THÈSE DE DOCTORAT DE L'UNIVERSITÉ PARIS XIII

Spécialité : **Informatique**

Présentée pour obtenir le titre de :
Docteur de l'Université Paris XIII

par

Elie NAULLEAU

Titre de la thèse :

Apprentissage et Filtrage Syntaxico-Sémantique de Syntagmes Nominaux Pertinents pour La Recherche Documentaire

Devant le jury composé de :

Jean-Pierre Desclés	Rapporteur
Benoît Habert	Examinateur
Daniel Kayser	Directeur
Adeline Nazarenko	Examinateur
Jean-David Sta	Examinateur
Bernard Victorri	Rapporteur

Janvier 1998

Remerciements

Qu'il me soit permis de remercier Daniel Kayser (LIPN, professeur à l'Université de Paris XIII) qui a accepté d'être mon directeur de thèse.

Benoît Habert (ELI, ENS de Fontenay St Cloud) a assuré l'encadrement scientifique de la thèse. Sans son soutien bienveillant, ses conseils, sa disponibilité, il est évident que je n'aurais pas été capable de poursuivre ce travail jusqu'à son terme. Je tiens à lui exprimer toute ma reconnaissance ainsi que le plaisir que j'ai eu à travailler avec lui.

J'adresse mes sincères remerciements à Bernard Victorri (Directeur de Recherche CNRS, ENS) et Jean-Pierre Desclés (professeur à l'Université Paris IV) qui ont accepté d'être mes rapporteurs.

J'ai eu la chance d'effectuer cette thèse CIFRE au groupe SID-ISI de la Direction des Etudes et Recherche d'Electricité de France. Ma gratitude s'adresse à Jean-Luc Sanson (chef du groupe SID-ISI) et à toute l'équipe, qui m'ont soutenu durant ces trois années enrichissantes.

En particulier, pour m'avoir fait confiance en me proposant ce poste de thésard, je remercie Marie-Gaëlle Monteil (chercheur à la DER-EDF).

Je remercie également Jean-David Sta (chercheur à la DER-EDF) qui a encadré ma thèse à EDF. Les discussions fructueuses que nous avons eues ont contribué à approfondir ce travail.

Je remercie Marie-Luce Herviou (chercheur à la DER) et Richard Quatrain (chercheur à la DER) pour leur assistance et leurs conseils avisés.

Mes remerciements vont encore à Adeline Nazarenko (LIPN, Université Paris XIII), Houssem Assadi (Némésia), Didier Bourigault (CNRS) pour leurs remarques et leurs conseils, ainsi qu'à Philippe Chassany (Unifix) pour ses tuyaux en programmation C et Tk/Tcl.

Enfin, je remercie ma femme Pénélope pour ses relectures et son soutien quotidien.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

1.1 Contexte et objectif

Ce travail de thèse a été mené dans le cadre d'une convention CIFRE ANRT entre l'ENS de Fontenay St Cloud et la Direction des Etudes et Recherches (DER) d'Electricité de France, au sein du département Système d'Information et de Documentation (SID) de la DER. L'activité du département SID s'articule autour de la gestion de l'information. Cela concerne principalement l'accès, le stockage, et la diffusion de l'information. Les applications d'informatique documentaire y occupent une place primordiale : bibliothèque électronique, recherche documentaire, diffusion ciblée d'information, veille technologique, extraction de connaissance[1].

Evolution rapide du contexte en recherche d'information A son début, cette thèse s'orientait vers une forme d'extraction terminologique, étant admis que les termes véhiculent des concepts majeurs abordés dans les documents. Ainsi, le thesaurus EDF exploité par plusieurs applications documentaires, et dont la complétude doit garantir la pertinence des résultats des traitements, devait être mis à jour : il s'agissait d'analyser de nouveaux textes, d'en extraire les nouveaux concepts qui, une fois introduits dans le thesaurus, auraient permis d'en actualiser la couverture documentaire.

Mais trois ans plus tard, cette démarche n'est plus envisageable; en raison de la production croissante et accélérée de documents électroniques, on est en droit de se demander si un thesaurus est adapté aux nouveaux textes que l'on souhaite indexer. Une mise à jour du thesaurus, processus long et coûteux, quand bien même serait-elle effectuée en un temps record, ne pourrait garantir une couverture documentaire optimale, ce thesaurus accusant inévitablement un retard face aux nouvelles données textuelles disponibles.

Dans un système documentaire où la collection de documents est ouverte, le contrôle lors de l'indexation automatique des textes n'est pas avantageux. Beaucoup

1. Le résumé automatique de texte, l'interrogation de bases multilingues et la traduction ne tarderont pas à faire leur apparition car il existe de réels besoins.

de systèmes documentaires, reposent encore sur des thesaurus mis sous forme électronique à partir d'un original en papier comme c'est le cas pour le thesaurus EDF. C'est le fruit d'un long travail, de nature le plus souvent terminographique. En numérisant le thesaurus papier, on fructifiait un travail important dont la pérennité était rendue possible grâce à une mise à jour régulière, mise à jour qui aujourd'hui n'est plus envisageable au regard de la quantité de documents à prendre en compte et des attentes des utilisateurs finaux. Ces derniers souhaitent le plus souvent un accès immédiat à l'information.

L'indexation automatique contrôlée n'est pas seulement mise en difficulté par la quantité importante de documents à traiter et les problèmes de mise à jour qui en résultent. Un autre aspect est le caractère éphémère des textes et par conséquent des index établis à partir de ces textes. Non seulement la production de documents s'accélère mais ces documents n'ont pas toujours une valeur de référence en soi : leur intérêt n'est parfois que de courte durée. L'investissement dans la constitution d'index contrôlés s'avère donc d'autant moins justifié.

Importance de l'activité TALN à la DER

Les études menées en TALN à la DER (voir entre autres [OHD94, HP96, MLH95, HM94, Sta94, AB96, Bou94a, Bou94b]) s'inscrivent dans une démarche de traitement massif de l'information textuelle : les techniques linguistiques sont perçues comme un moyen de produire automatiquement des représentations standardisées des textes, exploitables par des systèmes d'information, pour des applications documentaires appliquées aux textes scientifiques et techniques [OHD94]: classements automatiques, comparaisons, diffusion ciblée, routage, veille technologique, capitalisation du savoir-faire, gestion et accès à la documentation technique; mais aussi des applications d'analyse sociologique : dépouillement de résultats d'entretiens, analyses de réponses à des questions ouvertes. Le département SID gère pour la DER un important fond documentaire d'environ 600.000 références de documents (résumés et documents numérisés) qui croît régulièrement. L'équipe de recherche TALN compte environ une dizaine de chercheurs répartis entre les département SID et IMA (Informatique et Mathématiques Appliquées). Ces derniers travaillent en collaboration avec d'autres chercheurs dont l'activité est liée à des domaines connexes au TALN : structuration de documents, base de données, statistique, interfaces homme-machine.

Le projet ASTREE (Atelier Standardisé pour le Traitement Electronique de l'Ecrit) Ce projet, mené par SID (1993-1996) et dans lequel notre thèse a pris place, visait à structurer et modéliser l'information textuelle, massive et tout-venant, afin de l'exploiter au mieux. Le projet reposait sur la mise en place d'un atelier linguistique faisant appel à des techniques de traitement du langage naturel pour accéder au contenu des documents, de langue française ou anglaise.

De l'indexation manuelle à l'indexation automatique Un des premiers résultats du projet ASTREE est la mise en place d'un serveur d'indexation automatique.

Depuis plus de vingt ans, EDF a soutenu une activité de gestion de terminologie, qui a donné lieu à la création d'un thesaurus couvrant l'ensemble des intérêts relatifs à l'entreprise. Ce thesaurus (voir section 3.2.2), numérisé il y a quelques années, alimente depuis 1994 une application client-serveur linguistique, qui fournit un service d'indexation automatique de textes. Les applications documentaires qui ont besoin de points d'accès dans des documents se connectent donc à ce serveur [2].

Implication de la DER dans des projets européens Pour industrialiser les outils et réduire les coûts de développement, la DER a participé à plusieurs projets européens dans le domaine du TALN : GENELEX, GRAAL et TRANSTERM.

Le projet GENELEX a permis de développer un modèle de dictionnaire architecturé sur trois couches, correspondant aux niveaux morphologiques, syntaxiques et sémantique. Le principe fondamental est le suivant : à une unité morphologique correspond autant d'unités syntaxiques que de fonctionnements syntaxiques différents. Chaque unité syntaxique est reliée à autant d'unités sémantiques que de sens différents par construction syntaxique. Le modèle se présente sous la forme d'une DTD SGML. Pour nos expériences, nous avons eu à disposition le dictionnaire AlethDic (v1.5.5), document conforme à la DTD Genelex et mis au point par la société Erli. Le dictionnaire AlethDic est présenté en annexe A.

Le projet GRAAL (1992-1996) [St94] visait à construire des Grammaires Réutilisables pour l'Analyse Automatique des Langues. Les principaux résultats de ce projets sont [Her95] : une bibliothèque de grammaires au format DTD GRAAL facilement maintenables et réutilisables dans différents domaines applicatifs, une boîte à outils d'ingéniérie linguistique (moteurs linguistiques), mais aussi des ateliers de construction de grammaire, et de mise en place d'applications. Et enfin, des applications pilotes mises en place chez les partenaires utilisateurs : indexation automatique, extraction de connaissances, traduction assistée. La boîte à outils KES [3] [Her95], spécifiée et développée dans le cadre du projet européen GRAAL est dédiée à l'extraction et la structuration de connaissances à partir de corpus. Elle facilite l'ensemble des activités permettant de passer de textes bruts à un ensemble structuré de données. Elle est adaptable et configurable dans différentes configurations applicatives : extraction et mise à jour de terminologies, analyse de corpus à des fins d'interprétation sociologique, analyse fine du contenu des textes (repérage des concepts, des actions).

Le projet TRANSTERM (1993-1996) a conduit à la définition d'un modèle pour la représentation de terminologies multilingues utilisées dans les ressources de TALN. Succintement, les notions terminologiques (*terminological units*) sont regroupées dans des *containers* selon le domaine d'activité dont elles sont issues. Les mêmes notions terminologiques peuvent avoir des réalisations linguistiques dans plusieurs langues. Le modèle permet d'intégrer ces «représentants linguistiques» afin que les terminolo-

2. Dans son état actuel, le serveur repose sur l'application AlethIP issue du projet GRAAL et conçue par la société Erli.

3. Knowledge Extracting and Structuring.

gies soient utilisées de manière transparente par des applications de TALN comme les entrées lexicales dans un dictionnaire électronique. Contrairement au modèle e-TIF [Mel95] qui définit un format orienté vers l'échange de terminologies entre organismes, le modèle TRANSTERM est orienté vers l'utilisation des terminologies dans des applications de TALN.

1.2 L'informatique documentaire et l'accès à l'information pertinente : l'indexation automatique

Alors que de plus en plus de textes sont diffusés sous forme électronique, l'accès à l'information pertinente est un des principaux problèmes soumis à l'informatique documentaire. D'après la définition de Vickery [Vic60], l'analyse documentaire consiste à dériver d'un document un ensemble de mots qui lui sert de représentation condensée. Cette représentation peut être utilisée pour identifier le document, pour procurer des points d'accès dans la recherche de la documentation, pour renseigner sur son contenu, ou pour servir de substitut au document. Aujourd'hui, cette définition reste valable et le problème de l'analyse documentaire se pose par rapport aux techniques informatiques : dans un premier temps, il s'agit d'identifier de manière assistée ou automatique les mots ou expressions pouvant servir de représentation condensée au document, c'est la phase d'indexation. Dans un deuxième temps, ces mots ou expressions doivent être retrouvés dans le document ou une collection de documents ; c'est la phase de recherche documentaire.

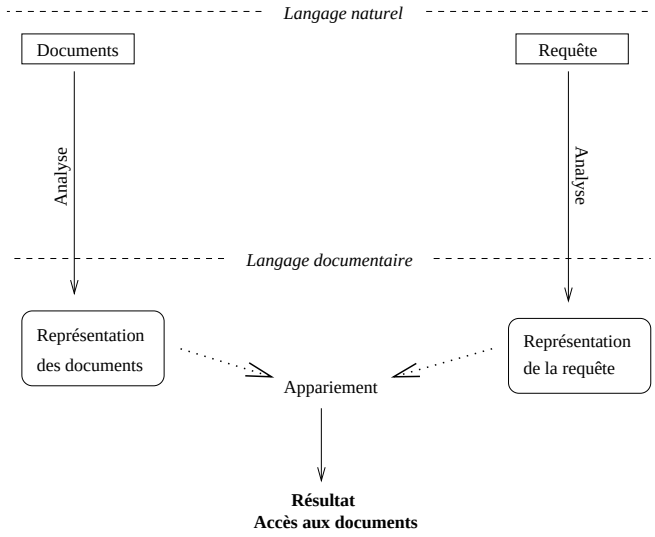
L'indexation de texte fournit des points d'accès précalculés dans un document. Le statut sémiotique de ces points d'accès et les méthodes utilisées pour les calculer définissent plusieurs formes d'indexation que nous décrivons maintenant. Chaque méthode d'indexation relève d'une stratégie documentaire car son coût et sa faisabilité dépendent de la nature des données à traiter et des résultats attendus.

Les indexations libres

- l'indexation plein-texte considère tous les mots simples du texte comme des index potentiels. Certaines formes graphiques peuvent en être exclues, comme les mots grammaticaux et les articles, voire d'autres formes mentionnées dans un anti-dictionnaire. Des techniques de statistique textuelle peuvent être appliquées pour sélectionner des index sur une pondération calculée d'après leur fréquence et leur distribution dans les documents [Sal89].

