération

L'acquisition de ressources ne peut être entièrement automatisée. Que ce soit pour l'étiquetage préalable des exemples d'apprentissage, pour contrôler en cours d'acquisition le processus de généralisation ou pour valider des résultats *a posteriori*, une part d'interprétation humaine est nécessaire dans le processus d'acquisition. Ce discours est difficile à tenir face à des partenaires industriels ou à des financeurs qui, pour des raisons évidentes de coût, veulent du « tout-automatique ». Il faut pourtant le redire : une terminologie ou une ontologie sont des constructions : elles ne sont pas cachées dans les textes. L'exploitation du corpus sert à guider ce travail de construction [Aussenac-Gilles et al., 2004] et à en assurer l'ancrage linguistique, ce qui est indispensable si les connaissances ainsi construites servent de ressources pour des outils de TAL.

Prendre conscience du rôle de l'interprétation humaine impose de concevoir des interfaces ergonomiques pour encadrer ce processus coopération. Ce souci est ancien, [Bourigault et Jacquemin, 2000] le souligne. Dans les projets présentés ici, nous avons introduit l'interaction de différentes manières : outil d'annotation dans le projet Caderige, graphes des similarités pour visualiser la cartographie lexicale d'un corpus, interfaces de validation de SynoTerm et de IndDoc (voir chapitre suivant). Les expériences demeurent cependant ponctuelles et isolées.

Aucune réflexion d'ensemble n'a été élaborée, à ma connaissance. Je pense que la communauté de la terminologie computationnelle en France est mûre pour recenser les expériences qui ont été faites et en faire le bilan mais cela suppose sans doute de collaborer avec des spécialistes d'autres communautés plus à même d'apprécier les processus d'interaction homme-machine.

L'expérimentation

Au final, l'acquisition de ressources apparaît comme un processus complexe, où l'on cherche à intégrer différentes méthodes et différentes sources de connaissances.

Au-delà de l'effort de développement de ces différentes méthodes, il faut savoir comment combiner ces outils et tester des scénarios d'acquisition variés. La diversité des méthodes utilisables, la sensibilité des données textuelles aux traitements, la méconnaissance de la typologie des textes à analyser et la dimension coopérative rendent les expériences coûteuses à réaliser et difficiles à analyser. En pratique, pour chaque application, on « bricole » une méthode d'acquisition en se fiant à l'intuition et faisant avec les outils disponibles. D'une application à l'autre, les résultats ne sont ni reproductibles

ni comparables et on a du mal à construire une expertise sur la stratégie à adopter.

Toutes les combinaisons ne pouvant pas être testées en grandeur nature, il est donc important de se doter des outils pour les tester *a priori* sur la base d'échantillonnage ou par simulation. Différentes voies sont envisageables. Avec Mo'K, [Bisson et al., 2000] a proposé un atelier de construction d'ontologies à partir de textes permettant d'analyser le comportement de différentes mesures de distances et différents algorithmes de classification sur un jeu de données particulier. Concernant le problème de la reconnaissance des entités nommées, T. Poibeau a proposé des mesures pour prédire l'efficacité de différentes stratégies en fonction de certaines caractéristiques des corpus à traiter [Poibeau, 2003]. Dans la même perspective, et pour guider le choix des ressources lexicales à utiliser pour une application donnée, j'ai commencé à définir[37] des mesures permettant d'apprécier *a priori* l'adéquation et la couverture d'une ressource lexicale pour un corpus donné. Ces expériences sont encore ponctuelles mais il est indispensable d'avancer dans cette voie pour sortir du domaine du « bricolage ».

[37] Ce travail a été effectué en collaboration avec Goritsa Ninova, dans le cadre de son stage de DEA (2004).

Chapitre 4

Donner accès au contenu des documents

Les connaissances acquises sont utilisées comme ressources dans des systèmes de TAL. Dans la lignée de mes travaux initiaux sur la compréhension de texte, je me suis intéressée aux méthodes qui permettent à l'utilisateur d'« accéder au contenu des documents ».

Une perspective appliquée

Parler d'« accès », de « contenu » et de « documents » plutôt que d'« analyse », de « sens ou de sémantique », et de « textes » traduit une perspective appliquée.

« Donner accès » fait implicitement référence à l'utilisateur : c'est mettre l'accent sur l'objectif plutôt que sur les méthodes. Il s'agit de développer des outils qui facilitent, accélèrent, éventuellement remplacent le travail de « lecture » des documents pour l'utilisateur qui recherche des informations précises. Les outils à développer doivent prendre en compte le point de vue des utilisateurs. C'est de l'analyse précise de la diversité de leurs besoins que peuvent naître de nouvelles applications. Celles que je présente dans ce chapitre ont des visées différentes : la recherche de documents dans une base documentaire hétérogène, la recherche d'un type particulier d'information dans une base documentaire, l'exploration d'un document scientifique ou technique volumineux.

Le « contenu » s'oppose au contenant. Là où la recherche d'information vise à retrouver des documents dans une base documentaire, les outils dont il est question ici visent leur contenu. Il s'agit de permettre à l'utilisateur de « rentrer dans » les documents, la lecture pouvant être rendue difficile

par la taille des documents, leur nombre et leur hétérogénéité[1]. Le terme de « contenu » est volontairement vague : dans certains cas, l'utilisateur a besoin de repérer quelques éléments précis, dans d'autres, il souhaite au contraire avoir une vue globale.

Le terme de « document » rattache ces outils à la tradition de la recherche documentaire, où le document est considéré dans son contexte social et dans l'exploitation qui en est faite [Amar, 2000, p. 330] : même en s'en tenant aux documents sur support électronique et à leurs parties textuelles, il y a peu de choses en commun entre une page web, un article scientifique ou une documentation d'entreprise.

Mettre ainsi l'accent sur la fonction de ces outils tend à occulter la question des méthodes d'analyse à employer ainsi que l'étude plus théorique du fonctionnement linguistique des documents eux-mêmes. L'analyse linguistique et les méthodes de TAL sont considérées ici non pour elles-mêmes mais en vue d'une application. Il ne s'agit pas d'opposer des études théoriques et des travaux plus appliqués mais de montrer comment les seconds permettent d'éprouver les premières, qui gardent tout leur intérêt par ailleurs.

Des outils multiples

Dans ce qui suit, je présente deux des applications sur lesquelles j'ai le plus travaillé : l'indexation fine de documents comme support pour la navigation et la consultation de documents (section 4.1), l'extraction d'information pour assister les biologistes dans leur consultation bibliographique (section 4.2). J'aborde aussi la question de la recherche d'information spécialisée (section 4.3). Même s'il ne s'agit pas à proprement parler de donner accès aux contenus des documents, il y a en réalité un continuum entre accéder aux documents d'une base documentaire et accéder à leur contenu. Du point de vue de l'utilisateur tout d'abord : une recherche de documents suffisamment précise (langage de requêtes et indexation adaptées à un domaine de spécialité) facilite évidemment l'accès à leur contenu. D'un point de vue technologique, il y a aussi des points communs entre les méthodes d'acquisition de connaissances et d'analyse des documents.

Les trois applications[2] envisagées ici sont très différentes dans leurs objectifs et leurs méthodes mais, d'une application à l'autre, les mêmes questions se posent. On retrouve ici les problématiques que j'ai soulevées dans

[1] La diversité des langues est également un handicap important mais je n'ai pas travaillé sur cette dimension.

[2] D'autres techniques pourraient être proposées. Parfois, c'est un résumé qui est utile. Dans d'autres cas, une traduction est indispensable. Proposer une cartographie des documents est également une piste prometteuse. Je n'ai pas travaillé sur ces questions.

les chapitres précédents : comment combiner acquisition de connaissances et exploitation de ces connaissances ? quel est l'apport des connaissances générales et comment rendre les systèmes adaptables à de nouveaux domaines ? comment intégrer la coopération avec l'utilisateur ?

4.1 Indexation fine de documents

Les index qui figurent à la fin des ouvrages destinés à être consultés, souvent appelés « index de fin de livre » (*back-of-book indexes*), sont des outils traditionnels d'accès à l'information. D'où l'idée de créer des index similaires pour des documents électroniques.

Problématique des index de documents

Les index portent généralement sur un texte (parfois un ensemble de textes) qui a une certaine unité et qui est publié de manière autonome. L'index fait partie intégrante du document publié (d'où le terme d'*index de document*, que je préfère à « index de fin de livre »). Il est généralement créé par l'auteur du document ou son éditeur.

Indexer est un processus qui associe à un texte un ensemble organisé de descripteurs pertinents. Un *index* est une liste des termes représentatifs du contenu d'un document ou d'une base documentaire. Par rapport à l'indexation d'une base documentaire, l'indexation d'un document se différencie par la granularité de sa description et le fait que les liens entre l'index et le document doivent être explicites et interprétables par le lecteur. Un index comporte une *nomenclature*, liste structurée de descripteurs, et, pour chaque entrée de l'index, une liste de *renvois* vers les occurrences pertinentes du descripteur considéré dans le document.

Cette pratique de l'indexation est peu répandue et moins normalisée en France dans que dans l'édition anglo-saxonne [Mulvany, 1993]. Même pour l'anglais, cependant, il existe peu d'outils permettant d'indexer automatiquement les documents, ou du moins d'assister le travail de l'indexeur. Les logiciels les plus sophistiqués proposent un jeu d'entrées d'index et de renvois au texte en s'appuyant sur la structure des documents. Ils gèrent le tri et la mise en forme de l'index mais les étapes les plus coûteuses (choix des descripteurs à faire figurer comme entrées de l'index, rattachement des variantes à une forme canonique et sélection des passages du document auquel il convient de renvoyer) restent à la charge de l'utilisateur.

Motivation et contexte de la recherche

A l'origine de ma réflexion sur les index et les outils d'assistance à la construction d'index, il y a une double motivation.

J'avais travaillé sur l'acquisition de connaissances terminologiques, notamment sur la structuration de terminologie, et se posait le problème de l'évaluation. Les méthodes proposées ayant souvent été conçues pour acquérir un type particulier de relation, il était difficile de se faire une idée globale du degré de structuration qu'on pouvait obtenir. Par ailleurs, je l'ai déjà souligné, on ne peut évaluer la qualité des connaissances acquises qu'en fonction d'une tâche et sur une application. La construction d'index de document m'a ainsi paru fournir un cadre approprié d'évaluation pour un ensemble d'outils terminologiques (extracteurs de termes et de relations). C'était l'occasion de faire le bilan (forcément partiel, car lié à une tâche particulière) des travaux effectués dans le domaine en France[3].

La seconde motivation était plus directement appliquée. Il me semblait que l'état de maturité des méthodes de TAL et des outils terminologiques permettait d'assister la construction d'index plus que ne le faisaient les outils commercialisés. Il fallait donc tenter de construire un prototype qui en fasse la démonstration.

Ce travail sur la construction d'index n'a réellement démarré qu'avec la thèse de T. Aït El Mekki (2000-2004) qui a permis d'élaborer le prototype d'un outil d'aide à la construction d'index de documents : IndDoc. Nous avons également participé au projet CEDERILIC[4], projet coordonné par Jean Charlet (STIM, AP-HP) et soutenu par France-Télécom, qui visait à construire un index thématique pour un ouvrage scientifique[5]. L'ouvrage doit être publié fin 2004 à la fois sous forme papier classique et sur cédérom : il existe donc deux versions de l'index pour ces deux supports.

Je décris ici le travail réalisé autour de IndDoc. Après avoir présenté l'approche proposée dans IndDoc (4.1.1) et la méthode d'acquisition des connaissances mise en oeuvre (4.1.2), je commente les résultats obtenus (4.1.3).

---

[3] D. Bourigault et J. Charlet avait élaboré un projet similaire en parallèle, mais avec une ambition moindre puisque seule l'extraction de termes étaient envisagée [Bourigault et Charlet, 1999]. Nous avons ensuite collaboré dans le cadre du projet CEDERILIC.

[4] CEDERILIC pour « CEDErom pour Indexer le Livre IC ».

[5] Cet ouvrage [Teulier et al., 2004] est constitué d'une sélection de vingt-et-un articles de trois éditions des journées Ingénierie des connaissances (1999-2001).

4.1.1 Approche générale

Un index se compose d'une nomenclature et d'un ensemble de renvois [Ait El Mekki et Nazarenko, 2001]. La nomenclature de l'index est une liste structurée des descripteurs qui constituent les entrées de l'index (voir figure 4.1). Ces entrées sont reliées les unes aux autres par des relations syntaxiques ou sémantiques. Un renvoi établit un lien entre une entrée de la nomenclature et une page, un groupe de pages ou une autre (sous-)entrée. Pour une entrée donnée, l'index ne renvoie qu'aux occurrences les plus pertinentes dans le document.

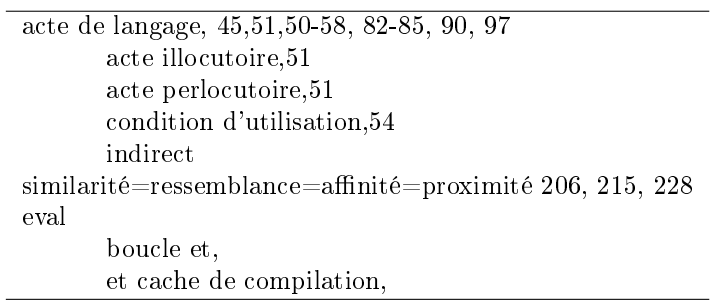

FIG. 4.1 – *Extrait d'index montrant des relations entre descripteurs.*

L'approche que nous avons proposée pour l'élaboration d'index se décompose en trois étapes (figure 4.2) :

- L'acquisition permet d'élaborer un index source qui est formé de l'ensemble des connaissances (descripteurs, renvois, relations) qui vont ensuite former l'index final.
- La génération permet de sélectionner dans l'index source le sous-ensemble des connaissances qu'on juge devoir être effectivement publiées.
- La visualisation, enfin, met en forme ces connaissances pour produire un index exploitable par le lecteur.

Dès lors qu'on considère l'index comme partie intégrante du document et donc de son cycle de vie, il convient de conserver un index source relativement stable, indépendant des contraintes de style et de volume imposées par l'éditeur, et qui puisse être repris et mis à jour pour une nouvelle publication.

4.1.2 Méthode d'acquisition

L'élaboration de l'index source constitue le coeur du processus de construction d'index. Elle est détaillée sur la figure 4.2 [Aït El Mekki et Nazarenko, 2004]. Elle comporte quatre étapes.

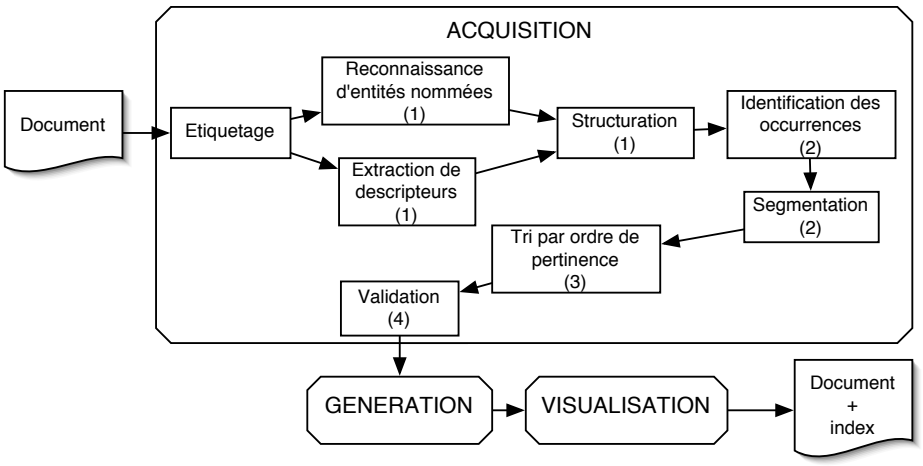

FIG. 4.2 – Architecture de IndDoc

L'analyse du document (1) identifie les descripteurs. Elle repose sur le repérage des termes et des entités nommées du document. La liste des descripteurs doit ensuite être structurée par l'ajout de relations sémantiques. Les techniques de structuration de terminologie présentées en section 3.4 sont mises en oeuvre ici : elles permettent d'associer à un descripteur l'ensemble de ses variantes, de ses synonymes, hyperonymes, meronymes.