- D'autres formes d'indexation libre font appel à des traitements linguistiques : des traitements morphologiques appliqués au texte permettent de sélectionner comme index des graphies lemmatisées voire catégorisées. Cela a pour effet de normaliser les formes graphiques et de regrouper des formes qui diffèrent par leur flexion. D'autres traitements, syntaxiques, permettent de repérer des

FIG. 1.1 – *Principe d'utilisation d'un langage documentaire d'après [Cou76]*



groupes nominaux polylexicaux. Ces derniers peuvent subir des transformations et des simplifications systématiques pour être normalisés (effacement de certains articles, adjectifs, distribution des groupes coordonnés). Des traitements statistiques basés sur l'information mutuelle peuvent être appliqués pour détecter des collocations remarquables [CH90] et sont utilisés de manière combinée avec les traitements linguistiques pour détecter des collocations (système *Xtract* [Sma93a]) ou des candidats termes [Dai94].

L'indexation contrôlée

Alors que les index libres sont tirés des documents à indexer, les index contrôlés proviennent d'un vocabulaire dont la pertinence est contrôlée. Ils constituent un langage documentaire. Ce langage artificiel se définit en prélevant dans le langage naturel les expressions dont il a besoin pour couvrir un certain univers conceptuel. Les composants de cet univers sont appelés descripteurs et peuvent être organisés grâce à des relations[4]. L'intégrité linguistique des descripteurs n'est pas nécessairement conservée : les expressions prélevées peuvent être simplifiées ou codées pour assurer l'univocité de l'association descripteur–concept. Avec un langage ainsi défini, on peut produire une représentation formalisée des documents et des requêtes sur les documents afin de faciliter leur appariement[5]. Le principe d'utilisation du langage

4. Par exemple les relations du thesaurus EDF sont: terme-spécifique—terme-générique (ainsi *voiture—véhicule*), voir-aussi (*registre voir-aussi microprocesseur*), employé-pour (*calculateur employé-pour ordinateur*.

5. C'est-à-dire l'opération qui identifie dans le document l'objet de la requête, par un calcul de similarité ou par une mise en correspondance.

documentaire est schématisé en figure 1.1.

D'après [Cou76], l'indexation automatique contrôlée a l'avantage de fournir des résultats de meilleure qualité : elle identifie dans un document des formes connues à l'avance dont on sait qu'elles sont pertinentes. De plus, le langage documentaire peut être organisé sous la forme d'un thesaurus, ce qui rend possible la sur-indexation par des concepts plus généraux que ceux trouvés dans le documents. Mais les difficultés sont nombreuses. Tout d'abord le coût : définir un langage documentaire est une lourde tâche. Il faut constituer un thésaurus (ou parfois des listes de descripteurs) à partir des documents à indexer. Si de nouveaux documents doivent être indexés, il faut s'attendre à une insuffisance de couverture du thesaurus. Dans ce cas, une mise à jour du vocabulaire de contrôle ou de la taxonomie des concepts du thesaurus est nécessaire. Une autre caractéristique de l'indexation contrôlée est qu'elle requiert un traitement complexe du flot textuel. En effet, le vocabulaire contrôlé est sujet à des variations dans les documents et peut être ambigu. L'insertion d'un adjectif, d'un adverbe ou le rattachement d'un groupe prépositionnel peut faire échouer la reconnaissance de l'index. Pour prendre en compte ces phénomènes de variation, le texte doit subir une analyse morpho-syntaxique afin que des traitements plus sophistiqués de recherche de variantes puissent lui être appliqué ([Jac96, Jac97, Dai94]). De plus, l'ambiguïté des unités lexicales qui entrent dans la composition du vocabulaire contrôlé est une source certaine de bruit. Bon nombre d'unités lexicales simples sont ambiguës (comme *base, régime, agent, code*) et certaines unités polylexicales le sont aussi [6]. Dans les meilleurs des cas, une désambiguïsation des unités lexicales est réalisée [7].

1.2.1 L'utilisation de ressources terminologiques en informatique documentaire

La masse croissante d'information textuelle produite par les sociétés industrielles est diffusée de par le monde grâce à l'expansion des réseaux. Une telle situation manifeste une intensification et une industrialisation des activités relatives à la terminologie et à la traduction. La terminologie, utilisée depuis longtemps pour faciliter la communication et la coopération entre les différents corps de métiers, souvent traduite pour faire face à la mondialisation des échanges, peut désormais être considérée dans une perspective nouvelle, celle de l'accès automatisé à l'information. En effet, les termes renvoyant à des notions spécialisées, constituent *a priori* des index pertinents. De plus, les terminologies mettant les termes en relation, ces relations sont utilisables en interrogation documentaire pour l'expansion de requêtes (les index

6. Par exemple *analyse de données* est un terme propre aux statistiques mais est aussi utilisé dans un sens général comme en linguistique

7. Les traitements de désambiguïsation nécessitent des ressources linguistiques importantes, quel que soit le modèle utilisé - stochastique ou linguistique. De ce point de vue les anglophones sont avantagés puisqu'ils disposent de ressources mises à disposition librement : Roget thesaurus, Base WordNet, Corpus catégorisés, etc. Ces dernières sont utilisées pour réaliser des évaluations standard de méthodes [Gre96] et notamment de méthodes de désambiguïsation [Yar92])

trouvés dans la requête sont remplacés par des termes synonymes ou plus généraux, ce qui permet l'extension du champ d'investigation de la requête).

Mais les pratiques terminographiques manuelles classiques ne sont plus adaptées au cycle de production de l'information et aux applications documentaires qui en découlent : masse textuelle trop importante, accélération de sa croissance. Il est difficilement envisageable de produire manuellement des thesaurus ou de les mettre à jour pour en actualiser constamment la couverture documentaire. Par exemple ceci est patent lorsque l'on considère l'évolution du thesaurus EDF. Les pratiques terminographiques manuelles pour la constitution d'index contrôlés ne sont plus d'actualité dans des systèmes dont les fonds documentaires sont mis à jour régulièrement : elles doivent être abandonnées ou épaulées par des outils spécifiques. Des outils d'extraction terminologique sont développés pour fournir une aide au terminologue et l'assister dans le dépouillement des sources, en lui proposant des groupes nominaux filtrés voire des candidats termes. Par exemple, le logiciel *Lexter* conçu par D. Bourigault [Bou94b] et utilisé à EDF, ou encore l'application AlethIP et la nouvelle version de la boîte à outils KES développée à SID.

L'indexation contrôlée à la DER-EDF

La solution choisie au service au département SID jusqu'à présent est d'utiliser sa terminologie comme un langage documentaire. C'est dans cette optique que le thesaurus papier EDF a été numérisé pour être utilisé comme un thesaurus d'indexation automatique contrôlée. Mais termes et descripteurs sont alors confondus à tort dans le thesaurus. D'un point de vue méthodologique, mettre sur le même plan sémiotique ces deux types d'objet – l'un documentaire, l'autre linguistique – conduit à une sous-évaluation de la complexité des traitements linguistiques qui doivent être envisagés pour tirer parti d'une terminologie en indexation contrôlée.

Le descripteur, vocable du langage documentaire, est un élément révélateur d'une notion. le descripteur n'est pas un objet linguistique à proprement parler, même si sa forme est inspirée d'une forme de la langue. Il n'est pas destiné à une interprétation naturelle. Il représente un niveau intermédiaire entre la langue et la représentation informatique pour rendre possible l'appariement.

Par exemple, dans un langage documentaire qui couvrirait chimie et géométrie, base1 et base2 réfèreraient respectivement à la SUBTANCE CHIMIQUE et à la base d'une figure. Lors de la phase de recherche documentaire, l'utilisateur fournirait donc base1 pour chercher les bases-SUBSTANCE et base2 pour rechercher les bases-OBJET GÉOMÉTRIQUE.

Le terme (ou unité terminologique) [W8̈1], contrairement au descripteur est un objet de la langue et désigne de façon univoque une notion dans un domaine de connaissance ou d'activité. Par exemple, le terme *base* en chimie est un nom de SUBSTANCE. Le terme est alors considéré comme une pièce dans le puzzle de la connaissance, connaissance elle-même compartimentée en domaines d'après les fondements de la terminologie.

La solution choisie à la DER montre que les problèmes de mise à jour et de maintenance ne sont pas simples [8] et qu'il n'est pas possible de transformer simplement une terminologie en un langage documentaire sans prendre en compte les deux phénomènes linguistiques suivants : ambiguïté sémantique et variation syntaxique.

D'une part, une stratégie de désambiguïsation est nécessaire pour résoudre la polysémie des nombreux unitermes. On trouve par exemple dans le thesaurus EDF des unitermes comme *code, base, agent, régime* qui, s'ils sont laissés polysémiques lors de l'indexation, risquent de créer simultanément du bruit ou du silence. Par exemple, le nom action est indexé 4710 fois dans notre corpus (défini au chapitre 3.2) comme un TITRE FINANCIER, alors qu'il n'est jamais question d'action au sens financier du terme. L'*agent* est-il un agent EDF ou une substance chimique? A quoi fait référence l'uniterme *régime*? Voici par exemple, quelques termes du thesaurus EDF qui pourraient être rencontrés dans les textes sous la forme elliptique *régime*.

Terme	**Thème du thésaurus**	**Champ du thème**
régime de retraite	Environnement social	Retraite
régime de marée	Sciences de la terre	Océanographie
régime de chauffage	Application de l'énergie	Chauffage des locaux
régime médical	Environnement social	Médecine
régime transitoire	Utilisation de système	Fonctionnement
régime alimentaire	Biologie	Physiologie
régime torrentiel	Mécanique des fluides	Ecoulement à surface libre
régime bornes-postes	Action commerciale	Vente d'énergie.

D'autre part, C. Jacquemin [Jac97] montre que le phénomène de modification–variation syntaxique de terminologie n'est pas mineur et qu'il a intérêt à être pris en compte aussi bien lors de l'acquisition terminologique (extraction de candidats termes) que lors de la phase de reconnaissance terminologique (indexation contrôlée). La prise en compte de la variation lors de l'acquisition permet comme cela est montré dans [Jac97] d'établir des relations notionnelles entre les candidats et les termes attestés qui servent de point de départ pour la recherche des variantes. Ainsi les syntagmes *réacteur à eau lourde* et *réacteur eau lourde eau lourde* pourraient être rapprochés, de même que *réacteur à eau légère* et *réacteur graphite à eau légère*. Et lors de l'indexation, la prise en compte de la variation permet d'assurer une meilleure reconnaissance des unités terminologiques. Par exemple, à partir du syntagme *support de ligne électrique aérienne en béton*, il sera possible de reconnaître les termes suivants du thésaurus : *support de ligne en béton* et *ligne aérienne*.

Devant la complexité et le coût des traitements à mettre en oeuvre, les alternatives à l'indexation contrôlée, c'est-à-dire les formes d'indexation libre et notamment d'indexation plein-texte, paraissent plus intéressantes : elles ne nécessitent pas de traitements complexes, elles sont moins coûteuses et sont immédiatement opérationnelles. Elles présentent toutefois un inconvénient majeur, qui apparaît lors de

8. Cela va de la validation des candidats termes par le comité thesaurus à l'intégration des nouveaux termes dans la taxonomie existante – mais plus toujours d'actualité – du thesaurus EDF.

l'utilisation de certains moteurs de recherche WEB : le bruit présent dans les résultats des requêtes augmente avec la quantité de documents indexés. Si cette forme d'indexation est une réponse efficace au caractère éphémère des documents (voir paragraphe 2.4), elle ne peut cependant founir des index d'aussi bonne qualité que des index attestés. La démarche que nous proposons dans le présent travail peut être présentée comme une forme d'indexation cherchant à tirer le meilleur de l'indexation contrôlée et de l'indexation libre.

1.3 Un système d'extraction de syntagmes nominaux pertinents ou d'indexation semi-contrôlée

L'approche que nous proposons se situe entre l'indexation libre et l'indexation contrôlée. Il s'agit d'une forme d'indexation libre qui est contrainte par des informations linguistiques, «contrôlées» par un utilisateur. Il s'agit de filtrer des groupes nominaux extraits de textes français, par rapport à un point de vue défini par l'utilisateur. Les groupes nominaux ainsi extraits servent alors d'index libres ou peuvent être introduits dans un vocabulaire contrôlé moyennant une validation.

1.3.1 Sortir du cadre de l'extraction de terminologie

Cette approche ne s'inscrit pas dans le cadre d'une méthode d'extraction ou d'acquisition de terminologie. Il y a plusieurs raisons à cela.

La première raison tient au rapport entre terminologie et linguistique qui n'est pas établi de manière satisfaisante. L. Guilbert [Gui76] en accord avec la doctrine wüstérienne écrit : «*l'unité terminologique est, par essence, monosémique alors que le mot en tant qu'unité linguistique est voué à la polysémie, parce qu'il est appelé à travers la diversité des actes de communication et des interprétations qu'il reçoit à se charger de diverses valeurs significatives. Le terminologisme (le terme) correspond à l'acte de dénomination, qui, dans son principe, procède d'une volonté de monosémie : l'industriel qui fabrique un produit lui donne un nom et il entend que ce nom lui reste attaché de telle manière que le produit ne puisse pas être confondu avec un autre*». Ce principe monosémique est local au domaine de spécialité, le mot retrouve légitimement sa polysémie naturelle dès qu'il est considéré hors d'un certain domaine de spécialité[9]. La terminologie repose ainsi sur une vision duale du monde : un monde scientifique ou industriel et le monde de la vie quotidienne, familier; chacun de ces mondes aurait ainsi sa langue propre, une langue de spécialité, au service de la science et de l'industrie et une langue générale. Mais que se passe-t-il si au cours d'«une conversation de langue générale» il est fait mention d'un terme d'une «langue de spécialité»? L'énoncé sera-t-il qualifié de général ou spécialisé?

9. La notion de terme repose sur la notion de domaine de connaissance qui permet d'assurer son caractère monoréférentiel. [Gou90], chapitre 3 définit ce nécessaire caractère monoréférentiel du terme, comme l'*utopie terminologique*.

Une telle vision dichotomique, à laquelle nous avons des difficultés à adhérer, permet toutefois d'enclore de manière artificielle les phénomènes linguistiques dans des domaines de connaissance, de chercher à les maîtriser lorsqu'il s'agit d'élaborer des systèmes automatiques qui s'appuient sur la réalité apparente, textuelle et descriptible de la langue. Il n'en demeure pas moins que seul un expert du domaine possède la compétence pour identifier précisément quels sont les termes de ce domaine à partir d'un texte qu'on lui fournit. Il convient d'admettre qu'il n'existe pas de modèle linguistique ou statistique capable de pronostiquer si un syntagme nominal est un terme, le résultat d'un acte de dénomination motivé. Car le terme est un objet sociolinguistique, il est défini à l'intérieur et pour l'usage d'une communauté scientifique. Cette dimension sociale et vivante du terme peut même se fonder jusque dans l'expérience professionnelle d'un seul individu [10]. Ce caractère consensuel et sociolinguistique du terme fait qu'il n'y a pas de «continuum» entre la description linguistique d'un terme et sa réalité terminologique. Il y a une part d'arbitraire dans le choix de la forme précise du terme au sein des dénominations possibles, lorsqu'on l'examine d'un point de vue strictement linguistique, indépendamment de toute connaissance de domaine. Il en résulte que si un terme peut être décrit d'un point de vue linguistique et formel, on ne peut décider d'après ce point de vue si un syntagme est un terme : *un modèle descriptif linguistique ne peut pas isoler une réalité terminologique.*

La deuxième raison tient aux fondements épistémologiques de la terminologie, «science» des termes. Les pratiques terminographiques conformes à la doctrine terminologique héritée de Wüster [W8̈1] et à partir de laquelle les normes ISO en vigueur ont été fixées, sont coûteuses et ne sont pas adaptées aux nouveaux cycles de production de l'information, ni aux besoins applicatifs qui en découlent. Cela ne signifie pas que ces pratiques – basées sur un travail lexicographique manuel– soient mauvaises, elles deviennent simplement inadaptées lorsqu'elles sont confrontées aux besoins et aux exigences informatiques. Comme l'explique Monique Slodzian [Slo95, Slo94, Slo93], la cause de ces problèmes réside moins dans les pratiques terminographiques et leurs difficultés méthodologiques que dans la doctrine terminologique même et ses fondements épistémologiques.

Il est montré dans [Slo95] que la terminologie est le vestige d'une conception scientifique universaliste et positiviste. La doctrine terminologique (la VGTT, Vienna General Theory of Terminology) est fondée en 1931 par Wüster et se définit elle-même comme une entreprise empiriste et positiviste. Elle naît dans un contexte où plusieurs tentatives de création de langues artificielles avaient déjà eu lieu. L'idée était de créer un système sémiotique optimal entièrement fondé sur la logique, une langue artificielle pour diffuser le savoir scientifique, une langue exempte de polysémie. L'unité

10. Par exemple «cycle à gaz» est un terme du thésaurus EDF. Nous avons joint la personne qui l'a créé pour lui demander la signification du terme. «Cycle à gaz», validé par le comité thesaurus, a été préféré à «cycle du gaz» pour éviter la confusion avec le cycle du gaz naturel. Le terme est utilisé dans un contexte de palier à gaz, palier hydrodynamique de compression. Notre avis est qu'il s'agit là d'une anomalie sémantique exploitée pour éviter une confusion avec un autre concept. On peut vraiment parler de dénomination motivée.

minimale en était le terme, précis et monoréférentiel.

La science des termes repose ainsi sur un vision mécanique et réductrice du couplage concept—mot [Slo95], elle propose un modèle qui ne prend pas en compte la polysémie naturelle de la langue et fait des notions et concepts des objets presque ordinaires, objectivables. Parce que les notions subissent une «logicisation[11]», les phénomènes langagiers ne sont pas pris en compte. Pour Wüster, la connaissance est objectivable à travers une terminologie. Pour reprendre le terme de M. Slodzian, l'«*objectologie*» de Wüster correspond ainsi au recensement cumulatif des objets et des concepts par domaine ou discipline scientifique, à partir de leur existence linguistique sous la forme écrite, les sources que le terminologue dépouille. Cette conception encyclopédiste fait de la connaissance un processus cumulatif plutôt que synthétique : domaine après domaine l'univers peut être décrit. De plus, le terme est inséré dans une vision duale de la connaissance selon laquelle il y aurait d'un coté les concepts scientifiques qui ont des réalisations linguistiques monoréférentielles et de l'autre les concepts de la vie quotidienne, qui ont des réalisations linguistiques polysémiques.

Face à l'absence d'une théorie de la connaissance unifiant les différentes sortes de concepts (scientifiques, techniques, philosophiques, moraux, familiers), la terminologie ne peut appréhender les concepts dans leur dimension psychologique, cognitive et leur préfère une existence «objectivée» dans la langue : «*Les notions ne doivent pas être comprises comme des "unités noétiques", plus ou moins obscures, mais comme des synthèses d'énoncés se rapportant à des objets donnés. Les notions constituent notre savoir sur l'univers, autrement dit, elles réunissent nos connaissances concernant les objets de l'univers.*» [Dah81]. Une notion est définie ici comme la synthèse d'énoncés sur un objet. La notion est formée de matériau linguistique, elle ne semble pas exister à l'extérieur de la langue. Derrière cette définition apparaît un aspect de la méthodologie du dépouillement terminologique, qui cherche à rassembler des précurseurs de définition des termes [Gou94] à partir d'énoncés collectés : contextes définitoires, explicatifs, associatifs [Gou93]. L'ensemble des notions, «la connaissance», est alors considéré comme un ensemble de synthèses d'énoncés. Il en résulte une confusion entre la connaissance et le medium linguistique qui permet de la faire naître ou revivre en soi, ce qui conduit à un savoir de type compilatoire, ayant pour équivalent une certaine quantité d'informations.

Aujourd'hui le concept est réhabilité en tant que constituant fondamental de la pensée par les sciences cognitives [Slo95]. La prise en compte de la dimension cognitive à l'oeuvre dans la construction du savoir a fait disparaître le clivage entre «savoir ordinaire» et savoir scientifique. Mais la logique reste la clef de décryptage de relations entre les concepts. Comme le signale F. Rastier [Ras91], les technologies

11. On trouve dans [Gou93] et dans [Dah81] des indications sur les méthodes de réduction des notions à une représentation logique. Les caractères notionnels (propriétés des objets décrits) doivent être listés et typés. Les rapports logiques entre les notions sont définis sur des opérations ensemblistes faites à partir des caractères typés (identité, implication, intersection, disjonction, négation). Les rapports entre les notions sont aussi de type hiérarchique (inclusion), méronymique-holonymique et fonctionnel. Outre les caractères notionnels, les notions se voient également attribuées un type, leur type notionnel : OBJET, PHÉNOMÈNE, ÉTAT, ACTION, PROPRIÉTÉ, RELATION ...

de l'IA, discipline fondatrice de la recherche cognitive, imposent un modèle logique et «*on assiste à une technologisation des théories*».

Monique Slodzian souligne que l'empirisme logique est maintenant dépassé et définit la situation actuelle ainsi : *Le travail scientifique est considéré comme en grande partie constitué par du langage, ou plus spécialement du texte, et la connaissance scientifique est elle-même considérée comme une information conceptuelle obtenue à partir des textes.*

Cette description correspond à la réalité des recherches actuelles en linguistique-informatique et en informatique documentaire. Mais est-elle au fond si différente de la conception terminologique? Certes, elle abandonne la contrainte logique pour une contrainte linguistique de langue écrite — on cherche à prendre en compte des phénomènes linguistiques symboliques ou quantitatifs, cependant elle reste clairement empirique : le texte induit de la connaissance. De l'empirisme logique, on est passé à un empirisme linguistique voire simplement sémiotique : les liens logiques sont remplacés par des associations linguistiques ou des relations de proximité graphique entre les signes.

Une telle conception de la connaissance repose sur la forme terminale que prend la connaissance à notre époque : la langue écrite, reflet partiel de la réalité de la langue et des processus cognitifs. C'est encore une conception du savoir «objectivé» dans la langue, avec en plus un corrélat numérique : le savoir appréhendé à travers les textes, les mots, peut être décrit sous une forme binaire, manipulable par des ordinateurs.

De l'appréhension de la connaissance comme des synthèses d'énoncés en terminologie, on est passé à l'appréhension de la connaissance comme une certaine quantité de textes en informatique documentaire. De fait, le principe du «recensement» de la connaissance à partir d'un support textuel, adopté par la terminologie, reste aujourd'hui le seul applicable en informatique documentaire. Dans certains secteurs d'activité (autres que la traduction de qualité), la participation du terminologue-lexicographe est minimisée ou supplantée par les techniques statistiques ou de reconnaissance de formes linguistiques : la transition textes—information conceptuelle est désormais assurée par une procédure automatique. Mais ces conceptions et ces méthodes de travail correspondent plus à des impératifs dictés par les récentes innovations technologiques que par des révélations scientifiques sur la nature profonde de la connaissance. La nécessité de trouver des solutions techniques impose un réductionnisme informatique, qui convertit la connaissance en information.

La science des termes n'a pas su répondre de manière satisfaisante à des questions fondamentales comme : «*Qu'est-ce qu'un terme? Qu'est-ce qu'un domaine?*». Elle n'a pas donné de solution à l'énigme de la nature des concepts. L'investigation qui a été faite des concepts et des notions s'est appuyée sur des méthodes utilisées par la science matérialiste, fondée à partir de l'observation du monde sensible (sciences naturelles, sciences physiques). Mais les concepts font-ils partie du monde sensible? L'analyse logique (décomposition d'une notion en caractères, définition d'opérations ensemblistes sur les caractères notionnels), la taxonomie, et l'idée selon laquelle la connaissance scientifique provient seulement d'un raisonnement logique n'ont pas

su élaborer une théorie de la connaissance capable d'éclaircir les rapports entre les concepts et la langue.

Une dernière raison pour laquelle nous sortons du cadre terminologique tient à l'importance que nous souhaitons donner à l'utilisateur dans le système d'indexation. Nous cherchons à montrer qu'il est possible de définir plusieurs points de vue sur un même document, en ne s'appuyant pas nécessairement sur les termes homologués présents dans ces documents. Nous souhaitons donner la possibilité à l'opérateur humain d'intervenir directement sur le jugement d'intérêt qu'il porte sur la forme des marques textuelles qui constituent ses indices de recherche dans un document. Cela revient à construire en quelque sorte des index personnalisés.

1.3.2 Aperçu de notre système

L'approche que nous proposons est comparable à une forme d'indexation libre sur des groupes nominaux complexes (les unitermes ne sont pas traités en raison de leur forte polysémie potentielle) : il n'y a pas de liste de contrôle, mais un contrôle linguistique des formes retenues. Les groupes nominaux retenus sont considérés comme des index libres ou peuvent être introduits après validation dans un vocabulaire contrôlé.

Autrement dit, il s'agit d'une indexation libre contrainte avec des informations morphologiques, syntaxiques et sémantiques. Ces contraintes linguistiques s'organisent sous la forme d'une «grammaire» du syntagme nominal qui s'accommode à l'objectif à atteindre. La possibilité d'adapter le système est assurée par un système d'apprentissage qui cherche à induire des morceaux de grammaire à partir d'une généralisation des propriétés linguistiques présentes dans des échantillons qu'on lui fournit. Ainsi la forme des objets à extraire n'est pas figée dans une grammaire d'extraction. Elle dépend d'un profil fourni par l'utilisateur qui fait intervenir sa compétence (de documentaliste, de terminologue, ou autre) en amont de l'extraction ou du filtrage. Les profils peuvent être constitués simplement en tirant parti de ressources déjà existantes (thesaurus, listes d'index) et peuvent bénéficier d'une mise à jour incrémentale.

Exemples de résultats

Nous donnons maintenant un exemple de résultats de filtrage des groupes nominaux sur un petit morceau de notre corpus. L'extrait qui est reproduit ci-après n'a pu être donné que sous une forme lemmatisée. En raison de disparition de matériel à la DER – et notamment de disques durs, nous n'avons pas eu la possibilité de récupérer dans un délai raisonnable le texte dans sa forme originale. L'extrait provient de sorties d'analyses dans lesquelles ont été effacées les informations de catégorie, de flexion et de structure syntaxique. Ainsi on n'y trouvera que des verbes à l'infinitif et des noms au masculin singulier. Ce type de représentation a le mérite de mettre en évidence les erreurs de lemmatisation. Ces résultats seront commentés au chapitre 8

en 8.2.