Le calcul des renvois consiste à segmenter le document (2) relativement à chaque descripteur pour identifier les passages devant figurer comme renvois dans l'index, la taille des segments de renvoi variant en effet selon qu'il faut renvoyer à un paragraphe précis ou à un passage plus large.

Pour identifier les termes et entités nommées les plus importants à faire figurer dans l'index, ainsi que les passages de texte à citer en priorité, il faut également trier ces éléments par ordre de pertinence (3). Différents indices sont exploités : fréquence et répartition dans le document, mise en relief à l'aide de marques typographiques ou linguistiques, structure du document, position des occurrences dans le document, etc.

Ces méthodes permettent d'élaborer une ébauche d'index, mais celui-ci doit être retouché par un indexeur. Une interface de validation (4) a été conçue pour permettre à l'indexeur de retravailler les résultats produits automatiquement [Aït El Mekki, 2004].

Au final, il faut souligner la complexité de ce processus d'indexation, la diversité des techniques auxquelles il fait appel et la variété des indices (structurels, lexicaux, discursifs, syntaxiques, statistiques) sur lequel il repose. Il s'agit là d'un premier travail significatif d'intégration des outils d'acquisition.

4.1.3 Résultats et évaluation

Méthode d'évaluation

Evaluer la qualité d'un index est une tâche difficile dans la mesure où il n'existe pas d'index de référence ou même de norme clairement établie. Comme tous les outils coopératifs, il est difficile de faire la part de l'outil et la part de l'humain. Nous avons cependant tenté d'apporter des éléments d'évaluation.

Evaluer précisément chacun des modules de IndDoc (extraction, structuration, segmentation, mesure de pertinence) est inutile dans la mesure où ils sont fortement interdépendants : le tri peut corriger les défauts de l'extraction terminologique, par exemple). C'est le comportement global du système d'acquisition automatique qui doit être évalué. Nous avons donc analysé la qualité globale des résultats obtenus après le calcul de pertinence, en utilisant une mesure classique de r-précision (précision triée ou *ranked-precision*[6]).

Deux points de comparaison ont été considérés. Nous avons travaillé sur des ouvrages déjà publiés avec un index pour comparer l'index obtenu avec IndDoc et l'index d'auteur. La comparaison permet d'apprécier l'apport d'un outil comme IndDoc pour un indexeur inexpérimenté et de vérifier que le silence reste faible dans les résultats produits automatiquement. Nous avons par ailleurs fait valider les index produits par IndDoc par des tiers[7]. Le résultats de ces validations permettent de mesurer la part de bruit dans les résultats de IndDoc.

Synthèse des résultats

T. Aït El Mekki a testé son système sur quatre corpus différents[8]. Ces résultats sont décrits en détail dans [Aït El Mekki et Nazarenko, 2004].

Certains critères de qualité sont garantis par construction. L'automatisation même du processus d'indexation assure une bonne couverture de l'index, la cohérence formelle de l'index généré, la possibilité d'avoir un index

[6]La r-précision est une mesure de précision qui tient compte du fait que les résultats sont ordonnés. La r-précision est d'autant meilleure que la précision simple est bonne et que le tri place les bons résultats en tête de liste et les mauvais en queue de classement. Cette mesure a été initialement mise au point pour apprécier la précision dans les résultats fournis par un moteur de recherche [Baeza-Yates et Ribeiro-Neto, 1999].

[7]Ce travail a été fait par Lise Pâris au cours de son stage de maîtrise [Pâris, 2003], par Jean Charlet et Régine Teulier dans le cadre du projet CEDERILIC.

[8]Il s'agit de recueils d'articles scientifiques en ingénieries de connaissances ([Bourigault et Charlet, 1999] et le corpus du projet CEDERILIC) et de deux monographies ([Nazarenko, 2000] et [Kayser, 1997]).

de bonne taille (l'index produit dans le cadre du projet CEDERILIC comporte ainsi environ 2 000 entrées).

Les différents modules de traitement produisent des résultats contrastés d'un corpus à l'autre, même abstraction faite de leur différence de taille. Le calcul des mesures de pertinence associées aux descripteurs et aux renvois repose sur un faisceau de facteurs : les marques d'emphase typographique, la structuration du document en différentes parties, la redondance et la distribution du vocabulaire, des tournures de mise en relief discusive, etc. Chacun de ces facteurs est plus discriminant pour certains textes et moins pour d'autres. C'est donc leur combinaison qui confère au système une certaine robustesse. La même remarque vaut pour la structuration de la nomenclature qui exploite quasiment tout l'éventail des méthodes de structuration présentées plus haut. C'est l'une des leçons qui peut être retenue de cette expérience d'intégration.

La confrontation des index sources aux index d'auteurs et l'analyse des résultats de validation apportent des éléments intéressants. Il est certes plus coûteux en temps de construire un index à l'aide d'IndDoc que de manière entièrement artisanale mais le résultat est plus riche et de meilleure qualité. Ces résultats, sans être encore probants, sont encourageants.

4.1.4 Bilan et perspectives

Ce travail sur l'indexation fine des documents ne s'achève pas avec la thèse de T. Aït El Mekki mais il est intéressant de faire un bilan partiel du travail effectué.

L'apport de la terminologie computationnelle

Nous avons exploité des outils terminologiques dans une optique d'indexation : il faut dresser le bilan de cette expérience. Divers outils ont été exploités pour extraire les termes, les mettre en relation et repérer leurs variantes.

Le bilan est globalement positif mais l'intégration de ces outils s'est avérée délicate. Ce sont pour la plupart des prototypes de laboratoire qui n'ont pas l'état d'achèvement souhaitable pour être facilement importés dans une chaîne de traitement complexe : leur disponibilité n'est pas toujours assurée, les formats sont hétérogènes et fluctuants. Certaines fonctionnalités font défaut. Ce problème d'intégration est particulièrement critique pour les outils de structuration, ce qui confirme l'analyse faite plus haut (section 3.4.4).

L'expérience de l'utilisation des outils terminologiques pour l'indexation de document et le bilan que nous avons fait est l'un des éléments qui nous ont amenés, au sein de l'équipe « Représentation des Connaissances et Langage

Naturel » (RCLN) du LIPN, à investir dans le redéveloppement de certains de ces outils, à commencer par un extracteur de termes.

Les limites de l'approche générique

Pour le premier prototype IndDoc, nous avons cherché à élaborer un outil d'indexation générique de documents. Pour améliorer la qualité de l'indexation, il faut maintenant viser des outils d'indexation spécialisés. Nous avons vu plus haut que les méthodes d'acquisition sont très dépendantes des corpus et des connaissances préalables.

C'est pour explorer cette voie, que j'ai proposé dans le cadre d'un Plan PluriFormations, qui devrait débuter en 2005, de travailler sur un outil d'indexation dédié aux documents médicaux. Il faut mesurer l'apport de connaissances extérieures : elles doivent permettre d'introduire dans l'index des descripteurs non attestés dans le document et être prises en compte pour les mesures de pertinence[9]. Il faudrait également réfléchir à une typologie des documents à indexer, cette question des genres des documents étant sans doute plus facile à aborder dans un domaine de spécialité que dans toute sa généralité.

Une évaluation plus systématique

Il faut tester IndDoc de manière plus systématique. Pour aller dans ce sens, j'ai déposé un projet de coopération Franco-Québécoise avec l'École de bibliothéconomie et des sciences de l'information de l'Université de Montréal, pour les années 2005-06 (ce projet est porté par Lyne Da Sylva pour la partie québécoise). Il s'agit en effet de profiter de l'expertise et de la pratique de l'équipe québécoise en matière d'indexation humaine et automatique pour tester et évaluer le prototype IndDoc par rapport aux pratiques des indexeurs professionnels et aux logiciels existants.

De nouveaux outils de navigation

Comment l'index produit à l'aide d'IndDoc peut-il être exploité ?

La dimension électronique du document et de son index doit permettre de concevoir des formes plus novatrices d'index. Nous avons négligé la question de la visualisation alors que les nombreuses propositions ont été faites pour l'indexation de bases documentaires (surlignage, vues panoramiques et

[9]Une première expérience a été faite par T. Aït El Mekki dans ce sens : elle a exploité la nomenclature du premier corpus IC [Bourigault et Charlet, 1999] dans le calcul du poids des descripteurs du deuxième corpus (CEDERILIC). La piste semble prometteuse.

zoom, focalisation à l'aide de loupes, notamment pour construire des classifications et cartographies des documents retournés par un moteur de recherche [Hearst, 1999].

IndDoc devrait pouvoir être exploité pour l'indexation de sites web, même si les indices de pertinence diffèrent. Il s'agit plus largement de toute la problématique de la navigation et de l'hypertextualisation des documents, bases documentaires ou sites web [Anick, 2001][Wacholder et Nevill-Manning, 2001].

Se pose également la question de la connexion des bases documentaires et des bases de données, question particulièrement sensible dans le domaine de la biologie moléculaire où la politique de publication en ligne des articles et des résultats scientifiques est bien établie. La méthode d'indexation fine proposée ici pourrait être adaptée à ce problème.

Si elle voit le jour, la collaboration avec l'École de bibliothéconomie et des sciences de l'information de l'Université de Montréal devrait permettre d'aborder ces questions liées à la problématique générale des outils et méthodes de navigation.

4.2 Extraction d'information

Dans le prolongement de mon travail sur la compréhension automatique de texte, je me suis intéressée à l'extraction d'information : j'ai encadré la thèse de Thierry Poibeau sur ce thème et monté différents projets en partenariat avec des partenaires extérieurs (Astuxe, Caderige[10], ExtraPloDocs[11]). Après les expériences des conférences MUC [MUC, 1995][MUC, 1998], l'adaptation des outils apparaît comme un enjeu important pour l'extraction d'information [Poibeau et Nazarenko, 1999]. L'essor des techniques d'apprentissage en extraction d'information [Riloff, 1993, Soderland, 1999] correspond à ce mouvement : l'adaptation nécessite de pouvoir apprendre ou plus largement acquérir à un coût raisonnable les ressources nécessaires au moteur d'extraction, ressources lexicales et règles d'extraction, notamment.

Le domaine de la génomique est apparu comme un champ d'expérimentation exemplaire. Du point de vue documentaire, la génomique fonctionnelle, qui étudie la fonction des gènes, se caractérise à la fois par l'ampleur des besoins et par la maturité du domaine.

Le développement des analyses à grande échelle a profondément modifié les pratiques des chercheurs. Il est désormais possible, pour un individu ou une petite équipe, d'analyser simultanément l'activité de plusieurs milliers de gènes dans une cellule ou un organisme biologique. Le travail bibliographique

[10] www-leibniz.imag.fr/SICLAD/Caderige

[11] http ://www-lipn.univ-paris13.fr/p̃oibeau/Extra/

associé doit donc impérativement être assisté d'outils de navigation et de consultation suffisamment puissants pour permettre une activité de veille scientifique efficace.

D'autres domaines des sciences et des techniques sont porteurs d'une telle demande mais la biologie est l'un des secteurs dans lequel elle s'y est d'abord manifestée. La maturité des besoins et la disponibilité de l'information textuelle et de bases de connaissances[12] font de la biologie moléculaire un champ d'investigation exemplaire pour les techniques d'accès au contenu.

Le projet ExtraPloDocs vise à élaborer un système d'extraction d'information qui soit adaptable et dont toutes les ressources (lexicales, syntaxiques et sémantiques, y compris les règles d'extraction elles-mêmes) puisse être acquises à partir de texte. Outre le LIPN qui en est le coordinateur, ce projet comporte deux autres partenaires : l'unité MIG de l'INRA représentée par Claire Nédellec (informaticienne, spécialiste de l'apprentissage symbolique) et Philippe Bessières (biologiste) ainsi que la société Isoft impliquée dans le secteur de la fouille de données en biologie. Une étude d'analyse de faisabilité avait été menée au préalable dans le projet Caderige qui rassemblait davantage d'équipes et plus de biologistes [Alphonse et al., 2004].

4.2.1 Approche générale : acquisition et production

Le projet ExtraPloDocs propose une approche ambitieuse pour l'exploration de la documentation scientifique : il s'agit de localiser et de formaliser l'information pertinente au regard du besoin exprimé par un utilisateur sous la forme d'un formulaire à remplir. L'approche d'ExtraPloDocs repose sur une démarche en trois étapes, l'extraction d'information étant la principale :

1. La *recherche d'information* permet de sélectionner dans la base bibliographique Medline[13] un ensemble de textes pertinents au regard du problème étudié par les biologistes. La requête « Bacillus subtilis transcription » permet ainsi d'isoler environ 2 700 articles dans la base de Medline. ExtraPloDocs repose sur le moteur de recherche Medline pour cette étape. On obtient des résumés comme celui de la figure 4.3, p. 69.
2. La *sélection des fragments* permet d'identifier dans les articles sélectionnés à l'étape précédente la portion de texte qui paraît pertinente

[12] L'essor d'internet et des services WWW a favorisé en biologie moléculaire l'apparition de bases de connaissance en lignes (MESH, GenOntology, par exemple) et la mise à disposition d'une information textuelle riche et très bien structurée dans des journaux scientifiques, également accessibles en ligne.

[13] www.ncbi.nlm.nih.gov

pour l'utilisateur d'après un jeu d'exemples qu'il fournit. Dans le cas des interactions géniques, au vu des expériences effectuées, nous évaluons à 3 % la portion de texte pertinente. Il est donc essentiel d'effectuer cette étape de filtrage avant d'aborder le traitement lourd d'extraction d'information à proprement parler.

3. Enfin, l'*extraction d'information* permet d'extraire et formaliser l'information pertinente. Par exemple, à partir du fragment souligné dans le résumé de la figure 4.3, le module d'extraction instancie le formulaire associé. C'est sur cette troisième étape que l'équipe du LIPN est le plus impliquée. Je mets l'accent sur celle-ci.

Le projet ExtraPloDocs repose sur deux idées force : la place centrale de l'acquisition dans le système d'extraction et le nécessité de procéder à une normalisation des fragments de texte en amont.

La place centrale de l'acquisition

L'extraction d'information met en oeuvre des ressources variées (lexique morpho-syntaxique, règles d'extraction, dictionnaires d'entités nommées, terminologies...) qui sont souvent dépendantes du type de corpus à traiter et plus globalement de la tâche à réaliser. Ces ressources sont un frein à l'adaptation des outils. Pour avoir un système d'extraction adaptable, il faut y incorporer des outils d'acquisition ou d'apprentissage des ressources. Le projet ExtraPloDocs s'inscrit ainsi dans la lignée de mon travail de recherche sur l'acquisition de connaissances.

L'architecture du système ExtraPloDocs distingue nettement les deux systèmes d'acquisition et de production :

- La chaîne de traitement en production comporte les étapes suivantes : pré-analyse des documents, sélection des fragments potentiellement pertinents, analyse des fragments et extraction. Chacun de ces modules fait appel à des connaissances qui sont acquises dans le sous-système d'acquisition (figure 4.4, p. 70) .
- Ce sous-système d'acquisition (figure 4.6, p. 73) repose sur les mêmes modules d'analyse que le sous-système de production. En effet, il faut que les connaissances soient acquises à partir de corpus homogènes à ceux pour lesquels les connaissances seront appliquées. On a donc les mêmes étapes d'analyse de texte en production et en acquisition et ce sont les mêmes modules d'analyse qui sont utilisés de part et d'autre.

Ces deux sous-systèmes de production et d'acquisition partagent les mêmes modules d'analyse mais ne s'appliquent pas aux mêmes données et ne répondent pas aux mêmes contraintes de temps. Le sous-système d'acquisition

Extrait d'un résumé de Medline

UI - 99175219
AU - Ichikawa H
AU - Halberg R
AU - Kroos L
TI - Negative regulation by the Bacillus subtilis GerE protein.
...
PT - JOURNAL ARTICLE
...
DP - 1999 Mar 19
TA - J Biol Chem
AB - GerE is a transcription factor produced in the mother cell compartment of sporulating Bacillus subtilis. It is a critical regulator of cot genes encoding proteins that form the spore coat late in development. Most cot genes, and the gerE gene, are transcribed by sigmaK RNA polymerase. Previously, it was shown that *the GerE protein inhibits transcription in vitro of the sigK gene encoding sigmaK*. Here, we show that GerE binds near the sigK transcriptional start site, to act as a repressor. A sigK-lacZ fusion containing the GerE-binding site in the promoter region was expressed at a 2-fold lower level during sporulation of wild-type cells than gerE mutant cells. Likewise, the level of SigK protein (i. e. pro-sigmaK and sigmaK) was lower in sporulating wild-type cells than in a gerE mutant.
...

Formulaire

Interaction	*type*	negative		
	agent	GerE protein		
	target	expression	*source*	sigK gene
			product	sigmaK

Fig. 4.3 – Exemple de remplissage d'un formulaire à partir d'un résumé de Medline. Le fragment pertinent figure ici en italique.

analyse un corpus d'acquisition construit pour être représentatif des données textuelles de l'application. Le sous-système de production doit à terme pouvoir analyser rapidement de gros volumes de textes.

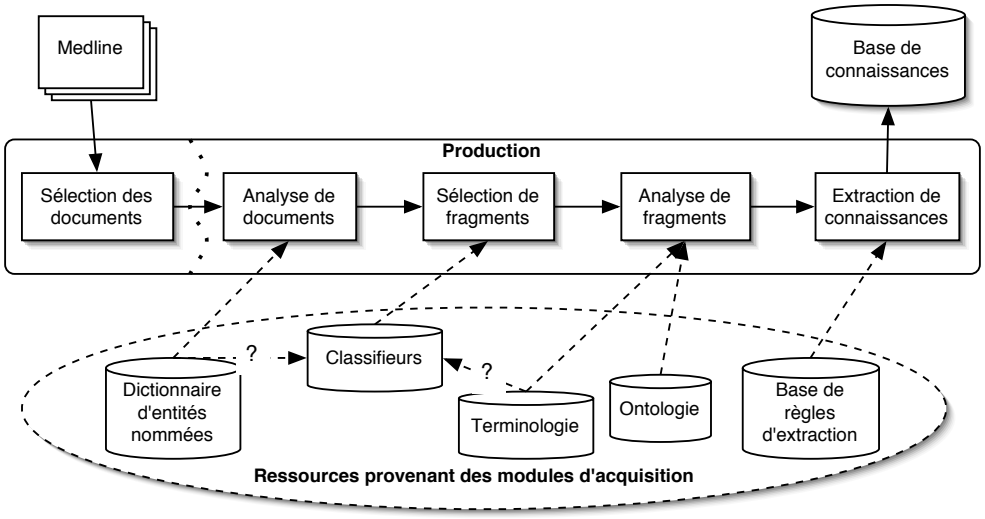

FIG. 4.4 – Architecture du système ExtraPlodocs

L'analyse de texte vue comme normalisation

Comme je l'ai mentionné plus haut (section 3.5), l'application des méthodes de TAL vise à donner une représentation normalisée des textes de manière à faciliter l'application et l'apprentissage de règles d'extraction.

Normaliser des phrases ou des fragments ne signifie pas directement réduire des paraphrases en ramenant les différentes formulations à une forme canonique. Normaliser des fragments de texte consiste ici à les enrichir d'annotations syntaxiques et sémantiques (étiquettes associées aux unités et relations entre ces unités). Ces annotations considérées isolément donnent une représentation abstraite des fragments et permettent de les rapprocher indépendamment de leur formes de surface.

L'analyse des fragments de texte consiste ainsi à enrichir les fragments de tout un ensemble d'annotations allant des catégories syntaxiques et relations de dépendance aux types et idéalement aux rôles sémantiques. La figure 4.5 présente un exemple de fragment annoté[14]. Comme je l'ai mentionné précé-

[14]Concrètement, nous avons défini pour cela un format de d'annotation pour ces corpus enrichis. Ce travail a été mené en collaboration avec Guillaume Vauvert, en postDoc dans

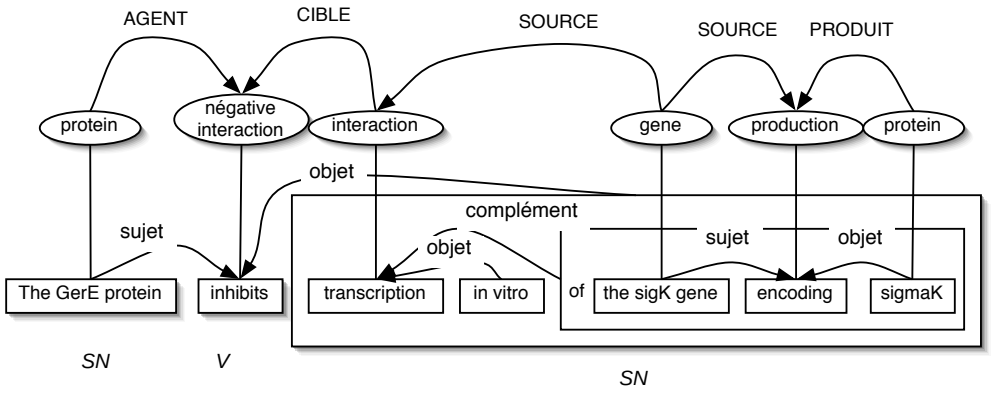

FIG. 4.5 – Extraction d'information : normalisation des fragments

demment (section 3.5), le niveau de normalisation à atteindre n'est pas fixé *a priori*, il est déterminé par l'interprétation que les biologistes donnent des exemples d'apprentissage. Si deux fragments différents sont interprétés de la même manière par l'utilisateur (dans notre cas le biologiste) ils peuvent être considérés comme équivalents relativement à l'application visée et les différences entre les deux fragments initiaux tendent à être considérés comme non significatives.

4.2.2 La mise en place d'une chaîne de traitement

La mise en place de la chaîne de traitement assurant l'analyse des textes nécessite un important effort d'ingénierie linguistique auquel l'équipe RCLN s'est attelée. Ce travail n'est possible que par l'investissement de plusieurs chercheurs et enseignants chercheurs (T. Hamon et T. Poibeau, notamment) et grâce au renfort d'« ingénieurs TAL » contractuels (S. Aubin et J. Derivière)[15].

Un tel effort ne se justifie que dans la mesure où il entre dans le projet plus global de développement d'une boîte à outils logiciels dédiée à l'analyse de corpus spécialisés. A quelques exceptions près, en effet, ce sont les mêmes modules d'analyse qui sont utilisés pour l'indexation fine de documents (sec-

l'équipe RCLN pendant 6 mois. Il a défendu l'idée d'un format d'annotation déporté (*stand-off annotations*) pour laisser le texte d'origine inchangé. Toutes les annotations sont stockées séparément avec des références aux positions des caractères dans le texte original [Vauvert, 2004] présente une première description de ce format d'annotation.

[15]Les projets ExtraPloDocs et Alvis permettent de financer la première partie de ce travail d'ingénierie.

tion 4.1) et pour la recherche d'information spécialisée (section 4.3). Le projet d'ensemble vise donc à construire des briques de base qui puissent être agencées de différentes manières pour tester différentes approches et différents outils d'accès au contenu.

La figure 4.6 montre en détail l'agencement de ces différents modules élémentaires dans la chaîne de traitement d'ExtraPloDocs. L'analyse repose sur quelques modules essentiels qui entrent dans la chaîne de traitement en production et auxquels correspondent différents modules d'acquisition :

- Le module de reconnaissance d'entités nommées exploite un dictionnaire de noms d'entités qui peut être acquis à partir de corpus ;
- Le module de sélection de fragments est constitué d'un classifieur qui est appris à partir d'un ensemble d'exemples de fragments étiquetés comme pertinents/non pertinents et qui permet en production de sélectionner les seuls fragments pertinents ;
- L'étiqueteur terminologique exploite une terminologie certifiée du domaine ou une liste de termes acquis à partir de corpus ;
- L'analyseur syntaxique est adapté pour les textes de biologie ;
- L'étiqueteur sémantique utilise les résultats d'un module d'acquisition d'ontologies [Faure et Nédellec, 1999], qui repose sur une méthode d'analyse distributionnelle ;
- Le module d'extraction d'information repose sur un ensemble de règles qui sont apprises en corpus par le module d'apprentissage de règles d'extraction (voir section 3.5, p. 47).

Dans ce qui suit, je décris le rôle et le fonctionnement de ces différents modules sur lesquels nous travaillons en mettant l'accent sur la partie analyse (production), les principaux mécanismes d'acquisition ayant été décrits dans le chapitre précédent.

Reconnaissance des noms d'entités

Le module d'étiquetage des noms d'entités (souvent appelés entités nommées, par abus de langage) a une place centrale dans les processus mis en œuvre pour accéder au contenu des documents (systèmes de question/réponse, indexation fine de documents, etc. [Nazarenko, 2005]). Cette place se justifie à la fois du point de vue technologique et du point de vue sémantique.

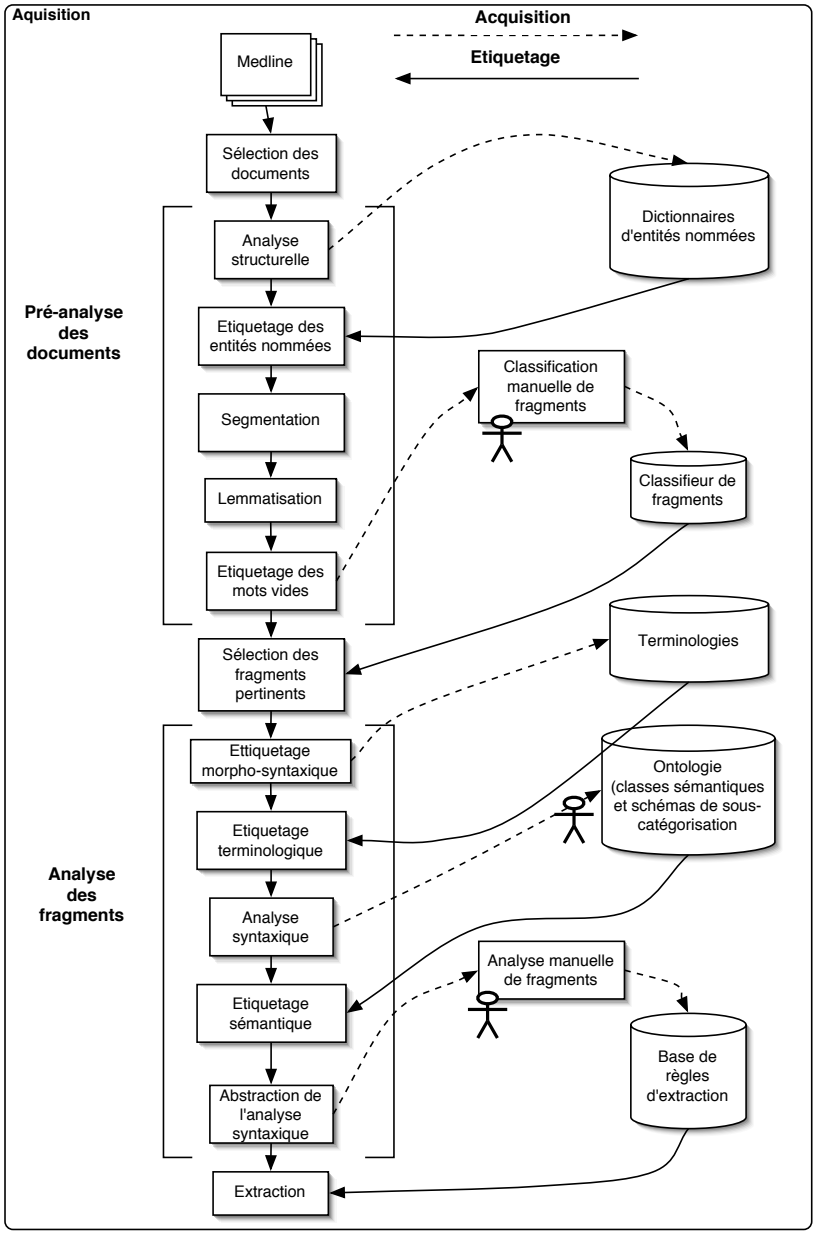

FIG. 4.6 – Les différentes étapes d'acquisition dans le système d'ExtraPlo-Docs.

Sur le plan technologique, la reconnaissance des noms d'entités est une étape critique. Les noms propres, les éléments chiffrés, les dates sont des éléments perturbateurs pour l'analyse des documents : il est donc impératif de les reconnaître. Par ailleurs, ce sont des îlots faciles à repérer dans le

flux textuel[16] et relativement fiables sur lesquels le reste de l'analyse peut s'ancrer.

Ces noms d'entités jouent en fait un rôle clef dans le flux informationnel que constitue un texte : de même que les noms de personnes ou d'organisations, les noms géographiques et les dates dans les textes d'actualité, les noms de gènes et de protéines sont importants à repérer dans les articles scientifiques de génomique. Toutes ces unités fonctionnent de la même manière, comme des noms propres *i.e.* comme des « désignateurs rigides » qui fixent une référence [Kripke, 1972]. C'est parce que ces noms d'entités servent d'ancres référentielles dans le flux informationnel qu'ils sont importants à repérer : savoir qu'un texte de presse parle de « Al Qaïda » et de « Madrid » ou qu'un article de génomique comporte les termes « Bacillus Subtilis » et « Spo3D » donne une indication assez précise sur leur contenu. La reconnaissance des noms d'entités permet d'identifier les référents du discours, ce qui est essentiel dès lors qu'on veut associer un sens référentiel au fragments de texte qu'on analyse[17].

Dans le cadre du projet ExtraPloDocs, un module de reconnaissance d'entité nommés assez classique a été développé au LIPN. Il a été réalisé par Jean-François Berroyer, dans le cadre de son stage de DEA, sous la direction de T. Poibeau [Berroyer, 2004]. Il fonctionne pour l'anglais et le français et repose sur des dictionnaires de noms d'entités classiques.