Extrait du corpus

DES MESURE DE TEMPERATURE ET DE ROTATION AVOIR ETE EFFECTUE PENDANT LE SOUDAGE. DES MESURE DE CONTRAINTE RESIDUEL PAR DIFFERENT METHODE AVOIR ETE AUSSI EFFECTUE. UNE SYNTHESE BIBLIOGRAPHIQUE SUR LES MESURE DE CONTRAINTE RESIDUEL PAR DIFFRACTION NEUTRONIQUE AVOIR ETE REDIGE. LES MESURE PAR DIFFRACTION X SUR LES ASSEMBLAGE SOUDE PAR FRICTION AVOIR ETE REALISE.
EFFET DE LES CONTRAINTE RESIDUEL SUR LA TENUE MECANIQUE DE LES COMPOSANTS.
LES ASSEMBLAGE HOMOGENE ET HETEROGENE AVOIR ETE FABRIQUE. LE TRAITEMENT THERMIQUE ET LES CARACTERISATION DE MATERIAU AVOIR ETE AUSSI EFFECTUE. UN PROGRAMME DE ESSAI ET DE CALCUL AVOIR ETE DEFINI.
OBJECTIF ET PRINCIPAL ETAPE DE LA ANNEE 1995. MESURE DE CONTRAINTE RESIDUELLES. ON UTILISER EVENTUELLEMENT DIFFERENT METHODE. CES ETUDE SE FAIRE EN RELATION AVEC LE CEA DANS LE CADRE DE LES FICHE BIPARTITE 2435. ON EXPLOITER LES RESULTAT FOURNI PAR LES MESURE SUR DES PLAQUE REVETU EFFECTUE EN 1994. LA COMPARAISON DE LES CONTRAINTE OBTENU PAR DIFFERENT METHODE SUR DES MAQUETTE IDENTIQUE FOURNIR DES INDICATION UTILE SUR LA FIABILITE DE CES METHODE. ON DEVOIR POUVOIR DETERMINER NOTAMMENT SI LES MESURE AINSI OBTENU POUVOIR SERVIR DE REFERENCE A LES CALCUL. ON POURSUIVRE LA MISE AU POINT DE UNE METHODE DE MESURE DE LES CONTRAINTE RESIDUEL DANS LA EPAISSEUR DE LES COMPOSANT PAR DIFFRACTION NEUTRONIQUE. LES MESURE ETRE EFFECTUE A LE LLB SACLAY. DES MESURE ETRE EFFECTUE SUR DES MAQUETTE PLAN BIMETALLIQUE ET SUR LES ASSEMBLAGE SOUDE PAR FRICTION. EN PEAU EXTERNE ET EN PEAU INTERNE APRES DECOUPE.
ON EFFECTUER 2 ESSAI DE FLEXION AVEC DES JOINT SOUDE PAR FRICTION. ON ANALYSER LES RESULTAT ET ON FAIRE DES CALCUL CORRESPONDANT. OBJECTIF ULTERIEUR : L'OBJECTIF ESSENTIEL ETRE LA ETUDE DE LE ROLE DE LES CONTRAINTE RESIDUEL SUR DES PHENOMENE TEL QUE LA RUPTURE, LA FATIGUE ET LA CORROSION SOUS CONTRAINTE. IL FALLOIR POUR CELA DISPOSER DE OUTIL NUMERIQUE PERMETTANT DE INTRODUIRE DES CHAMP DE CONTRAINTE RESIDUEL COMPLET A PARTIR DE QUELQUES MESURE PONCTUEL, CES CHAMP POUVANT ETRE SUPERPOSE ALORS A LES CHARGEMENT EXTERNE LORS DE CALCUL A LES ELEMENT FINI. CES OUTIL ETRE DEVELOPPE DANS LE CADRE DE UNE THESE. POUR ATTEINDRE CE OBJECTIF, IL APPARAITRE DE PLUS EN PLUS NECESSAIRE DE AUGMENTER LES COMPETENCE EN MESURE DE CONTRAINTE, NOTAMMENT EN COUPLANT LES APPROCHE EXPERIMENTAL ET NUMERIQUE.

Groupes nominaux retenus par le profil I

Le filtrage des groupes nominaux extraits du texte précédent a permi de retenir 32 groupes nominaux qui sont listés ci-dessous. Le profil I est défini en 2.2.1 et en 7.2.5. Il a été conçu pour retenir des groupes nominaux utilisés pour la veille technologique.

MESURE DE TEMPERATURE
MESURE DE ROTATION
MESURE DE CONTRAINTE RESIDUEL
MESURE DE CONTRAINTE RESIDUEL PAR DIFFRACTION NEUTRONIQUE
MESURE PAR DIFFRACTION X SUR LES ASSEMBLAGE SOUDE PAR FRICTION
EFFET DE LES CONTRAINTE RESIDUEL SUR LA TENUE MECANIQUE DE LES COMPOSANTS
ASSEMBLAGE HOMOGENE
ASSEMBLAGE HETEROGENE
TRAITEMENT THERMIQUE
CARACTERISATION DE MATERIAU
MESURE DE CONTRAINTE RESIDUELLES
FICHE BIPARTITE 2435
COMPARAISON DE LES CONTRAINTE
CONTRAINTE RESIDUEL
EPAISSEUR DE LES COMPOSANT
DIFFRACTION NEUTRONIQUE
MAQUETTE PLAN BIMETALLIQUE
LLB SACLAY ASSEMBLAGE SOUDE PAR FRICTION
PEAU EXTERNE
PEAU INTERNE APRES DECOUPE
ESSAI DE FLEXION
JOINT SOUDE PAR FRICTION
CONTRAINTE RESIDUEL SUR DES PHENOMENE
CORROSION SOUS CONTRAINTE
OUTIL NUMERIQUE
CHAMP DE CONTRAINTE RESIDUEL COMPLET
CHAMP EXTERNE
CALCUL A LES ELEMENT FINI
MESURE DE CONTRAINTE
APPROCHE EXPERIMENTALE
APPROCHE NUMERIQUE

Groupes nominaux retenus par le profil II

Le profil II, défini en 2.2.2 et 7.2.5, a un pouvoir filtrant beaucoup plus important. Il a retenu seulement 10 groupes nominaux d'après ses critères de pertinence.

MESURE DE TEMPERATURE
MESURE DE CONTRAINTE
CARACTERISATION DE MATERIAU
CORROSION SOUS CONTRAINTE
ESSAI DE FLEXION
PROGRAMME DE CALCUL
PROGRAMME DE ESSAI
METHODE DE MESURE
ELEMENT FINI
CHAMP DE CONTRAINTE

1.4 Organisation de la thèse

Le chapitre 2 est consacré à la notion de syntagme nominal pertinent. Nous cherchons à montrer que cette caractérisation documentaire du syntagme doit nécessairement prendre appui sur une description linguistique et notamment sémantique. Les descriptions sont rassemblées dans des profils qui constituent autant de points de vue différents sur un document. Dans le chapitre 3 nous exposons sur quelles ressources existantes en matière de traitement automatique du langage nous avons pris

appui. Nous décrivons également celles que nous avons dû élaborer pour résoudre le problème du filtrage de syntagmes nominaux pertinents. Nous terminons ce chapitre par une présentation de notre corpus de travail. Au chapitre 4, nous présentons nos choix en matière d'approche syntaxique et sémantique. Le chapitre 5 montre comment nous avons exploité sur le plan technique les approches syntaxique et sémantique abordées au chapitre 4 pour enrichir le texte d'informations linguistiques. Ensuite le chapitre 6, étant donné les options choisies pour les traitements syntaxique et sémantique, expose la solution retenue pour filtrer les syntagmes nominaux. Le chapitre 7, présente la procédure d'apprentissage utilisée pour construire les filtres appliqués aux syntagmes nominaux. Enfin le chapitre 8 montre les résultats des deux profils de filtrage qui ont été définis.

Chapitre 2

La notion de syntagme nominal pertinent

2.1 Pertinence documentaire lors des phases de recherche et d'indexation

2.1.1 Définition

En documentation est pertinent un document qui satisfait une requête documentaire. Le document doit apporter des éléments de réponse à la requête. La requête définit par son contenu sémiotique un certain champ d'investigation dont les marques de présence vont être recherchées dans les documents. Ainsi, la pertinence documentaire est établie pour la recherche documentaire à partir d'un rapport de satisfaction : combien de réponses données satisfont-elles les requêtes qui lui ont été adressées? Lors d'une indexation libre, qu'elle soit plein-texte, sur les lemmes ou sur les groupes nominaux, hormis les pondérations statistiques, on indexe sans poser le problème de la pertinence des index. Nous pensons qu'il est possible de prédéterminer dans une certaine mesure la forme des index[1]. Plusieurs groupes d'index, qui constituent autant de points de vue différents sur un même document, peuvent alors être générés. Nous souhaitons définir la pertinence documentaire d'un syntagne nominal par rapport à l'intérêt d'un individu pour un certain sujet ou l'objectif d'une application : «est pertinent un syntagme qui est porteur d'information par rapport à un certain champ d'investigation». Nous faisons l'hypothèse que **la pertinence d'un syntagme nominal peut être largement évaluée à l'aide de critères linguistiques**. Nous ne rejetons cependant aucunement l'intérêt des critères statistiques d'évaluation qui ont prouvé leur efficacité (comme la fréquence, la densité, et la répartition des candidats termes dans le corpus [Sta95]) pourvu que le corpus respecte une taille minimale. G. Grefensette [Gre94] montre notamment qu'une taille minimale de corpus est à respecter et qu'il faut choisir la méthode statistique en fonction de la taille du corpus.

1. Plus précisément que par l'usage de patrons syntaxiques

Nous optons d'emblée pour une solution symbolique du problème – utilisation d'informations linguistiques – plutôt que statistique, car nous voulons traiter des documents sans contrainte de taille. Ceci n'est pas possible avec les indicateurs statistiques : par exemple, un document de quelques lignes ne permettra pas de calculer des fréquences pertinentes. Pour déterminer la forme d'un syntagme pertinent, nous avons donc à considérer simultanément :

- Une description linguistique générale des syntagmes nominaux,
- Le but visé par l'application ou l'utilisateur, qui détermine les descriptions linguistiques «intéressantes».

2.1.2 Intérêt des profils de pertinence

Pour montrer l'intérêt de la notion de profil de pertinence, considérons trois individus qui interrogent un corpus traitant de matériel nucléaire (semblable à notre corpus EDF-ARD), pour lequel on ne dispose pas de thesaurus. La première personne cherche les types de métaux utilisés pour les cuves de réacteurs nucléaires : son chef de service lui a demandé de constituer une terminologie sur les alliages métalliques industriels. La seconde personne cherche des indices de références des documents cités dans le texte (rapports, notes, normes, états d'avancement) afin d'évaluer les contributions internationales dans le domaine. Enfin, la troisième personne est informaticienne et recherche tout ce qui concerne les techniques de modélisation (calculs, modèles).

On se rend compte d'après cet exemple fictif mais plausible que le caractère informatif du SN varie suivant l'exploitation qui en est faite, et que la pertinence du syntagme aurait intérêt à être définie pour une tâche particulière. Ainsi la première personne recherchera en priorité des noms de substances métalliques, la seconde recherchera des noms d'objets sémiotiques, la troisième s'intéressera à des noms de processus (mais pas seulement).

2.2 Deux profils de pertinence

Pour tester la notion de profil, nous appliquerons à un même corpus (défini au chapitre 3.2) deux profils de pertinence. Chaque profil est un point de vue particulier sur le document. Il est défini par un individu pour son usage personnel ou pour un groupe d'utilisateurs. Il est constitué d'un ensemble d'informations linguistiques (lexicales, morphologiques, syntaxiques et sémantiques) combinées entre elles (voir section 7.2.3). Il permet de distinguer les syntagmes nominaux non pertinents et pertinents, conformément aux exemples qui ont permis de le définir et qui sont placés dans des échantillons d'apprentissage.

2.2.1 Profil I : des syntagmes nominaux pour l'indexation libre et la veille technologique

Parmi les activités de consultant en information à la DER d'EDF, on trouve la veille technologique et le classement automatique de documents. Nous avons confié la réalisation de ce profil à un consultant en information intéressé par une extraction de groupes nominaux pertinents pour la veille technologique. Sa démarche habituelle consiste à soumettre des textes variés à un système d'extraction de groupes nominaux. Grâce à une boîte à outils nommée KES [OHD94, Her95] (voir page 3), le consultant sélectionne les groupes nominaux jugés les plus pertinents. Ces derniers, considérés comme des index caractérisant les documents, seront utilisés dans des procédures de classification automatique de documents. Cette phase de sélection–organisation, la plus coûteuse en temps, est manuelle ou semi-automatique (application de règles) et c'est celle-ci que le consultant souhaiterait sauter en recueillant du système d'extraction des syntagmes déjà pertinents. Nous lui avons donc proposé de constituer un profil de pertinence en fournissant au système un ensemble de syntagmes validés et éliminés par la méthode habituelle. Cet échantillon a été ensuite complété manuellement grâce à une application conçue dans ce but (voir en annexe C.2).