[16] La reconnaissance repose généralement sur des dictionnaires de noms d'entités et des règles implémentées sous la forme d'automates. C'est l'acquisition des dictionnaires d'entités et des règles qui peut soulever des difficultés. On obtient jusqu'à 95% pour une mesure combinée de précision et de rappel sur les tâches de type MUC [MUC, 1998]. La compétition que nous avons organisée lors du workshop BioNLP-NLPBA (International Joint Workshop on Natural Language Processing in Biomedecine and its applications, associé à COLING 2004) a montré qu'on atteint des scores de 75% pour la reconnaissances et le typage des entités biologiques [Kim et al., 2004, Zhou et Su, 2004].

[17] Cette approche de l'identification des référents du discours a cependant ses limites. Elle est partielle dans la mesure où seuls les référents nommés sont identifiés, les référents construits par détermination étant beaucoup plus difficiles à repérer. Elle est restreinte aux seules entités nominales, le plus souvent. Elle peut également être bruitée dans la mesure où identifier le référent d'une entité nommée n'est pas toujours immédiat : on sait que les noms propres sont ambigus, (si « Poincarré » paraît clair, c'est relativement à un contexte temporel et culturel donné), qu'ils sont sujets à variation (ce qui nécessite de repérer les formes équivalentes) et sont parfois employés de manière métaphorique (« le Einstein de l'économie ») ou métonymique («Le Goncourt est un jeune écrivain »[Kayser, 1988]). Cette opposition entre entités nommées (que l'on suppose être des désignateurs rigides, moyennant dans le meilleur des cas la vérification de quelques contraintes contextuelles de désambiguïsation) et d'autres groupes nominaux est davantage pragmatique que théorique.

Sélection de fragments

La sélection de fragment permet de focaliser l'analyse sur les seuls fragments pertinents.

Ce module a été développé par MIG, l'un des partenaires du projet ExtraPloDocs. Après étude de différentes méthodes d'apprentissage supervisé, il s'avère qu'un classifieur bayésien naïf couplé avec un mécanisme de sélection d'attributs permet d'obtenir un bon score de sélection (environ 85% de mesure combinée de précision et rappel) [Nédellec et al., 2001]. Il suffit d'avoir un corpus segmenté en phrases, dans lequel les noms d'entités ont été marqués et les mots vides éliminés. Une préparation linguistique plus sophistiquée du corpus ne semble pas avoir d'impact significatif[18].

Etiquetage terminologique

L'étiquetage terminologique consiste à identifier les segments de texte qui sont supposés correspondre à des termes du domaine. Il repose sur une liste de termes mais tout l'enjeu consiste à repérer en corpus les occurrences variantes de ces termes.

Le repérage de termes, notamment des termes polylexicaux, joue un double rôle dans le processus d'analyse :

- Il réduit la complexité de l'analyse syntaxique en identifiant certains groupes nominaux terminologiques et en supprimant des ambiguïtés de rattachement prépositionnel. Le gain en termes de qualité de l'analyse syntaxique est potentiellement important. Une étude menée par Danièle Cohen dans le cadre de son stage de fin d'études effectué au LIPN [Cohen, 2001] a montré sur un corpus différent que le taux d'erreur peut diminuer de 40% pour les relations de dépendance les plus importantes (relations sujet-verbe et verbe-objet, notamment). L'étude est en cours pour les textes de biologie.
- Il identifie les unités textuelles qui sont pertinentes sur le plan sémantique. Ce sont ces unités sémantiques (plus que les mots pris isolément) qui sont à prendre en compte pour l'acquisition d'ontologies et l'extraction d'information.

Les aspects terminologiques du projet ExtraPloDocs sont pris en charge par T. Hamon. L'étiquetage des termes repose sur une terminologie préexistante (terminologie certifiée du domaine ou acquise à partir de corpus). Les

[18] Des expériences ont notamment été faites par Zhyiu Qian dans le cadre de son stage de DEA pour mesurer l'impact de la lemmatisation et de la prise en compte des termes. Pour la sélection de fragments dans un domaine très spécialisé et avec un corpus d'apprentissage réduit, prendre en compte la terminologie me semblait en effet pouvoir améliorer les performances. Les premiers résultats semblent montrer que non.

variantes de ces termes sont calculées à l'aide de FastR [Jacquemin, 1997] et l'ensemble des termes initiaux et de leurs variantes est projeté sur le corpus pour permettre l'étiquetage des différentes occurrences de termes (avec ou sans variation).

Analyse syntaxique

L'analyse syntaxique entre dans le processus de normalisation des textes. Il s'agit de représenter le fragment comme un ensemble de relations de dépendance syntaxique. Ces relations de dépendance sont utilisées dans le processus d'acquisition de classes sémantiques qui reposent sur l'exploitation des contextes syntaxiques (comme dans l'expérience décrite en section 3.3). Elles sont également utiles lors de l'extraction d'information, puisque les règles d'extraction peuvent comporter des contraintes sur les dépendances syntaxiques que les éléments à prendre en compte entretiennent (voir l'exemple de règle donné en section 3.5).

Dans le projet ExtraPloDocs, Sophie Aubin est responsable du module d'analyse syntaxique. Après avoir comparé quelques analyseurs, elle a sélectionné le LinkParser à la fois pour la qualité de ses résultats, pour son adaptabilité et sa disponibilité. L'adaptation au domaine de la biologie suppose principalement d'enrichir le dictionnaire (par ex. *coenzyme*, *sucrose*), de modifier certaines étiquettes syntaxiques et certaines règles de grammaire (par ex. le verbe *initiate*, normalement transitif direct fonctionne ici comme un verbe intransitif) et d'intégrer les découpages syntaxiques fournis par l'analyse terminologique (voir figure 4.7).

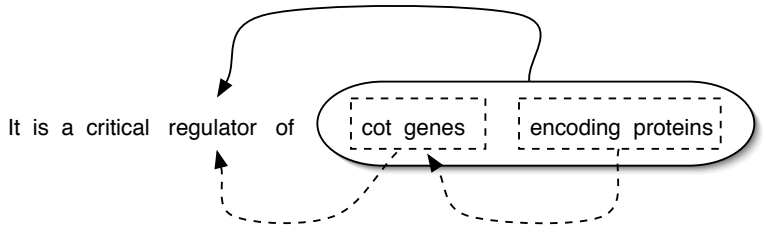

FIG. 4.7 – Rôle de la terminologie pour la désambiguïsation syntaxique. Dans les cas d'ambiguïté pour le rattachement des compléments, deux analyses sont syntaxiquement possibles (représentées ici en traits plein et traits pointillés. La possibilité du rattachement du groupe « encoding proteins » au nom « regulator » est bloquée dès lors que « cot genes encoding proteins » est identifié comme une expression usuelle du domaine.

Etiquetage sémantique

L'étiquetage sémantique consiste simplement à ajouter aux différentes unités textuelles les étiquettes sémantiques correspondant aux types des entités nommées et aux classes sémantiques de l'ontologie. Cet étiquetage permet dans les règles d'extraction d'exprimer des contraintes sur les catégories sémantiques des mots ou des termes plutôt que sur leur forme littérale.

4.2.3 Aller plus loin dans la normalisation

D'autres pistes de normalisation peuvent être envisagées.

La *résolution des anaphores* devrait permettre d'accroître l'homogénéité des formulations et d'augmenter le rappel de l'extraction d'information (elle a une importance moindre pour l'indexation de documents ou de bases documentaires). Diverses méthodes ont été proposées pour la résolution d'anaphores mais on observe que la fréquence, la répartition et le statut (anaphorique *vs.* déictique) des pronoms varient d'un genre de texte à l'autre (voir [Clemente Litrán et al., 2004] dans le domaine de la biologie). Dans un contexte très spécialisé comme le nôtre, il faut déterminer quelles sont les méthodes de résolution les plus fiables et les plus productives. Cette question a fait l'objet d'une étude préliminaire dans le cadre d'un stage de DEA qui a montré l'importance des phénomènes anaphoriques dans les textes de biologie [El Zant, 2002]. Ce travail se poursuit de manière plus approfondie dans la thèse de Davy Weissenbacher.

L'analyse des *nominalisations* est une question classique pour la réduction des paraphrases. Elle est critique pour l'analyse des textes scientifiques et techniques qui emploient beaucoup de tournures nominales. Cela suppose d'acquérir des schémas de sous-catégorisation. Ce travail sur l'analyse des nominalisations n'a pas encore démarré.

Les *attitudes propositionnelles* sont fréquentes dans les résumés d'articles de biologie : les auteurs émettent des hypothèses, prennent des précautions rhétoriques, émettent des doutes, etc. Ceci peut avoir une incidence directe sur la qualité de l'extraction d'information, les assertions n'ayant pas la même valeur que les hypothèses. Il faudrait idéalement repérer les marqueurs de modalités énonciatives et mesurer leur portée mais cela ne paraît pas réaliste. Un voie alternative, plus modeste, consisterait à entraîner un classifieur permettant de départager les vraies modalités des simples tournures rhétoriques et donc les phrases assertives des hypothèses. Un biologiste lisant des résumés d'article sait en effet faire la part des choses mais cette connaissance est difficile à formaliser, car elle repose sur un faisceau d'indices.

4.3 Vers la recherche d'information spécialisée ?

Une troisième méthode d'accès aux contenus des documents est abordée dans le cadre du projet Alvis[19] (2004-2006, projet STREP du 6ème PCRD) qui a débuté en janvier 2004 et doit durer 3 ans. Il s'agit cette fois de recherche d'information.

Ce projet associe différents partenaires européens universitaires et PME. Il vise au développement d'un moteur de recherche sémantique distribué. Les aspects sémantiques doivent assurer une meilleure adéquation des résultats aux requêtes formulées par les utilisateurs. La recherche distribuée, reposant sur la coopération de différents serveurs spécialisés dans des domaines différents, doit permettre au système Alvis global de monter en puissance et de s'enrichir progressivement de nouveaux moteurs spécialisés.

Le LIPN apporte à ce projet sa compétence en TAL et, plus spécifiquement, dans le traitement de documents spécialisés. Ce projet a débuté il y a moins d'un an et il est trop tôt pour présenter des résultats ou tirer un bilan. Je me contente d'expliquer mon intérêt pour ce projet et je décris l'approche proposée en montrant les synergies avec le projet ExtraPloDocs présenté ci-dessus.

Enjeu du TAL pour la recherche d'information spécialisée

Pour apparier de manière satisfaisante les documents et les requêtes, la recherche d'information doit établir une certaine forme de correspondance sémantique entre les deux. Comme la représentation des documents et des requêtes comme des sacs de mots donnent une piètre image de leur sens, de nombreuses expériences ont été menées pour incorporer des techniques de TAL aux méthodes de recherche d'information, depuis l'analyse syntaxique [Smeaton, 1997] jusqu'à la désambiguisation lexicale [Krovetz et Croft, 1992, Voorhees, 1994, Voorhees, 1998, Schütze et Pedersen, 1995, Shütze, 1998] et à la racinisation (*stemming*) [Hull, 1996]. Malgré le nombre et la diversité des tentatives allant dans ce sens, la plupart de leurs auteurs s'accordent pour dire que les résultats sont mitigés et que la contribution du TAL n'apparaît pas vraiment significative (voir [Smeaton, 1997, Sparck Jones, 1997], parmi d'autres). Même s'il est évident que le TAL et les ressources lexicales devrait améliorer la recherche d'information, "the impact of NLP on information retrieval tasks has largely been one of promise rather than substance" [Smeaton, 1997].

Le problème vient de ce que le modèle de langage utilisé en recherche d'information (le modèle vectoriel, notamment) tend à écraser la diversité

[19] http ://cosco.hiit.fi/search/alvis.html

des phénomènes. Une analyse précise (par ex. [Krovetz, 1993, Hull, 1996] pour l'analyse morphologique) montre que l'apport du TAL est réel sur certaines requêtes mais que l'impact global reste faible. De nombreux facteurs brouillent l'analyse des résultats : les unités prises en compte ont des fréquences très diverses, les collections de documents et les questions sont hétérogènes, on présente des scores de recherche d'information moyens qui occultent des écarts de performance, les outils robustes de TAL sont d'insuffisante qualité.

De ce point de vue, le projet Alvis ne comporte pas particulièrement d'innovation [Gaussier et al., 2003]. Le projet offre seulement un cadre intéressant pour tester de manière systématique l'intégration du TAL dans la recherche d'information en mettant l'accent sur la question des corpus spécialisés. Nous allons analyser la contribution des noms d'entités, des termes et mots-clefs, de la lemmatisation, de la racinisation, de la suppression des mots outils sur les performances de la recherche d'information et notamment sur l'indexation des documents et leur appariement aux requêtes. L'objectif est d'étudier la contribution de différentes combinaisons de facteurs. Il s'agit de mesurer l'impact sur les performances de recherche d'information des variations dans la qualité de l'indexation initiale des documents. La contribution du TAL devrait être plus significative pour des tâches de recherche d'information spécialisée. C'est cette hypothèse que je cherche à vérifier dans ce projet Alvis.

Sur un plan plus fondamental, le couplage de l'indexation sémantique des documents et de la gestion distribuée de ces documents pose des questions intéressantes qui vont dans le sens d'une gestion décentralisée du web sémantique plutôt qu'organisée de manière centralisée autour d'un référentiel ontologique stable. Si chaque collection de documents a son propre référentiel d'indexation comment peut-on naviguer au sein de ces différentes collections ? Il faut sans doute que ce référentiel soit public et qu'il soit exploitable par des tiers comme une grille d'interprétation de ladite collection. Ces questions, pour l'heure, sont à peine esquissées.

Approche

L'approche retenue pour tester l'intégration du TAL et de la sémantique dans le processus de la recherche d'information est très proche de celle d'ExtraPloDocs. Les synergies entre les deux projets sont grandes. A dessein, le domaine retenu pour l'expérimentation de l'indexation spécialisée est celui de la biologie.

Il s'agit comme précédemment d'élaborer une chaîne de traitements. Celle-ci est présentée dans la figure 4.8. Elle est développée en parallèle dans dif-

férentes langues (au moins partiellement) : l'anglais, le français et le slovène. Les documents récoltés par le butineur sont analysés, puis indexés avant d'être exploités par le moteur de recherche. L'un des objectifs d'Alvis étant de tester différents niveaux d'intégration du TAL dans la recherche d'information, différentes voies sont explorées. Ces options sont numérotées dans la figure 4.8. L'analyse peut être plus ou moins complète. *A minima*, on se contente d'analyser la structure des documents avant de procéder à l'indexation (option 1). On peut également procéder à une analyse morphologique robuste (option 2) ou à une analyse plus fine (options 3 à 7). On retrouve dans cette chaîne des traitement les modules importants de l'étiquetage de noms d'entités, étiquetage terminologique et sémantique. Comme précédemment, ces modules reposent sur des connaissances (dictionnaires d'entités nommées, terminologies et ontologies) qui doivent être acquises en fonction de la collection de documents à analyser (domaine, genre, degré d'analyse souhaité, etc.).

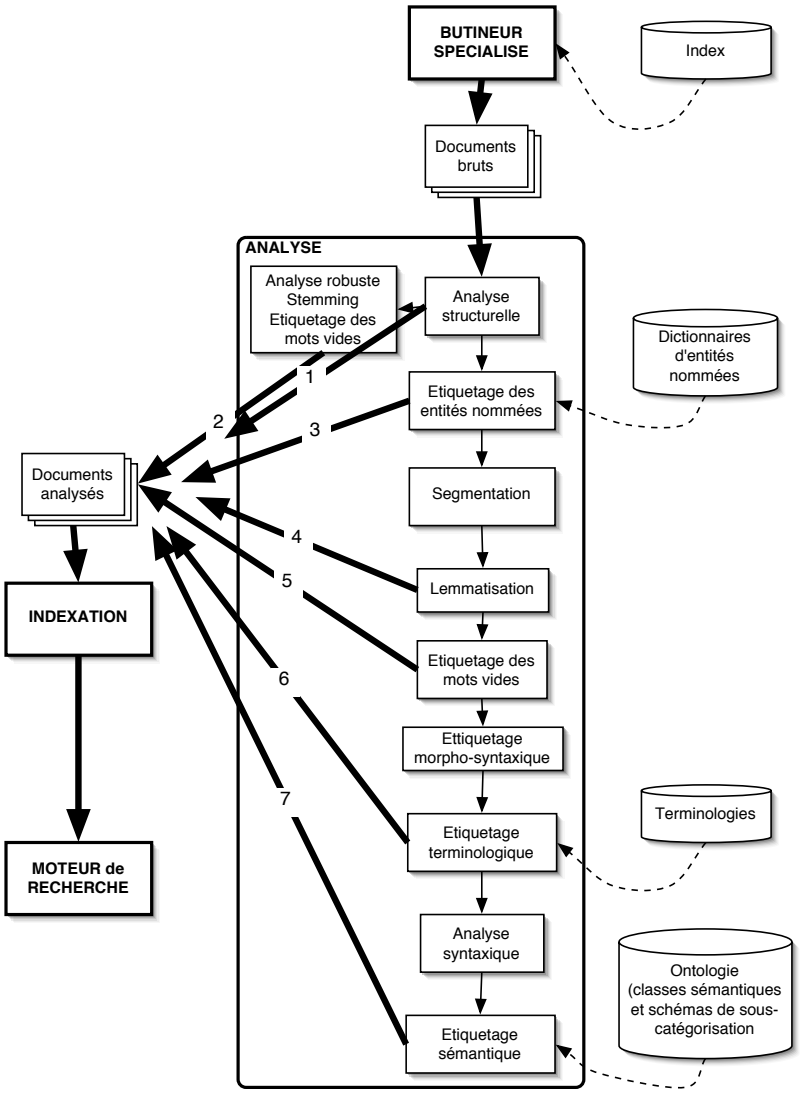

FIG. 4.8 – Chaîne de traitement dans Alvis. Les flèches numérotées 1 à 7 représentent les différents états de texte qui pourront servir de base à l'indexation. Ils correspondent à différents niveaux d'analyse des documents.

4.4 Conclusion

Les trois expériences décrites ici visent toutes à rendre compte du contenu des documents pour faciliter l'exploration de la documentation textuelle dans les domaines scientifiques et techniques, notamment la génomique.

Ces trois expériences sont très diverses, à la fois par leurs ambitions, leur ampleur et leur état d'achèvement. Elles s'inscrivent cependant dans une même perspective.

Sur le plan technologique d'abord, il s'agit à chaque fois de développer une chaîne de traitement couplant acquisition et analyse. Les différentes architectures présentées reposent pour l'essentiel sur les mêmes modules, à quelques spécificités près. Ce rapprochement justifie le projet de développement d'une boîte à outils d'acquisition et d'analyse que j'ai présenté. Ce travail est en cours.

Sur un plan plus fondamental, l'enjeu de ces différentes expériences est de mieux comprendre les point d'achoppement des techniques actuelles pour proposer des voies d'amélioration. Comme l'uniformisation des traitements dans la recherche d'information nuit à l'analyse détaillée des résultats, nous cherchons à décomposer le processus en différents niveaux de traitement et différents degrés d'adaptation au domaine. Dans cette perspective, le projet vise à définir des outils d'analyse de corpus spécialisés qui intègrent des procédures d'adaptation.

Nous avons commencé à travailler sur le repérage des entités nommées, l'analyse terminologique et le module d'extraction pour lesquels l'apport de ressources spécialisées est indispensable. D'autres chantiers restent à ouvrir pour la normalisation des tournures nominalisées, pour la résolution des anaphores et la prise en compte des modalités. L'adaptation de l'analyse syntaxique est plus problématique. Jusqu'à présent, les adaptations qui ont été faites sont ponctuelles et essentiellement manuelles. A court terme, le couplage de l'analyse terminologique et de l'analyse syntaxique devrait réduire l'ambiguïté syntaxique. Ces adaptations se font à la marge, en amont et en aval de l'analyse. A terme, il faudra peut-être reconsidérer l'ensemble de l'analyse syntaxique et concevoir une grammaire et un analyseur réellement adaptables, et de ce fait adaptés aux traitement des corpus spécialisés.

La collaboration avec l'Unité MIG a permis de mettre en oeuvre des techniques d'apprentissage pour certaines ressources (classifieur de fragments, ontologies, règles d'extraction) et d'en envisager d'autres (règles de résolution d'anaphores, classification des attitudes propositionnelles). Cette collaboration dans la durée permet de mieux comprendre la contribution respective des techniques de TAL et de l'apprentissage automatique. Dans un processus en spirale, les premières servent à normaliser les textes et à préparer les données des secondes, et les secondes constituent des connaissances qui sont exploitées à leur tour pour normaliser davantage les textes. Il faudra évidemment évaluer l'apport spécifique des différentes étapes de normalisation dans ce processus : la modularité de la chaîne de traitement proposée devrait le permettre.

Chapitre 5

Bilan et perspectives

L'ensemble de mes travaux portent sur le problème de la compréhension automatique de textes, vue à travers différentes méthodes d'accès au contenu des documents textuels. Comprendre des textes et acquérir des connaissances sont deux activités intimement liées. Acquérir c'est d'abord comprendre et comprendre suppose des connaissances qu'il faut avoir acquises au préalable. Après quelques années de recherche, je n'ai évidemment pas épuisé cette problématique paradoxale mais j'ai essayé de proposer des voies d'approche et des méthodes d'accès au contenu textuel qui combinent utilement acquisition de connaissances et analyse de corpus spécialisés.

Dans les chapitres précédents, j'ai présenté les différents chantiers sur lesquels j'ai travaillé :

- Côté acquisition, j'ai mis l'accent sur les classes sémantiques, les relations entre termes et les règles d'extraction mais j'ai également évoqué d'autres tâches d'acquisition : certaines sont classiques (acquisition de terminologie, de noms d'entités), d'autres plus hasardeuses (apprentissage de stratégies de résolution d'anaphores, acquisition de schémas prédicatifs pour l'interprétation des prédicats nominalisés voire apprentissage de classifieur de modalité).
- En ce qui concerne les outils d'accès au contenu textuel, j'ai montré la diversité des approches possibles, de l'indexation fine des documents, à l'extraction d'information et à la recherche d'information spécialisée tout en soulignant la parenté des traitements sous-jacents : tous les outils d'accès au contenu textuel exploitent des entités nommées et des termes ; la notion de pertinence et l'identification de zones pertinentes est également une constante, même si elle intervient de différentes manières et si elle repose sur des indices plus ou moins variés ; on retrouve aussi, d'une application à l'autre, la notion de classes ou de relations sémantiques sous différentes formes.

Le point commun à tous ces travaux est le choix délibéré de travailler sur des documents spécialisés, c'est-à-dire relatifs à un domaine de connaissance particulier, à une communauté précise et généralement à un type spécifique d'applications. Ce choix répond à une double motivation : il traduit une certaine analyse des besoins en matière de TAL et la conviction que les méthodes mises en oeuvre sont plus fructueuses si les corpus et la tâche sont mieux circonscrits.

En terminant la présentation de chacun des chantiers sur lesquels j'ai travaillé, j'ai tenté un bilan en montrant les limites et les perspectives des travaux effectués. Je me contente donc de synthétiser l'ensemble de ces remarques ici, sans les reprendre en détail.

Construire des outils de TAL adaptables

Les performances des outils de TAL génériques s'écroulent souvent sur les corpus spécialisés et il n'est pas imaginable de redévelopper de nouveaux outils pour chaque application. L'analyse de corpus spécialisés suppose donc de pouvoir paramétrer des outils génériques pour chaque nouvelle application.

Ce paramétrage peut se fait en amont par l'ajout de ressources dédiées à la tâche (lexiques et règles contextuelles), ou en aval, lors d'une étape de correction ou de validation, mais c'est souvent l'architecture elle-même de la chaîne de traitement qui doit être adaptée. Selon les applications visées et les corpus à traiter, la qualité relative et l'ordre d'application des traitements varient en effet.

Dans cette perspective, il est essentiel de construire des outils de TAL modulaires et adaptables, un chantier sur lequel nous avons commencé à travailler au LIPN et qu'il faut poursuivre. La modularisation consiste à découpler chaque fonctionnalité de ses voisines et à définir des interfaces claires. Elle doit permettre de combiner librement les modules élémentaires pour élaborer de nouvelles applications. Les développements commencés dans cette perspective ne visent pas nécessairement à implanter des méthodes originales mais à mettre au point une « boîte à outils » composée de modules Open-Source[1], faciles à intégrer, réutiliser et adapter. Les outils existant pour le traitement automatique du français étant souvent des outils propriétaires, difficilement compatibles entre eux et insuffisamment modulaires, les tentatives d'intégration sont difficiles.

La mise au point de nouvelles architectures nécessite de pouvoir tester et expérimenter facilement différentes solutions. La conception de méthodes, de métriques et d'outils qui permettent d'explorer « en laboratoire » avant de

[1]La licence OpenSource présente de bonnes garanties concernant l'utilisation effective, l'évolution et la pérennité du logiciel. C'est l'optique adoptée dans le projet Alvis.

tester en grandeur nature est probablement un des défis majeurs auxquels le traitement de corpus est confronté. J'ai esquissé quelques pistes mais la question demanderun investissement important, prix à payer pour sortir du « bricolage ».

Penser la coopération

Cela a été souligné à plusieurs reprises, l'acquisition n'est pas entièrement automatisable. Il ne s'agit pas de découvrir des connaissances dans les textes (termes, relations, classes sémantiques ou règles d'extractions) mais de les construire en s'appuyant sur le texte pour défricher et amorcer le travail ainsi que pour garantir la couverture, l'ancrage linguistique et la représentativité des connaissances élaborées.

Cela signifie que les procédures d'acquisition doivent intégrer la dimension coopérative. Celle-ci peut prendre différentes formes : dans l'apprentissage supervisé, elle intervient en amont par la construction d'exemples d'apprentissages ; dans les outils de terminologie comme SynoTerm ou IndDoc, elle intervient au contraire dans une validation a posteriori ; dans l'exploitation des graphes de similarités, elle a un rôle central. Dans tous les cas, elle doit être outillée (outils d'annotation, interface de validation, dispositifs de visualisation et de manipulation, etc.) et les outils doivent être élaborés avec soin si l'on souhaite qu'ils soient utilisés ! Les outils d'acquisition existants intègrent généralement cette dimension mais de manière diverse. La communauté du TAL n'a pas mené de réflexion de fond sur la place, le rôle et les modalités de cette coopération et ne pourra guère avancer sans interagir avec l'ingénierie des connaissances et l'ergonomie.

De la clarification du rôle de la coopération dépend la possibilité d'évaluer les outils d'acquisition de connaissances. Comme dans le cas de la construction d'index de documents, il faut pouvoir évaluer à la fois l'apport des traitements automatiques mais aussi la contribution de la coopération humain-machine en tant que telle : les outils d'aide accélèrent-ils le processus d'acquisition ? en quoi améliorent-ils (ou dégradent-ils !) la qualité globale des résultats ?

Concevoir de nouveaux modes d'exploration des documents

Les outils d'accès au contenu des documents sont encore très marqués par la tradition documentaire (index de fin de livre, indexation de bases documentaires, notamment). D'autres communautés réfléchissent à des usages plus innovants : au sein des sciences de l'information, autour du livre électronique et de l'hypertexte dynamique. Il me paraît intéressant d'explorer

certaines de ces voies, à la fois parce que le TAL peut avoir son rôle à jouer dans ces applications émergentes et parce que chaque nouvelle application permet d'éprouver et de refonder les méthodes de TAL proposées.

A court terme, je compte explorer la problématique de la *navigation dans les documents*. Avec T. Aït El Mekki, nous avons jusqu'à présent mis l'accent sur la construction des index de documents et partiellement négligé ce qui concerne l'exploitation de ces index par le lecteur. Le projet de collaboration franco-québécoise avec l'Ecole de bibliothéconomie et des sciences de l'information de l'Université de Montréal, s'il voit le jour, devrait apporter un éclairage intéressant sur ces questions, par sa dimension pluridisciplinaire.

Une piste plus exploratoire concerne le projet d'un *web sémantique*, où les travaux sur les index et la recherche d'information spécialisée se prolongent naturellement. L'objectif du web sémantique, concernant les documents textuels, est de définir un système formalisé de connaissances qui permette à une machine d'accéder, de manipuler et d'exploiter les documents textuels. Cela suppose d'associer des éléments de connaissances aux documents (on parle généralement de méta données) et de raisonner sur ces derniers. L'approche dominante, me semble-t-il, procède par annotation de texte, au sens où l'indexation contrôlée projette des connaissances connues au préalable sur les textes. Une fois le modèle fixé, l'ensemble des documents est annoté conformément à ce modèle et les services web peuvent communiquer. Une approche alternative procède en sens contraire en construisant des cartographies de texte (sortes de modèles régionaux comme par exemple les index induits à partir de l'analyse du document). La cohabitation de différents modèles régionaux pose la question leur interopérabilité. Il faut inventer des méthodes d'ajustement qui devraient s'avérer à terme plus robustes que l'approche fixiste, si le champ de connaissances à couvrir est large et hétérogène.

Comprendre ce qu'on fait

Il paraît enfin important d'analyser les fondements sémantiques de ces méthodes d'accès au contenu textuel. Les évolutions récentes ont été guidées le plus souvent par un esprit de pragmatisme. Elles reposent sur différents facteurs : l'essor des techniques d'apprentissage, la prise en compte de la structure du document, la mise à disposition de ressources lexicales, le besoin de traiter de gros volumes de documents.

L'analyse qui en résulte est complexe et hétérogène. Si on veut sortir du tâtonnement, il paraît essentiel d'en faire la synthèse et de comprendre comment s'articulent les différentes sémantiques à l'œuvre. Il faut prendre en compte les deux niveaux de l'acquisition et de l'analyse qui reposent en effet sur deux approches sémantiques différentes.

L'analyse est éclatée. Elle procède localement autour d'îlots de texte mais elle est aussi éclairée par l'analyse globale qui est faite lors de l'acquisition, analyse du texte pris dans son intégralité ou même d'un corpus plus large. L'acquisition construit ainsi des classes sémantiques qui servent à étiqueter sémantiquement les portions de texte analysés et à lever les ambiguïtés locales. On peut apprendre à partir d'un ensemble d'exemples étiquetés des règles d'extraction qui servent à interpréter les fragments de textes. On exploite de gros volumes de texte pour calculer des mesures de fréquence ou de pertinence (associées aux mots ou à toute autre unité textuelle) qu'on exploite ensuite pour sélectionner les zones de textes à analyser.

L'analyse repose essentiellement sur une sémantique référentielle : elle s'appuie sur les entités référentielles les plus faciles à repérer et à interpréter. On ne peut cependant réduire la sémantique des méthodes d'accès au contenu textuel à cette dimension-là, car l'acquisition procède au contraire par exploration de corpus. Elle repose davantage sur une conception différentielle où le sens des unités se définit non pas par rapport à l'extralinguistique (la référence) mais de manière interne, par un jeu de contrastes avec les unités voisines. Il s'agit plus largement d'une sémantique textuelle : dans la mesure elle met en jeu une analyse distributionnelle et où elle exploite tout un faisceau d'indices discursifs (balisage structurel, marqueurs d'emphase, fréquences des unités, etc.), c'est le texte dans son intégralité qui détermine le sens des unités. Dans certains cas, il est vrai, l'analyse sémantique est guidée par des ressources extérieures préexistantes (terminologies, thesaurus ou ontologies) qui peuvent relever d'une vision plus référentielle que textuelle du sens. C'est le cas de Wordnet, sans doute la ressource la plus fréquemment exploitée pour l'analyse de l'anglais. Pour être exploitables, cependant, ces ressources doivent être sélectionnées, adaptées et mises à jour à partir de corpus quand elles ne sont pas elles-mêmes directement élaborées à partir de corpus.