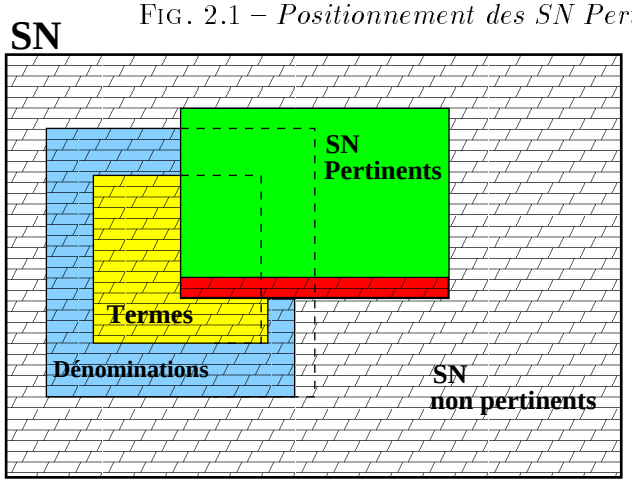
2.2.2 Profil II : des syntagmes nominaux pertinents pour le terminographe

Le deuxième profil de test vise à définir la pertinence des syntagmes plus strictement : est pertinent un syntagme nominal qui est un candidat terme pour un certain domaine d'activité. Comme nous n'avons pas pu faire appel à un terminologue EDF (les derniers sont partis à la retraite), nous avons extrait du thesaurus EDF la terminologie de deux domaines d'activités : APPAREILLAGE MÉCANIQUE (250 termes), SCIENCES PHYSIQUES (330 termes). Ces termes ont servi de matière à la construction d'un profil d'extraction de syntagmes nominaux ayant statut de candidats termes pour ces domaines d'activité. Cependant face à des difficultés, nous avons dû élargir la couverture de ce profil. Nous reviendrons au chapitre 7 sur la constitution de ce profil qui a posé quelques problèmes.

2.3 Caractérisation linguistique des SNP

Le Syntagme Nominal Pertinent (SNP) est un objet documentaire que l'on cherche à caractériser linguistiquement. Il est important pour nous que les SNP puissent être décrits sous forme de propriétés linguistiques codables et combinables. Sans cela, il ne serait pas envisageable de les identifier automatiquement.

FIG. 2.1 – *Positionnement des SN Pertinents*



2.3.1 Réalité linguistique de l'objet documentaire SNP

Le SNP est un objet documentaire, mais il n'est pas pour autant une réalité linguistique autonome ou isolable. Comme nous allons le voir, le SNP recouvre une pluralité d'objets linguistiques dont les caractéristiques ne convergent pas nécessairement en un modèle unifié.

Considérons l'ensemble des syntagmes nominaux d'un texte. Parmi ceux-ci on trouve des séquences ordinaires et d'allure dénominative. Ces dernières recouvrent les termes homologués par les terminologues. La figure 2.1 montre, parmi l'ensemble des SN d'un document (surface hachurée) et pour un certain profil de pertinence (un certain *a priori* sur ce qui est intéressant ou pas), un sous-ensemble de syntagmes nominaux considérés comme pertinents (fenêtre non hachurée). Cet ensemble de SNP recouvre tous les types de syntagmes (syntagmes «ordinaires», syntagmes d'allure dénominative, dénominations et termes homologués).

La question des séquences d'allure dénominative

Les séquences d'allure dénominative, qui peuvent représenter une large part des SNP dans un texte, sont définies par Kleiber [Kle84]. Il oppose relation de dénomination et relation de désignation entre un signe et un objet. Il définit la relation de dénomination comme *l'institution entre un objet et un signe d'une association référentielle* ***durable****. Cette association référentielle n'a pas pour but une désignation uniquement momentanée, transitoire et contingente de la chose, mais au contraire de l'établissement d'une règle de fixation référentielle qui permet l'utilisation ultérieure du nom pour l'objet dénommé.* Une autre caractéristique des dénominations est leur caractère **codé** : *Elles présupposent un* ***codage*** *antérieur à la première introduction de la séquence (par périphrase définitionnelle ou ostension).* En opposition, la relation de désignation autorise des expressions qui ne supposent aucun codage antérieur et n'impliquent pas un lien référentiel stable. La dénomination répond au

besoin de pouvoir référer à un élément du réel de manière durable par un signe. A ce titre le terme est une dénomination. Et comme en terminologie seul l'humain peut attester de l'existence d'une notion dans un domaine, en matière de dénomination, seul le jugement humain peut confirmer la stabilité de l'association signe–objet. On peut toutefois mentionner des critères linguistiques qui permettent de supposer que certains syntagmes ne sont pas dénominatifs.

Soit le syntagme *jeu considérable*, l'adjectif *considérable* qui a valeur d'évaluation subjective classe ce syntagme dans la catégorie des séquences désignationnelles, créées pour un usage limité et qui ne vivent pas au delà de la phrase ou du paragraphe. Autre exemple : le syntagme *résultats obtenus*, construit avec un participe passé à valeur de résultat, ne peut prétendre à être dénominatif. Dernier exemple : *les résultats de cette étude* est un syntagme qui n'est certainement pas dénominatif non plus. Le nom *étude* est introduit par un adjectif démonstratif. Cette valeur déictique interdit au syntagme tout caractère codé. Maintenant considérons les trois syntagmes suivants : *jeu inter-membranaire, résultats officiels, résultat d'étude*. Leur caractère codé et dénominatif est plus affirmé, pourtant ils sont construits sur les mêmes têtes nominales. Le modifieur du nom et la détermination jouent donc un rôle déterminant dans l'évaluation du caractère dénominatif.

L'évaluation du caractère dénominatif du syntagme est un problème, mais déterminer si une séquence, dénominative ou non, est pertinente est un autre problème qui est indépendant du premier. En demandant à l'opérateur de définir ce qui est pertinent et ce qui ne l'est pas, nous contournons le problème de la reconnaissance des séquences dénominatives. La tâche de reconnaissance des SNP est alors réduite à l'identification de formes conformes à une description existante et validée. Il n'est donc jamais question d'établir si telle ou telle séquence est dénominative ou ne l'est pas, le caractère dénominatif n'impliquant pas nécessairement la pertinence du syntagme et inversement.

Relativité de la notion de pertinence

Les tables 2.1 et 2.2 montrent des exemples de syntagmes nominaux considérés pertinents ou non relativement aux profils I et II. En table 2.1, dans la colonne des syntagmes pertinents, on trouve des termes attestés dans le thesaurus EDF (*réacteur à neutron rapide*), des syntagmes d'allure dénominative (*joint multi-passe*) et des syntagmes qui ne sont ni terminologiques ni dénominatifs (*aciers russes irradiés*). Dans la colonne des syntagmes classés non pertinents, on trouve également des termes attestés dans le thesaurus EDF (*groupe de travail, document de référence*) et des syntagmes ni terminologiques ni dénominatifs (*rédaction de la note de synthèse, les résultats de cette étude*). Si l'on compare le contenu des tables 2.1 et 2.2, il apparaît que le profil II a opéré un filtrage plus sévère. Des syntagmes retenus comme pertinents par le profil I ont été écartés dans le profil II : seul *corrosion sous contrainte* a été retenu parmi tous les syntagmes d'allure dénominative (La comparaison du pouvoir de filtrage des deux profils est faite au chapitre 8).

TAB. 2.1 – *Exemples de syntagmes nominaux pertinents ou non d'après le profil I*

Type d'objet	Pertinents	Non pertinents
Termes attestés	réacteur à neutron rapide, réacteur à eau pressurisée, cuve REP	document de référence, base de données, groupe de travail
Syntagmes d'allure dénominative	joint multi passe, plaque en acier ferritique, corrosion sous contrainte, maquette plane bimétallique	—
Syntagmes nominaux ni terminologiques, ni dénominatifs	mesure de teneur en ferrite, ?aciers russes irradiés	rédaction de la note de synthèse, les résultats de cette étude, seule voie possible

TAB. 2.2 – *Exemples de syntagmes nominaux pertinents ou non d'après le profil II*

Type d'objet	Pertinents	Non pertinents
Termes attestés	réacteur à neutron rapide, réacteur à eau pressurisée, cuve REP	document de référence, base de données, groupe de travail
Syntagmes d'allure dénominative	corrosion sous contrainte	joint multi passe, plaque en acier ferritique, maquette plane bimétallique
Syntagmes nominaux ni terminologiques, ni dénominatifs	–	mesure de teneur en ferrite, ?aciers russes irradiés, rédaction de la note de synthèse, les résultats de cette étude, seule voie possible

2.3.2 Modélisation de la pertinence à l'aide d'informations linguistiques : des SNP existants aux SNP possibles

Notre postulat de départ est le suivant : **la pertinence d'un syntagme ne peut pas être déterminée automatiquement**, elle est toujours déterminée par l'utilisateur humain qui sait ce qu'il cherche. Pour aider la machine, cet utilisateur donnera un certain nombre d'exemples des syntagmes qu'il juge pertinents et non pertinents. Il s'agit alors d'exprimer voire de modéliser, à l'aide d'informations linguistiques, la différence entre les objets pertinents et non pertinents pour l'utilisateur. De cette manière on doit être capable de prédire la pertinence d'une forme qui ne ferait pas partie des exemples fournis par l'utilisateur, c'est-à-dire extrapoler les formes possibles à partir des formes observées.

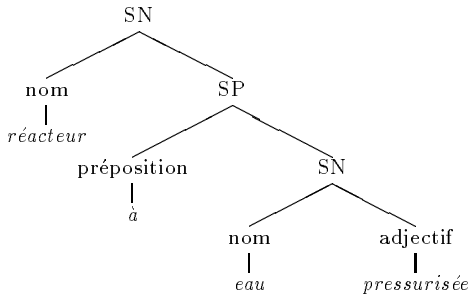
Description générale du groupe nominal

Le matériel lexical utilisé dans la formation des groupes nominaux est l'information distinctive la plus importante. On pourrait décrire les exemples de groupes nominaux pertinents et non pertinents donnés par l'utilisateur comme des listes de compatibilités lexicales. Par exemple *résultats* suivi de *obtenus* serait déclaré non pertinent, *réacteur* suivi de *à*, suivi de *eau*, suivi de *pressurisée* serait déclaré pertinent. Mais une telle description permettrait difficilement d'extrapoler des formes possibles à partir des formes existantes. Pour cela nous avons besoin d'une syntaxe du groupe nominal qui mette en relation les éléments lexicaux et qui identifie leurs positions sur l'axe syntagmatique. Tout groupe nominal, qu'il soit dénominatif, terminologique ou ordinaire peut être décrit en terme de dépendances lexicales exprimant des relations naturelles (et de compatibilité) entre les éléments lexicaux qui le composent. Ainsi dans *résultats obtenus*, nous dirons que le nom *résultats* est modifié par le participe passé *obtenus*; dans *réacteur à eau pressurisée*, nous dirons que le nom *réacteur* est modifié par le nom *eau* introduit par la préposition *à*, le nom *eau* étant lui-même modifié par l'adjectif *pressurisée*. La figure 2.2 montre deux représentations de ce syntagme, l'une en termes de constituants, l'autre en termes de dépendances naturelles. Nous ferons appel à ces deux représentations. Ces différences de formalismes syntaxiques sur lesquelles nous reviendrons en 4.1.3 sont exposées dans [Mel88b]. Nous généraliserons en considérant que le groupe nominal peut être décrit par un schéma {groupe régisseur, groupe régi}, le groupe régi pouvant être décrit récursivement par le même schéma[2]. Avec ce formalisme élémentaire, nous pouvons considérer le schéma *réacteur*$\longrightarrow$*à*$\longrightarrow$*eau*$\longrightarrow$ADJ et déclarer pertinents des modifieurs d'*eau* comme ADJ=$\{$*pressurisée, légère, lourde*$\}$.

La description du syntagme en terme de dépendances syntaxiques permet également la prise en compte de phénomènes d'insertion. Par exemple l'insertion d'un article dans *réacteur au graphite* ou d'un adjectif dans *réacteur à liquide organique*. Nous verrons au chapitre 5, comment ces variations sont prises en charge, grâce à une

2. Nous avons adopté ce type de description, qui outre son adéquation à la réalité, apporte des avantages en matière d'implémentation et de traitements informatiques.

FIG. 2.2 – *Représentation arborescente en constituants et représentation des dépendances syntaxiques dans le syntagme* «réacteur à eau pressurisée»



réacteur $\longrightarrow$ *à* $\longrightarrow$ *eau* $\longrightarrow$ *pressurisée*

décomposition du syntagme en relations syntaxiques élémentaires (voir paragraphe 5.2.1).

Création et modifications des syntagmes nominaux

Les syntagmes nominaux peuvent être créés de toute pièce (acte de dénomination) mais peuvent aussi être construits par composition avec d'autres syntagmes ou éléments lexicaux [Dai94, Jac97]. Il peut y avoir composition par juxtaposition, par exemple *aide à la décision* devient *système d'aide à la décision*, ou encore *image de marque* devient *image de marque de l'entreprise*. Il peut y avoir composition par recouvrement; ainsi à partir de *réseau de transmission* et *transmission de la voix*, peut être produit *réseau de* transmission *de la voix*. Il peut y avoir des modifications par insertions de modifieurs ou coordination; par exemple dans *direction de la production* **et** *du transport* seront reconnus *direction de la production* et *direction du transport*. Ou encore *haute tension* pourra être modifié par un adverbe de degré: *très haute tension*. Enfin, la mise sous forme d'abréviation et d'acronyme des noms de raisons sociales ou de concepts usuels est très fréquente, ainsi *IA* pour *Intelligence Artificielle* ou *EDF* pour *Électricité de France*. Ces acronymes sont couramment réinsérés dans des syntagmes plus complexes.