Cette réflexion a été initiée dans [Nazarenko, 2005] mais de nombreuses questions restent en suspens. Celle de la prise en compte des genres textuels est sans doute la principale. Si le genre est, comme l'avance [Rastier, 2001], une instance de normalisation linguistique, il devrait être pris en compte comme donnée de l'acquisition et paramètre de l'analyse, ce qui n'est pas le cas dans les travaux décrits ici. Les conclusions de l'Action Spécifique ASSTICCOT soulève ces questions. Je souhaite que le travail sur l'indexation fine des documents médicaux qui devrait commencer en 2005[2] permette de les aborder concrètement.

[2]Dans le cadre du plan Pluriformations que j'ai déposé pour le plan quadriennal 2005-08.

En guise de conclusion

Le TAL m'a amenée de la linguistique à l'informatique. Il m'a fait travailler avec des spécialistes de l'apprentissage et m'a fait rencontrer médecins et biologistes.

Ce mémoire a montré combien de telles collaborations pluridisciplinaires peuvent être fructueuses, pour la définition des méthodes et pour celle des objectifs. La confrontation avec les techniques d'apprentissage automatique a permis de renouveler les méthodes tout en mettant en évidence le rôle spécifique du TAL. Le dialogue avec des gens connaissant bien les domaines d'application aide à définir les besoins auxquels le TAL doit répondre. Il contrebalance de ce fait les protocoles d'évaluation usuels, qui mettent souvent l'accent sur les technologies pour encourager leur développement.

J'ai souligné les limites du « tout automatique » à plusieurs reprises. Il faut donc à la fois cerner au mieux le champ de l'automatisable et inventer des processus coopératifs, sans perdre l'utilisateur de vue. Là encore le dialogue pluridisciplinaire devrait être profitable : il faut, me semble-t-il, se tourner cette fois vers l'ergonomie et les sciences de l'information.

Annexe A

Présentation du formalisme des Graphes Conceptuels

Le formalisme des Graphes Conceptuels a été défini par J. Sowa [Sowa, 1984, Sowa, 1992] dans la lignée des travaux sur les réseaux sémantiques (voir [Chein et Mugnier, 1991] pour une analyse plus détaillée).

Un graphe conceptuel est un graphe orienté et étiqueté, comportant deux types de nœuds (les concepts et les relations) reliés entre eux par des arcs orientés. La figure A.1 présente deux exemples. Les concepts (représentés dans des rectangles) représentent d'après J. Sowa les objets (abstraits ou concrets) que l'on peut appréhender par la pensée. Un concept résulte de l'association entre un type de concept (VOITURE, AMORTISSEUR) et un référent qui désigne une entité particulière du monde (#12) ou une entité générique (*). Il peut également s'agir d'un ensemble de référents ({*}). Les relations qui relient ces concepts entre eux (notées dans des ovales) peuvent être de nature variée : 'partie' et 'caractéristique', dans les exemples de la figure A.1.

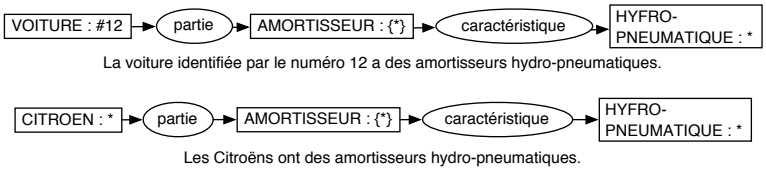

FIG. A.1 – **Exemples de graphes conceptuels**

On peut définir un langage de graphes pour un domaine de connaissance particulier. Ce langage se définit par un support :

– Une hiérarchie de types de concepts organisée autour de la relation de spécificité. Un type de concept t_j est dominé par un (ou plusieurs)

type(s) t_i s'il est plus spécifique que t_i. Dans l'exemple de la figure A.1, on pourrait poser la relation CITROËN<VOITURE pour indiquer le fait que CITROËN est un type de concept plus spécifique que VOITURE et donc que les Citroëns forment un sous-ensemble des voitures. Le type ENTITY est le type le plus général.

- Un ensemble de relations, qui peut également être organisé en hiérarchie.
- Un ensemble de marqueurs ou référents qui représentent les entités du domaine que l'on cherche à modéliser.
- Une relation de conformité qui associe un marqueur au(x) type(s) de concept dont il relève.
- La signature des types de concepts et des relations, ensemble des graphes qui expriment les contraintes liées à chaque type ou relation. On peut ainsi exprimer le fait que dans le domaine considéré, un amortisseur est toujours une partie d'une voiture. Pour une relation, la signature fixe le nombre et le type de ses arguments. La relation 'partie' est ainsi une relation binaire qui relie deux entités concrétes.

Les graphes de signature constituent des graphes élémentaires à partir desquels ont peut construire des graphes plus complexes. J. Sowa introduit un ensemble d'opérations de combinaison de graphes, qui permettent de dériver un graphe à partir d'un ou de plusieurs graphes élémentaires [Sowa, 1984, Catach et Fargues, 1985].

Annexe B

Liste des contrats

Je présente ici la liste des contrats auxquels j'ai participé. Dans tous sauf le premier, j'ai assuré la responsabilité scientifique du contrat au sein du LIPN.

Contrats terminés

Action de Recherche Concertée (ARC 4, 1995-97) de l'AUPELF-UREF pour l'évaluation des systèmes de compréhension de textes, sous la direction de Paul Sabatier (CNRS, Laboratoire d'Informatique de Marseille). Cette étude avait pour but d'étudier les conditions d'une évaluation de tels systèmes et de jeter les bases d'un protocole d'évaluation. La contribution du LIPN a porté sur les aspects proprement sémantiques et pragmatiques de la compréhension de textes

Contrat de d'Études de Recherche et Développement entre la DER-EDF et l'Université Paris 13 pour l'aide à la structuration de terminologies. Ce contrat de 40 KF a duré un an (1998). Le stage de DEA de Thierry Hamon a été effectué dans ce cadre. Cette collaboration a permis de conduire des expériences intéressantes concernant l'étude de la synonymie entre termes.

Projet RNRT ASTUXE, associant Spartacom (PME, éditeur de logiciel, coordinateur), le Laboratoire Central de Recherche de Thomson-CSF, le LRI et le LIPN. Ce projet précompétitif vise à développer un module d'analyse automatique des mails de demande d'intervention adressé à un centre de support technique. Le financement prévu pour le LIPN s'élevait à 1 125 KF pour 27 mois (2000-02). Ce projet a connu des difficultés suite à la défection et au remplacement du coordinateur et au changement d'objectifs qui en a

découlé. Il a été interrompu au bout de 15 mois. En dépit de son relatif échec, ce projet a permis de jeter les bases d'une collaboration durable avec le LRI. Il a été relancé sur des nouvelles bases dans le cadre du projet ExtraPloDocs.

Projet de bioinformatique inter-EPST CADERIGE 1 et 2, associant différents partenaires dont le LIPN. Ce projet a pour objectif de concevoir des bases de connaissance d'interactions géniques grâce à l'automatisation de l'extraction d'information dans la bibliographie de MedLine (novembre 2001 à octobre 2003). Il a permis de jeter les bases de mon travail sur la biologie.

Contrats en cours

Projet CEDERILIC (2003-04), associant l'ERSS, l'unité STIM de l'AP-HP, FranceTélécom et le LIPN pour la construction d'index pour des ouvrages au format électronique et papier. Ce projet, modesterment financé par France-Télécom, a résulté en la publication d'un ouvrage scientifique doté d'un riche index. Cet ouvrage à été publié à la fois sur CD-rom et sous forme papier [Charlet et al., 2004].

Projet RNTL ExtaPloDocs (30 mois, 2002-2005), associant le LIPN qui en assure la coordination, l'Unité Mathématique Informatique et Génome de l'INRA et la société Isoft (PME, éditeur de logiciel de datamining). Ce projet exploratoire vise au développement d'un outil d'extraction d'information destiné à faciliter les recherches bibliographiques effectuées par les biologistes en génomique. Le financement pour le LIPN s'élève à 170 Keuros.

Projet STREP Alvis (6ème PCRD, 36 mois, 2004-2006), associant différents partenaires européens académiques et PME. Ce projet est en cours de négociation après des premières évaluations positives. Il vise au développement d'un moteur de recherche sémantique et distribué. Les aspects sémantiques assurent une meilleure adéquation des résultats aux requêtes formulées par les utilisateurs. La recherche distribuée repose sur la coopération de différents serveurs spécialisés dans des domaines différents. Le financement pour le LIPN s'élève à 210 Keuros.

Plan Pluriformations pour le plan quadriennal 2005-2008 en collaboration avec 2 autres laboratoires de l'Université Paris 13 (le Lim&Bio et le LLI). Le LIPN est coordinateur. Ce projet a pour but de développer des outils facilitant l'accès à l'information médicale. Ce projet a été accepté. Il doit débuter en janvier 2005. Le financement pour le LIPN doit s'élèver à 70

Keuros sur toute la durée du projet.

Contrats en cours d'évaluation

Coopération franco-québécoise entre le LIPN et l'Ecole de bibliothéconomie et des sciences de l'information de l'Université de Montréal (2005-2006). Ce projet vise à faire tester et évaluer le prototype d'indexation fine de documents développé au LIPN par des professionnels des sciences de l'information. Il devrait déboucher sur des propositions d'amélioration. Ce projet a été déposé en juin 2004. Il est en cours d'évaluation.

Bourse Cifre avec la société Sinequa, pour la thèse de Frederik Cailliau, dans le prolongement de son stage de DEA. Le dossier est en cours d'évaluation.

Bourse Cifre avec la société ELDA pour la thèse d'Olivier Hamon, dans le prolongement de son stage de DEA. Le dossier est en cours de constitution.

Bibliographie

[Ait El Mekki et Nazarenko, 2001] Ait El Mekki, T. et Nazarenko, A. (2001). Quel index pour le document électronique ? In Mojahid, M. et Virbel, J., editors, *Actes du 4ème Colloque International sur le Document Electronique (CIDE'01)*, pages 147–161, Paris. Europia.

[Alphonse et al., 2004] Alphonse, E., Aubin, S., Bessières, P., Bisson, G., Hamon, T., Lagarrigue, S., Manine, A. N. A.-P., Nédellec, C., Vetah, M. O. A., Poibeau, T., et Weissenbacher, D. (2004). Event-based Information Extraction for the biomedical domain : the Caderige project. In Collier, N., Rush, P., et Nazarenko, A., editors, *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications (COLING'2004)*, pages 43–49, Geneva.

[Amar, 2000] Amar, M. (2000). *Les fondements théoriques de l'indexation : une approche linguistique*. ADBS Editions, Paris.

[Anick, 2001] Anick, P. G. (2001). *The automatic construction of faceted terminological feedback for interactive document retrieval*. In Bourigault, D., Jacquemin, C., et L'Homme, M.-C., editors, *Recent Advances in Computational Terminology*, volume 2 of *Natural Langage Processing*, pages 29–52. John Benjamins, Amsterdam.

[Assadi, 1997] Assadi, H. (1997). Knowledge acquisition from Texts : Using an automatic Clustering Method Based on Noun-Modifier Relationship. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistic - Student Session*, Madrid, Spain.

[Aussenac-Gilles et al., 2004] Aussenac-Gilles, N., Biébow, B., et Szulman, S. (2004). *Modélisation du domaine par une méthode fondée sur l'analyse de corpus*. In Teulier, R., Charlet, J., et Tchounikine, P., editors, *Ingénierie des Connaisances*. Eyrolles, Paris.

[Aït El Mekki, 2004] Aït El Mekki, T. (2004). *Construction semi-automatique d'index de fin de livre*. Thèse d'Informatique . Université de Paris XIII.

[Aït El Mekki et Nazarenko, 2004] Aït El Mekki, T. et Nazarenko, A. (2004). L'index de fin de livre, une forme de résumé indicatif ? *Traitement Automatique des Langues*, 45(1).

[Baeza-Yates et Ribeiro-Neto, 1999] Baeza-Yates, R. et Ribeiro-Neto, B., editors (1999). *Modern Information Retrieval*. Addison-Wesley.

[Barwise et Perry, 1983] Barwise, J. et Perry, J. (1983). *Situations and Attitudes*. MIT Press, Cambridge MA.

[Basili et al., 1998] Basili, R., Pazienza, M.-T., Stevenson, M., Velardi, P., Vindigni, M., et Wilks, Y. (1998). An Empirical Approach to lexical Tuning. In Velardi, P., editor, *Proceedings of the Workshop on Adaptating Lexical and Corpus Ressources to Sublanguages and Applications (First International Conference on Language Resources and Evaluation)*, Grenade.

[Basili et al., 1997] Basili, R., Rocca, M. D., et Pazienza, M. (1997). Contextual Word Sense Tunig and Disambiguation. *Applied Artificial Intelligence*, 11 :235–262.

[Benjamins et Fensel, 1998] Benjamins, R. et Fensel, D. (1998). The Ontological Engineering Initiative-KA. In Guarino, N., editor, *Proceedings of the 1st International Conference on Formal Ontologies in Information Systems, FOIS'98*, pages 287–301, Trento, Italy. IOS Press.

[Bensch et Savitch, 1995] Bensch, P. A. et Savitch, W. J. (1995). An occurrence-based model of word categorization. *Annals of Mathematics and Artificial Intelligence*, 14 :1–16.

[Berland et Charniak, 1999] Berland, M. et Charniak, E. (1999). Finding Parts in Very Large Corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*.

[Berroyer, 2004] Berroyer, J.-F. (2004). *Elaboration d'un système d'entités nommées*. Mémoire de DEA d'intelligence Artificielle et d'Optimisation Combinatoire. Université de Paris XIII.

[Bessières et al., 2001] Bessières, P., Nazarenko, A., et Nédellec, C. (2001). Apport de l'apprentissage à l'extraction d'information : le problème de l'identification d'interactions géniques. In *Actes du Colloque International sur le Document Electronique – Méthodes, Démarches et Techniques Cognitives (CIDE'2001)*.

[Bisson et al., 2000] Bisson, G., Nédellec, C., et Canamero, L. (2000). Designing clustering methods for ontology building - The Mo'K workbench. In Staab, S., Maedche, A., et Nedellec, C., editors, *Proceedings of the Ontology Learning ECAI-2000 Workshop*, Berlin.

[Blache et al., 2000] Blache, P., Guizol, J., Lévy, F., Nazarenko, A., N'Guéma, S., Rolbert, M., Pasero, R., et Sabatier, P. (2000). *Evaluer des systèmes de compréhension de textes.* In et al, C., editor, *Ressources et Evaluation en Ingénierie de la langue*, pages 265–275. Duculot-De Boeck-Université.

[Blashke et al., 1999] Blashke, C., Ouzounis, M., et Valencia, A. (1999). Automatic Extraction of biological information from scientific text : protein-protein interactions. In *Proceedings of International Symposium on Molecular Biology, (ISMB'99).*

[Borillo, 1996] Borillo, A. (1996). Exploration automatisée de textes de spécialité : repérage et identification de la relation lexicale d'hyperonymie. *LINX*, 34-35 :113–124.

[Bouaud et al., 1995] Bouaud, J., Bachimont, B., Charlet, J., et Zweigenbaum, P. (1995). Methodological principles for structuring an "ontology". In *Proceedings of the Workshop on "Basic Ontological Issues in Knowledge Sharing" (IJCAI'95)*, Montréal.

[Bouaud et al., 2000] Bouaud, J., Habert, B., Nazarenko, A., et Zweigenbaum, P. (2000). *Regroupements issus de dépendances syntaxiques sur un corpus de spécialité : catégorisation et confrontation à deux conceptualisations du domaine.* In Charlet, J., Zacklad, M., Kassel, G., et Bourigault, D., editors, *Ingénierie des connaissances : évolutions récentes et nouveaux défis*, chapter 17, pages 275–290. Eyrolles, Paris.

[Bourigault, 1994] Bourigault, D. (1994). *LEXTER un Logiciel d'EXtraction de TERminologie. Application à l'extraction des connaissances à partir de textes.* Thèse en mathématiques, informatique appliquée aux sciences de l'homme, École des Hautes Études en Sciences Sociales, Paris, France.

[Bourigault, 2002] Bourigault, D. (2002). Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In *Actes de la 9ème conférence annuelle sur le Traitement Automatique des Langues (TALN 2002)*, pages 75–84, Nancy.

[Bourigault et Charlet, 1999] Bourigault, D. et Charlet, J. (1999). Construction d'un index thématique de l'Ingénierie des connaissances. In *Actes de la conférence Ingénierie des Connaissances.*

[Bourigault et Jacquemin, 2000] Bourigault, D. et Jacquemin, C. (2000). *Construction de ressources terminologiques.* In Pierrel, J.-M., editor, *Industrie des langues*, pages 215–233. Hermès, Paris.

[Bourigault et Slodzian, 2000] Bourigault, D. et Slodzian, M. (2000). Pour une terminologie textuelle. *Terminologies Nouvelles*, (19) :29–32.

[Brachman, 1979] Brachman, R. J. (1979). *On the epistemological status of semantic nets.* In Findler., N. V., editor, *Associative Networks : Representation and Use of Knowledge by Computers*, pages 3–50. Academic Press, New York.

[Bérard-Dugourd et al., 1989] Bérard-Dugourd, A., Fargues, J., Landau, M.-C., et Rogala, J.-P. (1989). *Un système d'analyse de texte et de question/réponse basé sur les graphes conceptuels.* In Degoulet, P., editor, *Informatique et Gestion des Unités de Soins*, Coll. Informatique et Santé 1, pages 223–233. Springer-Verlag, Paris. Comptes rendus du Colloque AIM-IF, Paris Juin 1989.

[Bès et Guillotin, 1992] Bès, G. G. et Guillotin, T., editors (1992). *A Natural Language and Graphics Interface (Results and Perspective from the ACORD Project).* Number 393 in Research Report ESPRIT. Springer-Verlag.

[Cabré, 1998] Cabré, M. T. (1998). *Terminology - Theory, methods and applications.* Terminology and Lexicography - Research and Practice. John Benjamins.

[Catach et Fargues, 1985] Catach, L. et Fargues, J. (1985). Déduction et opérations pour le modèle des graphes conceptuels. Etude F087, Centre Scientifique IBM France, Paris.

[Cerbah, 2000] Cerbah, F. (2000). Exogeneous and endogeneous approaches to semantic categorization of unknown technical terms. In *Proceedings of the 17th conference on Computational linguistics*, Saarbrücken, Germany.

[Chang et al., 2002] Chang, J., Schüter, H., et Altman, R. (2002). Creating an online dictionary of abreviations from Medline. *Journal of the American Medical Informatics Association*, pages 612–620.

[Charlet et al., 2004] Charlet, J., Aït El Mekki, T., Bourigault, D., Nazarenko, A., Teulier, R., et Toledano, B. (2004). CEDERILIC : constitution d'un livre et d'un index numériques. In *Actes du Colloque International sur le Document Electronique.*

[Charlet et al., 1996] Charlet, J., Bachumont, B., Bouaud, J., et Zweigenbaum, P. (1996). *Ontologie et réutilisabilité : expérience et discussion.* In Aussenac, N., Laublet, P., et Reynaud, C., editors, *Acquisition et Ingénierie des Connaissances*, pages 69–88. Cépaduès-Editions, Toulouse, France.

[Chaudiron, 2004] Chaudiron, S., editor (2004). *L'évaluation des systèmes de traitement de l'information.* Lavoisier.

[Chein et Mugnier, 1991] Chein, M. et Mugnier, M.-L. (1991). Graphes conceptuels de Sowa : *Notions fondamentales (1).* R.R. 93, CRIM, Montpellier.

[Clemente Litrán et al., 2004] Clemente Litrán, J. C., Satou, K., et Torisawa, K. (2004). Improving the Identification of Non-Anaphoric *it* using Support Vector Machines. In Collier, N., Rush, P., et Nazarenko, A., editors, *Proceedings on International Joint Workshop on Natural Language Processing in Biomedecine and its Applications (BioNLP&LNPBA)*, pages 58–61, Geneva, Switzland. International Conference on Computational Linguistics (COLING'04).

[Cohen, 2001] Cohen, D. (2001). *Analyse linguistique et terminologique de corpus techniques pour l'acquisition de connaissances.* Mémoire de Fin d'étude d'ingénieur. CNAM.

[Condamines, 1997] Condamines, A. (1997). *Langue spécialisée ou discours spécialisé ?* In Lapierre, L., Oore, I., et Runte, H., editors, *Mélanges de linguistique offerts à Rostislav Kocourek*, pages 171–184. Les presses d'Alfa, Université de Dalhousie.

[Condamines, 2003] Condamines, A. (2003). Sémantique et corpus spécialisés : constitution de bases de connaissances terminologiques. Carnets de grammaire 13, Equipe de Recherche en Syntaxe et Sémantique, UMR 5610 – CNRS & Université de Toulouse-Le Miral, Toulouse, France.

[Cruse, 1986] Cruse, D. A. (1986). *Lexical Semantics.* Cambridge University Press.

[Côté et al., 1993] Côté, R. A., Rothwell, D. J., Palotay, J. L., Beckett, R. S., et Brochu, L., editors (1993). *The Systematised Nomenclature of Human and Veterinary Medicine : SNOMED International.* College of American Pathologists, Northfield.

[Dachelet, 1994] Dachelet, R. (1994). *Sur la notion de sous-langage.* Thèse de doctorat en sciences du langage, Université Paris VIII. Directeur : N. Ruwet.

[Dagan et al., 1999] Dagan, I., Marcus, S., et Markovitch, S. (1999). Similarity-based models of word coocurrence probabilities. *Machine Learning*, 34(1-3) :43–69.

[Dague et al., 2004] Dague, P., Kayser, D., Lévy, F., et Nazarenko, A. (2004). Raisonnement Causal. *Intellectica*, 1(38).

[El Zant, 2002] El Zant, M. (2002). *Normalisation de textes techniques pour l'extraction d'information.* Mémoire de DEA d'intelligence Artificielle et d'Optimisation Combinatoire. Université de Paris XIII.

[Faure et Nédellec, 1999] Faure, D. et Nédellec, C. (1999). Knowledge acquisition of predicate argument structures from technical texts using Machine Learning : the system ASIUM. In *Proceedings of the 11th International Conference on Knowledge Engineering and Knowledge Management.*

[Freitag, 1998] Freitag, D. (1998). Toward General-Purpose Learning for Information Extraction. In *Proceedings of the Seventeenth International Conference on Computational Linguistics (COLING-ACL-98).*

[Fukuda et al., 1998] Fukuda, K.-I., Tsunoda, T., Tamura, A., et Takagi, T. (1998). Toward Information Extraction : Identifying protein names from biological papers. In *Proceedings of the Pacific Symposium on biocomputing (PSB'1998).*

[Gangemi et al., 2001] Gangemi, A., Guarino, N., et Oltramri, A. (2001). Conceptual Analysis of Lexical Taxonomies : The Case of WordNet Top-Level. In *Proceedings of FOIS'2001*, Ogunguit, Maine.

[Gaussier et al., 2003] Gaussier, E., Jacquemin, C., et Zweigenbaum, P. (2003). *Traitement automatique des langues et recherche d'information.* In et Marie-Hélène Stefanini, E. G., editor, *Assistance Intelligente à la recherche d'information.* Lavoisier.

[Gayral, 1998] Gayral, F. (1998). Créativité du sens en contexte et hypothèse de compositionnalité. *Traitement Automatique des Langues*, 39(1).

[Gayral et al., 1994] Gayral, F., Grandemanche, P., Kayser, D., et Levy, F. (1994). Interprétation des constats d'acccidents : représenter le réel et le potentiel. *traitement automatique des langues (t.a.l.)*, 35(1) :65–82.

[Grefenstette, 1994a] Grefenstette, G. (1994a). Corpus-derived first, second and third order affinities. In *EURALEX*, Amsterdam.

[Grefenstette, 1994b] Grefenstette, G. (1994b). *Explorations in Automatic Thesaurus Discover.* Kluwer Academic Publisher.

[Grishman, 1997] Grishman, R. (1997). *Information Extraction : Techniques and Challenges.* In Pazienza, M. T., editor, *Information Extraction : a Multidisciplinary Approach to an Emerging Information Technology*, pages 10–27. Springer, Berlin.

[Gruber, 1993] Gruber, T. R. (1993). A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2) :199–222.

[Guarino, 1995] Guarino, N. (1995). Formal Ontology, Conceptual Analysis and Knowledge Representation. *International Journal of Human-Computer Studies*, 43 :625–640.

[Habert et Fabre, 1999] Habert, B. et Fabre, C. (1999). Elementary dependency trees for indentifying corpus-specific semantic classes. *Computers and the Humanities*, 33(3) :207–219.

[Habert et al., 1996] Habert, B., Naulleau, E., et Nazarenko, A. (1996). Symbolic word clustering for medium-size corpora. In *Proceedings of the 16th*

International Conference on Computational Linguistics, volume 1, pages 490–495, Copenhagen, Denmark.

[Habert et Nazarenko, 1996] Habert, B. et Nazarenko, A. (1996). La syntaxe comme marche-pied de l'acquisition des connaissances : bilan critique d'une expérience. In *Journées sur l'acquisition des connaissances*, pages 137–142, Sète. AFIA.

[Habert et al., 1997] Habert, B., Nazarenko, A., et Salem, A. (1997). *Les linguistiques de corpus*. U Linguistique. Armand Colin/Masson, Paris.

[Habert et Zweigenbaum, 2002] Habert, B. et Zweigenbaum, P. (2002). *Contextual Acquisition of Information Categories : what has been done and what can be done automatically ?* In Nevin, B. E. et Johnson, S. B., editors, *The legacy of Zellig Harris : language and Information into the 21st century*. John Benjamins, Amsterdam.

[Hamon, 2000] Hamon, T. (2000). *Variation sémantique en corpus spécialisés : Acquisition de relations de synonymie à partir de ressources lexicales*. Thèse d'Informatique . Université de Paris XIII.

[Hamon et Nazarenko, 2001] Hamon, T. et Nazarenko, A. (2001). *Detection of synonymy links between terms : Experiments and results*. In Bourigault, D., Jacquemin, C., et L'Homme, M., editors, *Recent Advances in Computational Terminology*, volume 2 of *Natural Langage Processing*, pages 185–208. John Benjamins, Amsterdam.

[Harris et al., 1989] Harris, Z., Gottfried, M., Ryckman, T., Mattick, J. P., Daladier, A., Harris, T., et Harris, S. (1989). *The Form of Information in Science, Analysis of Immunology Sublanguage*. Kluwer Academic Publisher.

[Harris, 1991] Harris, Z. S. (1991). *A theory of language and information. A mathematical approach*. Oxford University Press, Oxford.

[Hatzivassiloglou et McKeown, 1993] Hatzivassiloglou, V. et McKeown, K. R. (1993). Towards the automatic identification of adjectival scales : clustering adjectives according to meaning. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, USA.

[Hearst, 1992] Hearst, M. A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the 15th International conference on Computational Linguistics*, volume 2, pages 539–545, Nantes.

[Hearst, 1999] Hearst, M. A. (1999). *User Interfaces and Visualization*. In Baeza-Yates, R. et Ribeiro-Neto, B., editors, *Modern Information Retrieval*, chapter 10. Addison-Wesley, Wokingham, UK.

[Herzog et Rollinger, 1991] Herzog, O. et Rollinger, C.-R., editors (1991). *Text Understanding in LILOG.* Lecture Notes in Artificial Intelligence. Springer-Verlag, Heidelberg.

[Hindle, 1990] Hindle, D. (1990). Noun Classification from Predicate-Argument Structures. In *Proceedings of the Association for Computational Linguistics*, pages 268–275.

[Hishiki et al., 1998] Hishiki, T., Collier, N., Nabata, C., Ohta, T., Ogata, N., Sekimizu, T., Steiner, R., Park, H., et Tsujii, J. (1998). Developping NLP tools for Genome Informatics : An Information Extraction Perspective. *Genome Informatics*, 9 :81–90.

[Hull, 1996] Hull, D. A. (1996). Stemming Algorithms – A Case Study for Detailed Evaluation. *Journal of the American Society of Information Science*, 47(1) :70–84.

[Jacquemin, 1997] Jacquemin, C. (1997). *Variation terminologique : reconnaissance et acquisition automatique de termes et de leurs variantes en corpus.* Habilitation à diriger des recherches en informatique, Université de Nantes, Nantes.

[Jacquemin, 2003] Jacquemin, C. (2003). Syntagmatic and paradigmatic representations of term variation. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 341–348, University of Mariland.

[Kamp, 1981] Kamp, H. (1981). Evénements, représentations discursives et référence temporelle. *Langages*, (64) :34–64.

[Kavanagh, 1995] Kavanagh, J. (1995). *The Text Analyser : A Tool for Extracting Knowledge From Text.* Master of computer science thesis, University of Ottawa, Ottawa, Canada.

[Kayser, 1987] Kayser, D. (1987). Une sémantique qui n'a pas de sens. *Langages*, Sémantique et intelligence artificielle(87) :33–45.

[Kayser, 1988] Kayser, D. (1988). What kind of thing is a concept ? *Computationnal Linguistics*, 4 :158–165.

[Kayser, 1997] Kayser, D. (1997). *La représentation des connaissances.* Collection informatique. Hermès, Paris.

[Kayser et Abir, 1991] Kayser, D. et Abir, H. (1991). Lexical Semantics and Shifts in Meaning. In *Actes du 1er séminaire de Sémantique Lexicale du PRC-GDR Communication Homme-Machine*, pages 89–99, Toulouse.

[Kim et al., 2004] Kim, J.-D., Ohta, T., Tsuruoka, Y., Tateisi, Y., et Collier, N. (2004). Introduction to the Bio-Entity Recognition Task at JNLPBA.

In Collier, N., Rush, P., et Nazarenko, A., editors, *Proceedings on International Joint Workshop on Natural Language Processing in Biomedecine and its Applications (BioNLP&LNPBA)*, pages 70–75, Geneva, Switzland. International Conference on Computational Linguistics (COLING'04).

[Klavans et Resnik, 1996] Klavans, J. et Resnik, P., editors (1996). *The Balancing Act – Combining Symbolic and Statistical Approaches to Language.* Language, Speech, and Communication. MIT Press.