Catégorisation sémantique des composants

Si nous avons à comparer deux syntagmes de structure identique, nous pouvons remplacer leurs éléments sémantiquement similaires par leur hyperonyme commun. Soit les syntagmes *réacteur à eau* et *réacteur à graphite*, nous pouvons leur trouver un schéma commun qui est : *réacteur à* NOM-DE-MATIÈRE. Ce schéma, qui généralise la forme de deux syntagmes existants, permet de générer d'autres syntagmes sur le même modèle; par exemple *réacteur à uranium, réacteur à plutonium*. Cette

productivité ne garantit par l'acceptabilité des syntagmes générés. Mais si de tels syntagmes sont rencontrés dans le texte — ils sont *a priori* acceptables — le schéma en question permet de les classer comme pertinents ou non pertinents, selon les choix que l'utilisateur aura faits. La catégorisation sémantique n'est pas la seule manière de passer d'un syntagme existant aux syntagmes possibles. La suffixation peut aussi être utilisée (voir chapitre 5).

Condition de description du possible à partir de l'observé

Pour illustrer le fait que c'est par une description fine des séquences observées qu'un contrôle fin des séquences possibles peut être effectué, nous étudions maintenant à partir d'un certain nombre d'exemples tirés du thesaurus EDF, des séquences conformes au schéma N_1 À N_2. Si nous étudions le schéma N_1 À N_2 dans le thesaurus EDF, nous remarquons qu'il est productif et qu'il présente une très grande variété de configurations. L'attachement prépositionnel «À N_2» correspond à trois fonctionnements distincts :

1. La préposition introduit un modifieur adverbial figé, comme dans *filage à sec* ou *collage à chaud*. Ce comportement adverbial ne s'observe qu'avec des noms de processus ou d'activité, et permet de spécifier les conditions particulières dans lesquelles se déroule le processus ou l'activité.

2. La préposition introduit un nom objet comme dans *roue à blé* ou *brosse à main*. Dans ce cas l'attachement prépositionnel est en rapport avec un prédicat sous-jacent implicite qui met en relation le premier nom avec un objet qui lui est extérieur : *une roue qui sert à moudre du blé*, ou *une brosse pour nettoyer les mains*.

3. La préposition introduit un nom modifieur. Dans ce cas il n'y a pas de prédicat implicite, le deuxième nom vient modifier les caractéristiques référentielles du premier. C'est le cas de *perceuse à main* ou *roue à aube*.

Dans un certain nombre de cas, parmi lesquels nous excluons les cas de figements, le schéma N_1 À N_2 manifeste une forme de composionnalité, comme le montrent [BT91] et [Cad92] en dégageant deux schémas d'interprétation fondamentaux, qu'ils nomment «programmes», et qu'ils mettent en évidence par des tests syntaxiques :

N_1 avec N_2 Ce schéma peut être mis en évidence avec tous les emplois où la préposition *à* introduit un N_2 modifieur (*roue à aube*).

N_1 pour N_2 Ce schéma est valable pour tous les cas ou le N_2 est considéré comme un «objet» par rapport au N_1 (*roue à blé*).

Cependant cette forme d'interprétation compositionnelle, intelligible, n'est pas systématique. Elle correspond à la **vraie compositionnalité sémantique** comme la définit P. Cadiot [Cad92]:209. Elle s'observe par exemple dans *verre à pied* et

verre à dents. Et c'est seulement lorsque cette forme de compositionnalité s'observe – comme c'est le cas le plus fréquent dans les textes scientifiques et techniques – que nous sommes en mesure de construire un schéma des séquences possibles à partir des séquences observées : un schéma comme *réacteur à* NOM-DE-MATIÈRE produit des syntagmes cohérents parce qu'il est régi par un principe de compositionnalité. Si nous reprenons maintenant la classification de Cadiot dans [Cad92] et ses exemples, cela apparaît clairement; les types de construction se répartissent entre opacité sémantique et compositionnalité. Il définit un type de construction qui aboutit à une **opacité sémantique**, et où aucune compositionnalité ne se manifeste, comme par exemple dans le nom composé *pied-à-terre*. Ainsi **pied-à-mer* ou **cheville-à-terre* ne sont pas acceptables. Un autre type de construction aboutit à une **semi-compositionnalité métaphorique**. Par exemple *sac à vin, de la chair à canon*. Le caractère métaphorique est induit par la recatégorisation radicale du référent du premier nom : *sac*, nom d'artefact devient un humain et la *chair*, substance organique vivante est recatégorisée comme humain. Un autre type de construction aboutit à une forte **compositionnalité sans transfert métaphorique** comme dans *rouge à lèvres*. L'**altération référentielle** est un autre type de construction où le référent d'un des noms est altéré pour s'ajuster à l'autre. Par exemple, un *arbre à came* est arbre, mais pas au sens habituel. Un dernier type de construction relève d'un figement du statut sémiotique. C'est le cas de nombreux titres, ainsi «*un américain à Paris*», «*Les demoiselles de Rochefort*». Ces séquences observées ne sont absolument pas productives, il n'y aurait pas de sens à construire des syntagmes comme *un français à New-York* ou *Les messieurs de La Rochelle*.

Les configuration de N_1 À N_2 qui apparaissent dans le thesaurus sont en majorité compositionnelles, plutôt qu'«opaques». On trouve une unique occurrence du schéma NOM-DE-MATIÈRE À NOM-DE-MATIÈRE avec *fer à béton*, qui s'interprète comme «*fer* pour *béton*» : il s'agit de tiges de fer traversant et consolidant le béton coulé dans des coffrages. Le schéma NOM-DE-MATIÈRE À NOM-D'ARTEFACT par exemple *acier à outil* s'interprète de la même manière. Dans des syntagmes comme *frettage à froid, attente à chaud, arrêt à froid, essai à vide, fonctionnement à vide* (NOM-DE-PROCESSUS À NOM-D'ÉTAT[3]), l'attachement prépositionnel s'interprète comme un modifieur adverbial. Le schéma «N_1 pour N_2» est également commun aux syntagmes suivants : *cable à huile, cuve à lisier, canne à sucre*[4] (NOM-D'ARTEFACT À NOM-DE-MATIÈRE), *chambre à fusion, tube à luminescence* (NOM-D'ARTEFACT À NOM-DE-PROCESSUS/PHÉNOMÈNE). En revanche c'est le schéma «N_1 avec N_2» qui convient à des syntagmes comme *moteur à explosion, turbine à réaction , générateur à induction, embrayage à friction* (NOM-D'APPAREIL À NOM-DE-PROCESSUS/PHÉNOMÈNE), *robinet à soupape, moteur à collecteur* (NOM-D'APPAREIL À NOM-D'APPAREIL). De même «N_1 avec N_2» convient à *roue à aube* (NOM-D'ARTEFACT À NOM-D'ARTEFACT)

3. Le froid, le chaud, le vide, ... peuvent être considérés comme des noms de phénomènes physiques, des noms d'états physiques

4. Le cas de *canne à sucre* correspond à une composition puis à un transfert métaphorique. «Une plante qui ressemble à une canne lorsqu'elle est coupée et dont on extrait du sucre». De plante, la canne devient matière première.

et *hangar à poutrelle* (NOM-BÂTIMENT À NOM-D'ARTEFACT) et «N_1 pour N_2» pour *garage à bateau* (NOM-BÂTIMENT À NOM-DE-VÉHICULE).

Il ne faudrait pourtant pas en déduire que le type d'interprétation est prédictible à partir d'un schéma abstrait substituant aux lexèmes leur catégorie conceptuelle. Par exemple dans les syntagmes *roue à blé, ensacheur à lait, four à chaux, silo à cendre, enceinte à glace, contact à mercure, condensateur à eau, scie à métal, filtre à air* (NOM-D'ARTEFACT À NOM-DE-MATIÈRE), le schéma «N_1 pour N_2» convient, sauf pour *contact à mercure, interrupteur à mercure* où le schéma «N_1 avec N_2» s'applique (interrupteur qui fonctionne avec du mercure). Les noms dits abstraits posent aussi des difficultés. Dans *réglage à distance, vente à crédit* (NOM-DE-PROCESSUS À NOM-ABSTRAIT), la glose est aisée, mais l'interprétation en *pour* ou *avec* est problématique.

Il ressort des exemples de composés binominaux de type N_1 À N_2 que des régularités d'interprétation peuvent être mises en évidence et formalisées sous la forme de schémas abstraits. La prédictibilité de l'interprétation des schémas n'est pas garantie, mais ces derniers constituent une matière descriptive que nous exploiterons pour tenter de généraliser les différences entre syntagmes pertinents et non pertinents.

Nous ne cherchons pas à fournir une description linguistique suffisament complète pour garantir une interprétation correcte du syntagme. Il s'agit seulement de décrire les syntagmes pour les distinguer, s'ils sont pertinents, de syntagmes non pertinents, et réciproquement.

Les gloses que nous mentionnons parfois visent à souligner le caratère productif du syntagme ramené à une description plus abstraite et donc, l'intérêt qu'il y a le décrire de la sorte. Comme les précédents, les exemples suivants sont tirés du thesaurus EDF.

Rôle de la préposition dans l'interprétation des syntagmes

Outre le matériel lexical des groupes nominaux, nous accorderons une grande importance aux prépositions qui constituent un apport d'information distinctive dans l'expression de la différence entre un syntagme pertinent et un syntagme non pertinent. Les prépositions contribuent à fournir les clefs d'interprétation du mode de compositionnalité du syntagme, elles instaurent une certaine relation entre les lexèmes dotés d'un pouvoir référentiel.

Comme dans l'exemple précédent de la préposition *à*, d'autres prépositions peuvent manifester des schémas interprétatifs réguliers. C'est cette régularité, cette productivité possible, qui nous importe. Nous en donnons trois exemples avec les prépositions *sous, sur* et *en*.

La préposition *sous* permet d'établir une relation de localisation spatiale. Par exemple dans *culture sous serre*. Mais ce n'est pas la seule relation. Ainsi dans *malade sous perfusion* (NOM-HUMAIN SOUS NOM-PROCESSUS), *soudage sous argon* (NOM-PROCESSUS SOUS NOM-SUBSTANCE), on peut gloser par «N_1 est sous l'influence chimique de N_2». L'exemple *cable sous tube* donne à la préposition *sous* l'interprétation «est à l'intérieur de». Alors que dans *comportement sous pollution, soudage sous pression* (NOM-PROCESSUS SOUS NOM-PROCESSUS), la glose est «N_1 a lieu dans les

conditions imposées par N_2». Il y a une productivité paradigmatique pour chaque type d'interprétation (*sous pression, sous vide, sous contrainte ..., sous perfusion, sous valium ...*).

La préposition *sur* est en symétrie spatiale avec *sous*. Les exemples *travail sur écran, opération sur machine* (NOM-PROCESSUS SUR NOM-ARTEFACT) mettent en évidence une relation de localisation spatiale figurée : on ne travaille pas directement sur l'écran, mais le regard s'y porte pour travailler. En revanche, *transformateur sur socle* (NOM-ARTEFACT SUR NOM-ARTEFACT) s'interprète avec la relation spatiale conventionnelle. Dans *taxe sur salaire*, la relation spatiale est transposée au niveau abstrait en une relation de prélèvement (taxe prélevée sur le salaire).

Dans les exemples *tube en hélice, structure en épi, voûte en arc*, la préposition *en* suggère l'interprétation «N_1 a la forme de N_2» alors que dans les exemples *lait en poudre, béton en masse* la glose est «N_1 est présenté sous la forme de N_2». Dans *essai en laboratoire*, *en* a une valeur de localisation spatiale «à l'intérieur de» : *essai qui a lieu en laboratoire* et est productif : *en cabine, en chambre, en enceinte*. Dans *énergie en réserve*, le N_2 donne une indication qualitative sur l'état de N_1 : N_1 en {réserve, stock, pénurie, attente, progression, ...}.

Constructions binominales

Les constructions de type NOM NOM comme *tuyauterie vapeur*, fréquentes dans notre corpus technique, ont un caractère dénominatif plus marqué : leur caractère codé est manifeste. Il peut s'agir parfois d'abréviations. Plus encore que pour les exemples construits à partir des prépositions *à, sous, sur, en*, les exemples de constructions binominales montrent que la connaissance du domaine est indispensable pour l'interprétation du syntagme. Avec les syntagmes prépositionnels, le sens communément attribué aux prépositions et aux lexèmes n'est pas toujours suffisant pour résoudre le caractère codé de la dénomination. Avec les constructions binominales, cela est encore plus vrai car la préposition, indice d'interprétation, a disparu et cela laisse au lecteur plus de liberté ou de flou dans le choix de ses mécanismes d'interprétation. Sur le modèle de *débit turbine* sont construits *débit réseau, débit robinet, consommation gaz*. La préposition *de* semble avoir été omise par souci de brièveté. Dans *pylone chat, serre tunnel* c'est la forme ou le type du N_1 qui est caractérisé par le N_2.