[Kripke, 1972] Kripke, S. (1972). *La logique des noms propres.* Editions de minuit.

[Krovetz, 1993] Krovetz, R. (1993). Viewing morphology as an inference process. In *Proceedings of the ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 191–202, Pittsburgh, USA.

[Krovetz et Croft, 1992] Krovetz, R. et Croft, B. W. (1992). Lexical Ambiguity and Information retreival. *ACM Transactions on Information Systems*, pages 145–161.

[Landau et al., 1993] Landau, M.-C., Sillion, F., et Vichot, F. (1993). EXOSEME : a Thematic Document Filtering System. In *Proceedings of Avignon'93*, Avignon.

[Lebart et Salem, 1994] Lebart, L. et Salem, A. (1994). *Statistique textuelle.* Dunod.

[Lehnert, 1977] Lehnert, W. (1977). Human and computational question answering. *Cognitive Science.*

[Lenat et Guha, 1990] Lenat, D. et Guha, R. V. (1990). *Building Large Knowledge Based Systems : Representation and Inference in the Cyc Project.* Addison-Wesley.

[Lyons, 1980] Lyons, J. (1980). *Sémantique linguistique.* Collection "langue et langage". Larousse Université.

[Mahon et Smith, 1996] Mahon, J. G. M. et Smith, F. J. (1996). Improving Statistical Language Model Performance with Automatically Generated Word Hierarchies. *Computational Linguistics*, 22(2) :217–247.

[Manning et Schütze, 1999] Manning, C. et Schütze, H. (1999). *Foundations of statistical Natural Language Processing.* MIT Press, Cambridge, MA.

[Meyer et al., 1992] Meyer, I., Skuce, D., Bowker, L., et Eck, K. (1992). Towards a new generation of terminological resources : an experiment in building a terminological knowledge base. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING'92)*, pages 956–960, Nantes, France.

[Minel, 2002] Minel, J.-L. (2002). *Filtrage sémantique du résumé automatique à la fouille de textes.* Hermes.

[Morin, 1999] Morin, E. (1999). Acquisition de patrons lexico-syntaxiques caractéristiques d'une relation sémantique. *Traitement Automatique des Langues*, 40(1).

[Morin et Jacquemin, 1999] Morin, E. et Jacquemin, C. (1999). Projecting Corpus-Based Semantic Links on a Thesaurus. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, University of Maryland.

[Morin et Martienne, 1999] Morin, E. et Martienne, E. (1999). Raffinement de patrons lexico-syntaxiques par un système d'apprentissage. In *Actes de IC-99 (Ingénierie des Connaissances)*, Palaiseau, France.

[MUC, 1995] MUC (1995). *MUC 1995 Proceedings of the Sixth Message Understanding Conference.*

[MUC, 1998] MUC (1998). *MUC 1998 Proceedings of the Sixth Message Understanding Conference.*

[Mulvany, 1993] Mulvany, N. C. (1993). *Indexing Books.* The University of Chicago Press.

[Nazarenko, 1994] Nazarenko, A. (1994). *Compréhension du Langage Naturel : le problème de la causalité.* Thèse d'Informatique. Université de Paris-Nord XIII. Directeur : Daniel Kayser.

[Nazarenko, 1996] Nazarenko, A. (1996). Structurer les graphes pour comprendre la causalité dans les textes. *Revue d'Intelligence Artificielle*, 10(1) :163–198.

[Nazarenko, 2000] Nazarenko, A. (2000). *La cause et son expression en français.* L'essentiel Français. OPHRYS.

[Nazarenko, 2005] Nazarenko, A. (2005). *Sur quelle sémantique reposent les méthodes automatiques d'accès au contenu textuel ?* In Condamines, A., editor, *Sémantique et Corpus.* Hermes.

[Nazarenko et Hamon, 2002] Nazarenko, A. et Hamon, T. (2002). Structuration de terminologie : quels outils pour quelles pratiques ? *Traitement Automatique des Langues*, 43(1) :7–42.

[Nazarenko et Poibeau, 2004] Nazarenko, A. et Poibeau, T. (2004). *Evaluation des systèmes d'analyse et de compréhension de texte.* In Chaudiron, S., editor, *L'évaluation des systèmes de traitement de l'information.* Lavoisier, Paris.

[Nazarenko et al., 2001] Nazarenko, A., Zweigenbaum, P., Habert, B., et Bouaud, J. (2001). *Corpus-based extension of a terminological semantic lexicon.* In Bourigault, D., Jacquemin, C., et L'Homme, M., editors, *Recent Advances in Computational Terminology*, volume 2 of *Natural Langage Processing*, pages 327–351. John Benjamins, Amsterdam.

[Nédellec, 2000] Nédellec, C. (2000). Knowledge Extraction from Text, a Machine Learning Approach. In *Proceedings of the Third International Conference on Human-System Learning (CAPS'3 Learning WWW)*, Paris, France. Europia Production.

[Nédellec et Nazarenko, 2004] Nédellec, C. et Nazarenko, A. (2004). *Ontologies and Information Extraction.* In Staab, S. et Studer, R., editors, *Handbook on Ontologies in Information Systems.* Springer Verlag. Article non publié à ce jour, suite à une erreur des éditeurs.

[Nédellec et al., 2001] Nédellec, C., Vetah, M. O. A., et Bessières, P. (2001). Sentence Filtering for Information Extraction in Genomics : A Classification Problem. In Raedt, L. D. et Siebes, A., editors, *Principles of Data Mining and Knowledge Discovery (Proceedings of the 5th European Conference Practical Knowledge Discovery in Databases, PKDD'2001)*, number 2167 in Lecture Notes in Artificial Intelligence, page 326–338, Freiburg, Germany. Springer Verlag.

[Pazienza, 1997] Pazienza, M. T., editor (1997). *Information Extraction : a Multidisciplinary Approach to an Emerging Information Technology.* Springer, Berlin.

[Pearson, 1998] Pearson, J. (1998). *Terms in Context.* Studies in Corpus Linguistics. John Benjamins.

[Pillet, 2000] Pillet, V. (2000). *Méthodologie d'extraction automatique d'information à partir de la littérature scientifique en vue d'alimenter un nouveau système d'information.* Thèse de Doctorat en informatique. Université de droit, d'économie et des sciences d'Aix-Marseille.

[Ploux et Victorri, 1998] Ploux, S. et Victorri, B. (1998). Constructions d'espaces sémantiques à l'aide de dictionnaires de synonymes. *Traitement Automatique des Langues (T.A.L.)*, 39(1) :161–182.

[Poibeau, 2003] Poibeau, T. (2003). *Extraction d'information à base de connaissances hybrides.* Thèse d'Informatique . Université de Paris XIII.

[Poibeau et Nazarenko, 1999] Poibeau, T. et Nazarenko, A. (1999). L'extraction d'information, vers une nouvelle conception de la compréhension de textes ? *Traitement Automatique des Langues (T.A.L.)*, 42(2) :87–115.

[Proux et al., 1998] Proux, D., Rechenmann, F., Julliard, L., Pillet, V., et Jacq, B. (1998). *Detecting Gene Symbols and Names in Biological Texts :*

A First Step toward Pertinent Information Extraction. In *Genome Informatics*. Universal Academy Press, Inc.

[Pâris, 2003] Pâris, L. (2003). *Construction automatique d'index : validation de résultats et évaluation.* Mémoire de Maîtrise de Linguistique Informatique. Université de Paris VII.

[Péry-Woodley, 1995] Péry-Woodley, M.-P. (1995). Quels corpus pour quels traitements automatiques ? *Traitement Automatique des Langues*, 36(1-2) :213–232.

[Rastier, 1991] Rastier, F. (1991). *Sémantique et recherches cognitives.* Presses Universitaires de France.

[Rastier, 2001] Rastier, F. (2001). *Arts et sciences du texte.* Presses Universitaires de France.

[Rastier et al., 1994] Rastier, F., Cavazza, M., et Abeillé, A. (1994). *Sémantique pour l'analyse : de la linguistique à l'informatique.* Coll. Recherches cognitives. Masson, Paris.

[Riloff, 1993] Riloff, E. (1993). Automatically constructing a Dictionary for Information Extraction Tasks. In *Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI-93).*

[Riloff et Jones, 1999] Riloff, E. et Jones, R. (1999). Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. In *Proceedings of AAAI-99*, pages 474–479.

[Riloff et Schmelzenbach, 1998] Riloff, E. et Schmelzenbach, M. (1998). An Empirical Approach to Conceptual Case Frame Acquisition. In *Proceedings of the Sixth Workshop on Very Large Corpora.*

[Riloff et Shepherd, 1997] Riloff, E. et Shepherd, J. (1997). A corpus-based approach for building semantic lexicons. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP-97)*, pages 117–124, Providence, RI.

[Rousselot et al., 1996] Rousselot, F., Frath, P., et Oueslati, R. (1996). Extractinf Concepts and Relations from Corpora. In Wahlster, W., editor, *Proceedings of the ECAI-96 Workshop on Corpus-Oriented Semantic Analysis.* ECAI, John Wiley & Sons.

[Récanati, 1979] Récanati, F. (1979). *La transparence et l'énonciation : pour introduire à la pragmatique.* Coll. L'ordre philosophique. Le Seuil, Paris.

[Schütze et Pedersen, 1995] Schütze, H. et Pedersen, T. (1995). Information retrieval based on word senses. In *Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, Las Vegas, NV.

[Séguéla, 1999] Séguéla, P. (1999). Adaptation semi-automatique d'une base de marqueurs de relations sémantiques sur des corpus spécialisés. *Terminologies Nouvelles*, 19 :52–60. Acte du colloque Terminologie et Intelligence Artificielle, 10-11 mai 1999.

[Shütze, 1998] Shütze, H. (1998). Automatic Sense Discrimination. *Computational Linguistics*, 24(1) :97–124.

[Simmons, 1965] Simmons, R. F. (1965). Answering English Questions by Computer : A Survey. *Communications of the ACM*, 8(1) :53–700.

[Sinclair, 1996] Sinclair, J. M. (1996). CEE - Preliminary Recommendations on Corpus Typology. Technical report, EAGLES (Expert Advisory Group on Language Engineering Standards).

[Smeaton, 1997] Smeaton, A. F. (1997). *Using NLP or NLP resources for information retrieval tasks.* In Strzalkowski, T., editor, *Natural language information retrieval*, pages 99–111. Kluwer Academic Publishers, Dordrecht, NL.

[Soderland, 1999] Soderland, S. (1999). Learning Information Extraction Rules for Semi-Structured and Free Text. *Machine Learning Journal*, 34 :233–272.

[Sowa, 1984] Sowa, J. F. (1984). *Conceptual Structures. Information Processing in Mind and Machine.* The system programming series. Addison Wesley Publishing Company, Reading, MA.

[Sowa, 1992] Sowa, J. F. (1992). Lettre électronique du Forum "Conceptuel Graphs" datée du 26 juillet 1992.

[Sparck Jones, 1997] Sparck Jones, K. (1997). *What is the role of NLP in text retrieval ?* In Strzalkowski, T., editor, *Natural language information retrieval.* Kluwer.

[Stapley et Benoit, 2000] Stapley, B. et Benoit, G. (2000). Bibliometrics : Information Retrieval and Visualization from co-occurrence of gene names in MedLine abstracts. In *Proceedings of the Pacific Symposium on biocomputing (PSB'2000)*, Honolulu.

[Szulman et al., 2002] Szulman, S., Biébow, B., et Aussenac-Gilles, N. (2002). Structuration de Terminologies à l'aide d'outils de TAL avec TERMINAE. *Traitement Automatique des Langues*, 43(1) :103–128.

[Teulier et al., 2004] Teulier, R., Charlet, J., et Tchounikine, P. (2004). *Ingénierie des connaissances.* L'harmattan, Paris.

[Thomas et al., 2000] Thomas, J., Milward, D., Ouzounis, C., Pulman, S., et Caroll, M. (2000). Automatic Extraction of Protein Interactions from

Scientific Abstracts. In *Proceedings of the Pacific Symposium on biocomputing (PSB'2000)*, volume 5, pages 502–513, Honolulu.

[Vauvert, 2004] Vauvert, G. (2004). Format for NLP Annotation - Progress Report. Technical report, Alvis Project (STREP).

[Victorri et Fuchs, 1996] Victorri, B. et Fuchs, C. (1996). *La polysémie, construction dynamique du sens*. Editions Hermès. 220 pp.

[Voorhees, 1994] Voorhees, E. (1994). Query Expansion Using Lexical-Semantic Relations. In *Proceedings of the ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 61–70.

[Voorhees, 1998] Voorhees, E. M. (1998). *Using WordNet for text retrieval*. In Fellbaum, C., editor, *WordNet : an Electronic Lexical Database*, chapter 12, pages 285–303. MIT Press, Cambridge, MA.

[Voorhees et Tice, 1999] Voorhees, E. M. et Tice, D. M. (1999). The TREC-8 question answering track evaluation. In *Proceedings of the 8th Conference (TREC-8)*, pages 83–105, Maryland. NIST.

[Wacholder et Nevill-Manning, 2001] Wacholder, N. et Nevill-Manning, C. (2001). Workshop report : The Technology of Browsing Applications, Workshop held in conjunction with JCDL 2001. *SIGIR Forum*, 35(1) :16–19. http ://www.acm.org/sigir/forum/S2001-TOC.html.

[Weissenbacher, 2004] Weissenbacher, D. (2004). La relation de synonymie en génomique. In *Actes des Journées nationales sur le Traitement Automatique des Langues Naturelles (TALN'04)*, Fès.

[Yarowsky, 1992] Yarowsky, D. (1992). Word-Sense Disambiguation Using Statistical Models of Roget's Cateories Trained on Large Corpora. In *Proceedings of COLING'92*, pages 454 – 460, Nantes, France. International Conference on Computational Linguistics.

[Yu et Agichtein, 2003] Yu, H. et Agichtein, E. (2003). Extracting synonymous gene and protein terms from biological literature. *Bioinformatics*, 19(Suppl 1).

[Yu et al., 2002] Yu, H., Hatzivassiloglou, V., Friedman, C., Rzhetsky, R., et Wilbur, W. J. (2002). Automatic Extraction of gene and protein synonyms from medline and journal articles. In *Proceedings of AMIA Symposium*, pages 413–423.

[Zhou et Su, 2004] Zhou, G. et Su, J. (2004). Exploring Deep Knowledge Resources in Biomedical Name Recognition. In Collier, N., Rush, P., et Nazarenko, A., editors, *Proceedings on International Joint Workshop on Natural Language Processing in Biomedecine and its Applications (BioNLP&LNPBA)*, pages 96–99, Geneva, Switzland. International Conference on Computational Linguistics (COLING'04).

[Zweigenbaum, 1999] Zweigenbaum, P. (1999). Encoder l'information médicale : des terminologies aux systèmes de représentation des connaissances. *Innovation Stratégique en Information de Santé*, (2-3) :27–47.

[Zweigenbaum et al., 1997] Zweigenbaum, P., Bouaud, J., Nazarenko, A., et Habert, B. (1997). Coopération apprentissage en corpus et connaissances du domaine pour la construction d'ontologies. In *Actes des 1ères Journées scientifiques et techniques FRANCIL 97*, pages 501–508, Avignon. AUPELF-UREF.

[Zweigenbaum et Consortium MENELAS, 1994] Zweigenbaum, P. et Consortium MENELAS (1994). MENELAS : an Access System for Medical Records using Natural Language. *Computer Methods and Programs in Biomedicine*, 45 :117–120.