La détermination du point de vue énonciatif

Après les prépositions, la détermination est un paramètre important de l'analyse des syntagmes nominaux. L'analyse de la linguistique énonciative offre une approche intéressante pour l'étude de la détermination nominale. Nous en énonçons les principes, tirés de la linguistique énonciative de A. Culioli qui sont expliqués dans l'ouvrage [BC90] : un même nom peut avoir tantôt les propriétés du discontinu (dénombrable), tantôt du continu dense, ou du continu compact (non dénombrable), lorsque ses propriétés notionnelles sont combinées avec celles d'un déterminant. Il

s'agit du discontinu lorsque les noms sont des unités discrètes et individuables, qu'on peut en compter les occurences et faire des prélèvements sur une classe construite. (par exemple : *un chien, deux chiens*). Il s'agit du continu dense lorsque l'on a affaire à du non-individuable sur lequel on peut quand même opérer un prélèvenent à l'aide d'un dénombreur (*une cuillerée de soupe, un bol de soupe*). Enfin, il s'agit du continu compact lorsqu'on a des prédicats nominalisés. Aucun prélèvement n'est alors possible, sauf en introduisant une propriété différentielle. La combinaison des propriétés du nom et des opérations de déterminations qui l'introduisent dans le texte lui permettent de prendre ces différentes valeurs (par exemple : *la santé -être sain, la pitié - avoir pitié*). Voici quelques exemples :

Je n'aime pas le lapin; le lapin	nourriture, continu dense
J'ai caressé trois lapins; lapins	discontinu
Tu veux du fromage? fromage	continu dense
J'ai fait trois fromages; fromages	discontinu
Il a une bonne santé; santé	continu compact
Donne-moi un peu de ta santé	«un peu de» : propriété différentielle : le continu compact devient dense.

Les valeurs des principaux déterminants Les articles *le, la, les* ont une même valeur généralisante dans certains contextes. Par exemple : *accident du travail, affectation au service publique*, ou encore :

Le lapin [ce qui est lapin] *est un animal vif.*
La propolis [ce qui est propolis] *a un pouvoir antibiotique.*
Les eaux profondes [ce qui est «eau profonde»] sont dangereuses.

L'article *un* + un nom singulier discontinu produit un prélèvement ou la validation d'une occurrence de la notion associée au nom. Par exemple : *propagation d'un feu.*

Dans les groupes nominaux, l'article zéro (absence de déterminant) renvoie toujours à la notion. Il s'agit d'une valeur qualitative sans aucune spécification de quantité. Par exemple : *accélérateur de particules, séparation isotopique par laser.*

Les déterminants sont donc un élément d'information important pour décrire les dépendances lexico-syntaxiques des syntagmes nominaux. Nous en ferons un usage systématique (voir paragraphe 5.2.1) en enregistrant le type de déterminant qui introduit le deuxième nom dans les dépendances élémentaires du type NOM_1 PREP NOM_2, y compris lorsqu'il s'agit d'adjectifs démonstratifs, indéfinis ou possessifs.

Impact de l'article défini dans les constructions N à N

L'article défini qui détermine le deuxième nom dans les constructions du thesaurus de type N_1 À LE N_2 modifie le schéma d'interprétation. Alors que l'absence d'article renvoie à la notion, comme dans *acier à silicium* (acier destiné pour la fa-

brication ou l'utilisation du silicium), l'utilisation de l'article extrait une occurrence de la notion plaçant les images référentielles des deux noms sur le même plan, ainsi dans *acier au silicium* (alliage), le silicium est mélangé à l'acier. L'interprétation est valable pour les syntagmes de type NOM-DE-MATIÈRE À LE NOM-DE-MATIÈRE : *acier au carbone, graisse au mobylène, verre au plomb*. Dans *filage au sec*, les conséquences de l'utilisation de la notion ou d'une occurrence de la notion comme modifieur sont visibles : dans *filage au sec* le filage se fait à l'abri de l'eau (d'une quantité d'eau) alors que dans *filage à sec*, le filage se fait sans utilisation d'eau (de la notion d'eau). Pour des schémas de type NOM-DE-PROCESSUS À NOM-DE-MATIÈRE/NOM-D'APPAREIL, l'usage de l'article est incontournable pour extraire un nom continu dense ou dénombrable : *brasage au fer* (le processus N_1 utilise la substance N_2), *cuisson au four, soudage au chalumeau, brasage au four* (le processus N_1 utilise l'outil ou l'appareil N_2). La présence de l'article n'est plus nécessaire si le deuxième nom est modifié par un adjectif : *cuisson à four chaud*.

Bilan

Nous avons énuméré un certain nombre de caractéristiques linguistiques des syntagmes nominaux. Nous feront appel à ces dernières pour chercher à modéliser la notion de syntagme nominal pertinent (SNP). Pour caractériser certaines constructions, nous avons fait appel à la notion de schéma d'interprétation. Ces schémas nous montrent qu'il y a des régularités d'interprétation identifiables dans des configurations linguistiques données, ce qui est le constat d'une certaine compositionnalité. Cependant à auncun moment il ne sera question pour nous d'interprétation automatique comme c'est le cas dans les travaux de C. Fabre [Fab96]. Dans [Fab96], une modélisation du groupe nominal est développée pour l'interprétation automatique de séquences binominales. Cette modélisation, plus complexe et complète que la nôtre, vise à rendre compte des phénomènes d'ajustements sémantiques entre les unités lexicales. Les descriptions linguistiques que nous utilisons visent seulement à décrire les compatibilités lexicales entre les unités.

2.4 Réagir vite au changement

Comme nous le soulignions en introduction, les textes n'ont pas toujours une valeur de référence en soi, leur intérêt n'est parfois qu'éphémère, ce caractère étant accentué par leur production toujours plus rapide et massive sur support électronique. Par exemple, les dénominations qui font leur apparition dans ces documents n'ont plus valeur de référence mais d'actualité, elles ne sont plus que la trace d'une activité ou d'un événement peut-être voué à l'oubli. Si l'on considère par exemple toutes les terminologies associées à l'évolution informatique de ces dernières années, nombre de termes et d'acronymes, parlants ou synonymes d'enjeux majeurs il y a 10 ans (*CP/M, DOS, GWBasic, CPU 8088* et *tube à vide* à l'époque de l'ENIAC...) n'ont plus aujourd'hui qu'un intérêt historique. Il faut donc relativiser la durabilité de l'association signe—objet extra-linguistique par rapport à la durée de vie du texte

(son actualité) et la durée d'existence des objets extra-linguistiques dénommés : dans le cadre de la technologie, c'est le plus souvent la disparition rapide de ces derniers qui entraîne la suppression de l'usage du signe linguistique associé. L'évolution technologique entraîne de nombreux actes de dénomination pour décrire des choses qui n'existaient pas peu de temps auparavant mais aussi la disparition de termes inusités.

Dans ce contexte, il faut pouvoir facilement et avec un faible coût redéfinir les points de vue sur les documents. Le système d'indexation doit réagir vite aux changements de contenu. Ainsi le caractère éphémère des signes et des référents donne un avantage aux techniques qui mettent en oeuvre de l'apprentissage et qui minimisent le codage lexical[5].

2.5 Compétence requise pour le repérage des SNP

La compétence requise pour identifier des groupes nominaux est linguistique et automatisable. En revanche, l'identification des termes requiert une compétence humaine spécialisée, puisque *in fine* il s'agit d'identifier les concepts propres à un domaine à partir des formes candidates retenues.

L'identification de SNP requiert elle aussi une compétence humaine puisqu'il y a une dépendance entre : la forme linguistique des SNP, les domaines d'activités abordés par le texte, l'objectif de l'application (pour quoi sont faits les SNP extraits, à quel type de traitement les destine-t-on ?). Le type de compétence humaine est reflété par le sens donné à la pertinence des SN. Ainsi la compétence sera celle du terminologue pour extraire des candidats termes, celle du consultant en information pour extraire des syntagmes destinés à la veille technologique, ou toute autre compétence pour tout autre profil de pertinence.

Nous pensons que la compétence de l'expert doit intervenir le plus tôt possible, bien avant l'étape finale de validation et de filtrage; elle doit intervenir dès la conception du système de filtrage. Dans les systèmes d'extraction de groupes nominaux comme *AlethIP*, la compétence de l'expert intervient *a posteriori*, c'est-à-dire après qu'ait eu lieu l'extraction des syntagmes ou des candidats termes. Par contre, un système d'extraction comme *Lexter* [Bou94b] incorpore un certain profil de pertinence pour l'identification des candidats termes. Les résultats dépendent alors d'une certaine grammaire du syntagme et d'un certain point de vue sur ce qui est pertinent ou ne l'est pas. Le terminologue ou l'expert vont donc valider ces résultats : ils vont rejeter les candidats qui d'après eux ne relèvent pas du domaine qu'ils cherchent à caractériser, et retenir les candidats termes qui sont des dénominations à homologuer.

Si l'intervention d'une compétence *a posteriori* est requise lorsqu'il s'agit de construire une terminologie homologuée, une intervention de la compétence *a priori* nous paraît nécessaire pour calibrer le système d'extraction : le système d'extraction ou de filtrage doit s'adapter aux types de textes et à l'objectif de l'application. Il ne doit pas être figé dans une unique grammaire d'extraction. Seule la mise en oeuvre d'un **apprentissage** du modèle linguistique des SNP permet une adaptation avec

5. Nous verrons en conclusion que nous n'avons pas pu tenir strictement cet objectif.

un investissement raisonnable. La compétence humaine requise se concentrera donc sur la constitution d'échantillons d'apprentissage, plutôt que sur l'écriture de règles de filtrage ad hoc.

2.6 Conclusion

Nous avons défini les SNP (Syntagmes Nominaux Pertinents) comme des objets documentaires qui recouvrent des réalités linguistiques diverses. Ces SNP peuvent être décrits par des informations linguistiques : le matériel lexical utilisé, les prépositions, la détermination sont des éléments à prendre en compte et seront formalisés en terme de dépendances et de compatibilités lexicales. Nous sommes partis du constat que les SNP ne peuvent être identifiés automatiquement à partir de critères linguistiques absolus, étant donné le caractère non prédictible de ce qui doit être considéré comme pertinent. Nous faisons donc intervenir le jugement de l'expert pour définir des échantillons de SNP. A partir de ces données rassemblées par l'expert, nous faisons l'hypothèse qu'il est possible de formaliser la forme des dépendances lexicales possiblement pertinentes à partir des dépendances lexicales pertinentes observées. Une étude partielle des syntagmes du thesaurus EDF et la réalité des corpus techniques montrent que l'on peut tabler sur une forme de compositionnalité des dépendances élémentaires. En effet lorsque l'on formalise ces dernières sous la forme de schémas plus abstraits, il est souvent possible de leur assigner un schéma d'interprétation.

Dans cette formalisation de l'observé, le lexique joue un rôle d'importance en fournissant les indices de formes pertinentes ou non pertinentes les plus évidents. La difficulté à généraliser de manière inductive des formes observées tient au fait qu'il faut passer par un niveau d'abstraction hyperonymique : à partir de *réacteur à eau pressurisée*, on réécrit *réacteur à* NOM-DE-MATIÈRE ADJ grâce au lien notionnel d'hyperonymie entre l'eau et les noms de matière. Cette généralisation présente un certain danger de surcouverture mais lorsqu'elle est utilisée en conjonction avec d'autres dépendances lexico-syntaxiques, sa portée est plus restreinte. Nous verrons au chapitre 5 (paragraphe 5.2.1) que cette forme de généralisation, qui vise à étendre la couverture des profils définis pour faciliter la prise en compte de nouveaux syntagmes, n'est pas la seule possible.

Dans l'immédiat nous présentons les outils linguistiques utilisés et développés, ainsi que notre corpus d'expérimentation.

Chapitre 3

Outils et ressources

3.1 Environnement de traitement linguistique

L'enrichissement linguistique d'un texte ASCII nécessite des moyens importants de traitement automatique du langage. Le texte doit être lemmatisé, catégorisé, subir une analyse syntaxique, etc. Nous avons construit notre système à partir des sorties d'analyses des outils linguistiques utilisés au département SID, c'est-à-dire la boîte à outils *Aleth* développée par la société Erli. Le moteur d'analyse qui assure lemmatisation, catégorisation et analyse syntaxique est nommé *AlethIP*. Il utilise une grammaire d'analyse robuste (*AlethGram*) développée dans le cadre du projet GRAAL déja mentionné. Cette robustesse lui permet d'analyser des textes tout-venant. *AlethIP* s'appuie aussi sur un dictionnaire nommé *AlethDic* que nous présentons brièvement en annexe A. Notre choix s'est porté sur cet outil industriel car il était le seul analyseur disponible au département SID, utilisé dans plusieurs applications documentaires comme l'indexation automatique.

3.1.1 Les pré-découpeurs et *AlethIP*

La robustesse de l'analyseur permet de traiter toutes sortes de textes sans limitation de taille. Toutefois les textes à analyser doivent être normalisés et mis au format de la DTD *GraalDoc*. Cette tâche est effectuée par des outils développés au groupe ISI et appelés *pré-découpeurs*. Développés à partir de grammaires *lex-yacc*, ils appauvrissent la typographie, retirent des parties de textes que l'on ne souhaite pas analyser (tableaux par exemples), normalisent la typographie (ponctuation), redressent la forme de certaines dates, cherchent à repérer les titres, etc. Chaque prédécoupeur traite un certain type de texte. Après cette phase de pré-découpage, le texte est soumis à *AlethIP*. Ensuite, il est lemmatisé, catégorisé et analysé syntaxiquement. Différentes options sont disponibles : extraction de groupes nominaux pour un export vers KES, indexation du document à partir du thesaurus EDF, analyse simple. Nous exploitons les sorties d'analyse simple, c'est-à-dire que chaque phrase est analysée sous la forme d'un arbre syntaxique annoté.

Un exemple de phrase analysée par *AlethIP*

Nous avons reproduit ci-dessous la phrase analysée par *AlethDic*. En l'état actuel, *AlethIP* n'est capable de traiter que des textes en typographie pauvre. Ainsi dans les exemples d'analyse que nous présenterons ne figurent aucune accentuation, ni distinction entre majuscules et minuscules). La phrase qui a été analysée figure entre les balises $< SEQ >$. La phrase analysée figure entre les balises $< RES >$. L'analyse représente un arbre syntaxique annoté avec des informations fonctionnelles (phrase, sujet, complément d'objet, etc., celles-ci ne correspondent pas toujours aux fonctions communément admises) et morphologiques. Les fonctions et catégories grammaticales sont précédées du symbole «slash»: `/gF_OBJ2` (objet), `/frNom` (nom). Les crochets signifient que l'on descend d'un niveau dans l'arbre (relation de dominance). Sur les noeuds terminaux, des traits (entre accolades) peuvent être attachés au catégories grammaticales: ils indiquent le genre, le nombre ou le mode la personne, le temps: `/frNom{GEf,NBs}`.

```
<ALETH> <TXTSRC> <SEQ NUM="XX-00001.00">cette reflexion est menee en liaison avec
l' ard e4101r sur les simulations numeriques de la turbulence et leurs applications
aux ecoulements externes &period; </SEQ> </TXTSRC> <ZRES>
<RES TY="resReprInterne">/gSentence [/gF_SUJ [/gNP [/gDet [/adjDem
CETTE]/frNom{GEf,NBs} REFLEXION]]/gVP [/gVtenseC [/gVtenseS
[/frVerbe{MODEind,TPSpres,PERS3,NBs} ETRE]/partPass{GEf,NBs} MENE]/gF_OBJ2 [/gPP
[/frPrep EN/gNP [/frNom{GEf,NBs} LIAISON]]]]] /frPrep AVEC /gDet [/frArt LE]
/gInc [/xx ARD/xx E4101R] /gPP [/frPrep SUR/gNP [/gDet [/frArt LES]/gNP [/gNP
[/frNom{GEf,NBp} SIMULATION/adjStd{GEf,NBp} NUMERIQUE]/gPP [/frPrep DE/gNP [/gDet
[/frArt LA]/frNom{GEf,NBs} TURBULENCE]]]]] /frConj ET /gNP [/gNP [/gDet [/adjPoss
LEURS]/frNom{GEf,NBp} APPLICATION]/gPP [/frPrep A/gNP [/gDet [/frArt LES]/gNP
[/frNom{GEm,NBp} ECOULEMENT/adjStd{GEm,NBp} EXTERNE]]]]</RES> </ZRES> <ZAPP>
</ZAPP> </ALETH>
```

La figure 3.1 donne la représentation graphique[1] correspondant à cette analyse. Comme on peut le remarquer, l'analyseur n'a pas correctement analysé la phrase. Une tentative a été faite pour attribuer des fonctions grammaticale (Sujet, Objet), puis le reste de la phrase a été analysé à part sous la forme d'arbres distincts du premier.

3.1.2 Les outils développés

Notre chaîne de traitement prend en entrée les phrases analysées par *AlethIP*. Nous n'intervenons pas sur ces analyses, bien qu'on y relève de nombreuses erreurs, étant donné le caractère robuste de l'analyseur. Nous acceptons l'imperfection actuelle pour donner la priorité aux autres traitements : ces analyses constituent l'entrée des processus d'enrichissement propres à notre système. Nous décrivons maintenant en quoi consiste ces derniers.

1. Parce que l'analyse a produit une forêt d'arbres, et pour une question d'occupation de la page, l'analyse est reproduite en mode paysage sur deux lignes

FIG. 3.1 – *Représentation graphique de l'analyse de la phrase* «cette réflexion est menée en liaison avec l' ARD E4101R sur les simulations numériques de la turbulence et leurs applications aux écoulements externes.»

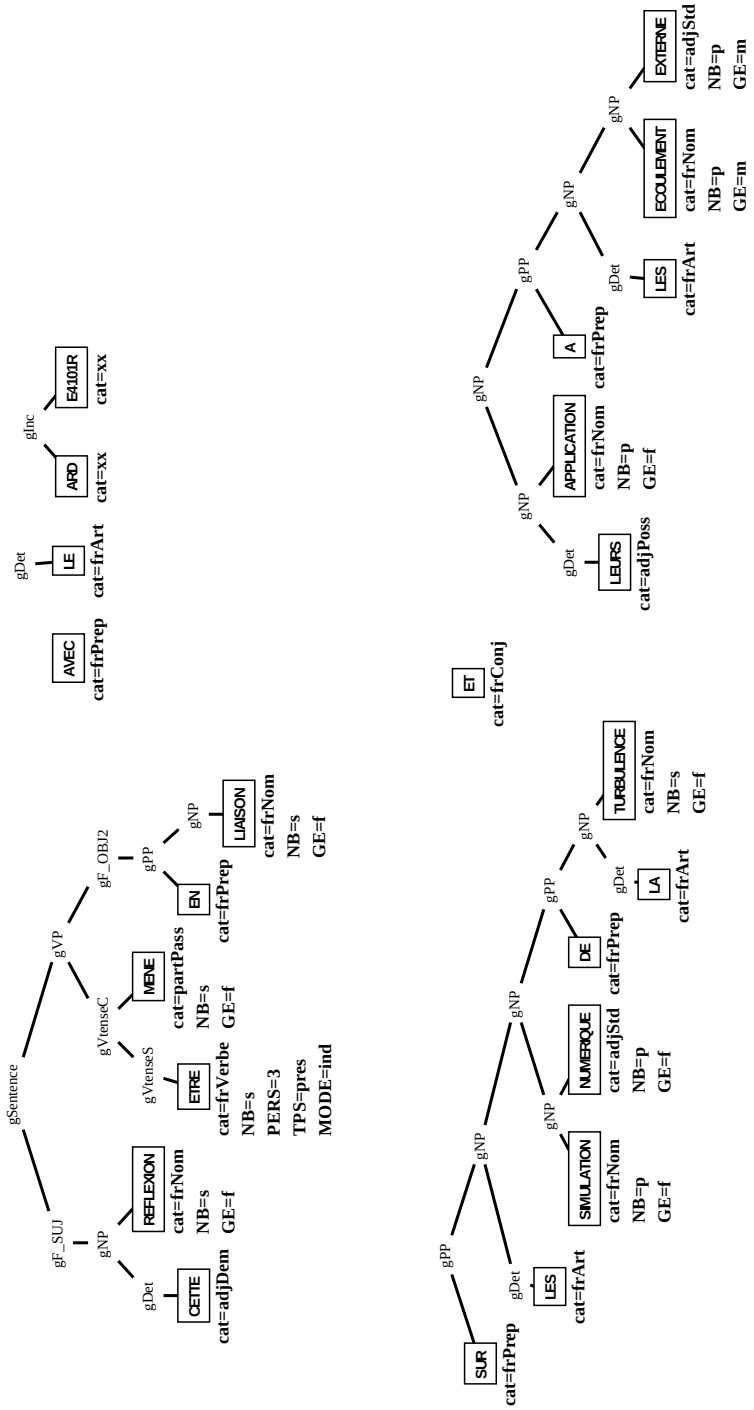
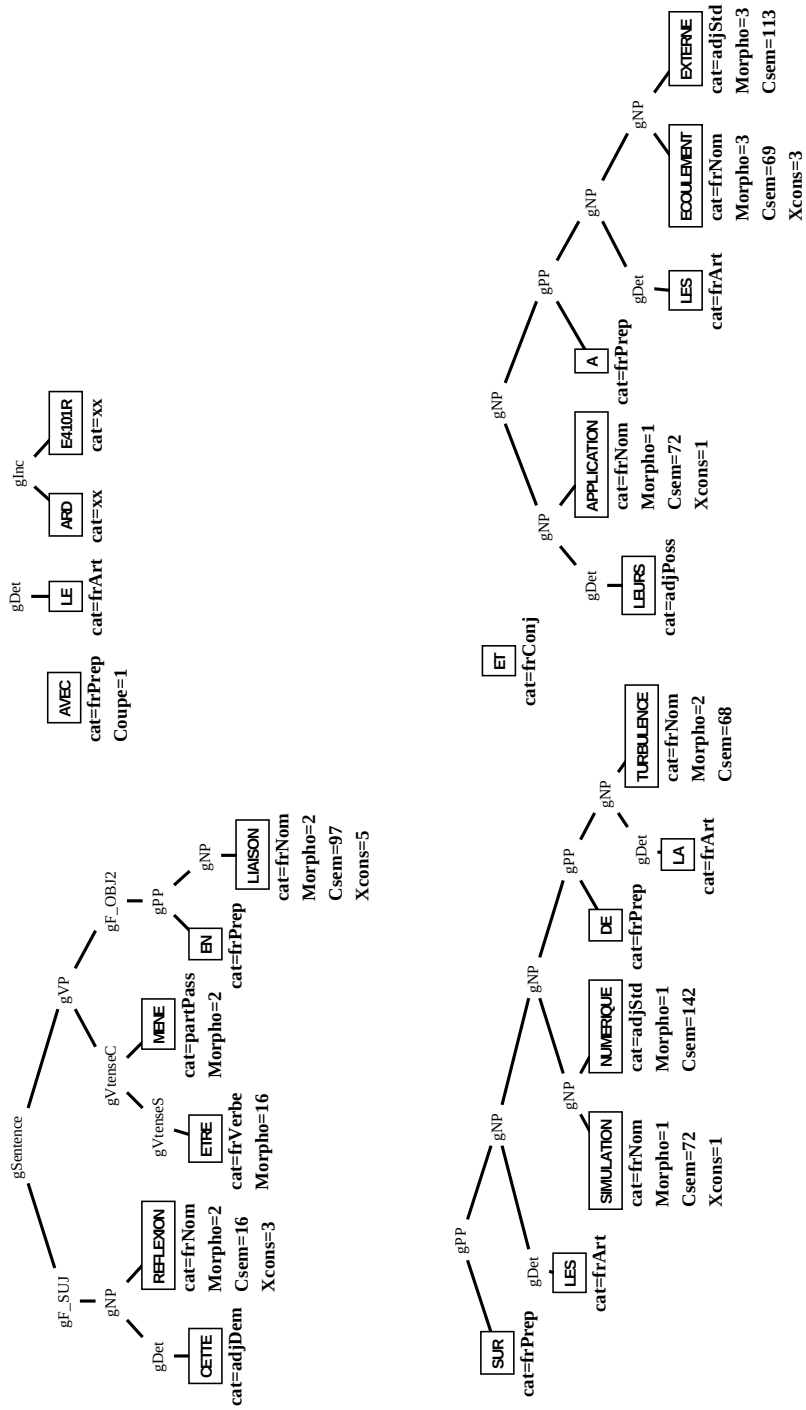


FIG. 3.2 – *Version enrichie de l'arbre de la figure 3.1. Seuls les traits attachés aux noeuds ont changé.*



Un exemple de phrase enrichie

L'analyse de la figure 3.2 montre la phrase précédente analysée par *AlethIP* puis enrichie avec notre système. Les traits apparaissent toujours sous les noeuds. La signification des traits associés aux catégories grammaticales est donnée en annexe A. Aucune information de suffixe n'apparait dans les arbres, elle est recalculée lorsque cela est nécessaire.

La chaîne de traitement

Voici les différentes étapes de notre chaîne de traitement :

E	Document/corpus analysé par d'*AlethIP*. Format ASCII
1	Normalisation du format d'analyse des phrases (réexpression du format des traits, etc.)
2	Numérisation du corpus : suppression de l'information de structure dans l'analyse et résolution des références avec le dicitonnaire AlethDic. De plus, chaque lexème est mis en relation avec le noeud correspondant dans l'arbre d'analyse d'*AlethIP*. Résultat : le document est représenté sous la forme d'une suite d'entiers.
3	Désambiguïsation lexicale du document (voir 5.3.2)
4	Répercussion des catégories sémantiques attribuées dans l'arbre d'analyse syntaxique
5	Extraction des groupes nominaux dans l'arbre d'analyse. Leur présence est marquée par dans l'arbre *AlethIP* par l'étiquette `/gNP`.
6	Extraction des dépendances syntaxiques élémentaires dans l'arbre d'analyse (voir 5.2.1)
7	Filtrage des groupes nominaux du corpus à partir des dépendances extraites et par rapport à un profil existant (voir 6.2.1).
8	Production des syntagmes nominaux pertinents (voir 6.2.2)
S	Syntagmes nominaux dits pertinents par rapport au profil de filtrage

L'étape 3 de désambiguïsation Cette étape est réalisée sur le document mis au format «numérique» (étape 2). Elle prend entrée un dictionnaire de désambiguïsation (constitué d'unités lexicales associées à des règles de désambiguïsations) et le document. Le document numérisé est modifié et ces modifications sont répercutées (étape 4) dans l'arbre d'analyse (stocké sous forme de fichier texte). Nous avons par ailleurs développé un environnement pour l'écriture et la mise au point de règles de désambiguïsation. Celui-ci est exposé en annexe B.

L'étape 5 d'extraction des groupes nominaux Nous nous appuyons sur la reconnaissance des constituants faite par *AlethIP* : les groupes nominaux sont étiquetés `/gNP`. Cependant, étant donné la structure de ces groupes nominaux et de fréquentes

erreurs d'analyse, nous avons dû introduire la notion de frontière comme dans *Lexter* [Bou94b]. Ainsi des coupures sont effectuées sur les pronoms relatifs, les groupes verbaux, ou des caractères typographiques particuliers (comme ceux marquant une puce). La détection de frontières est faite avec deux méthodes. L'une est spécifique et adaptée aux erreurs d'analyse d'*AlethIP*. Elle est réalisée par un parcours et une exploration des arbres d'analyses au moment de l'extraction des groupes nominaux. La seconde est plus générale et s'appuie sur le «pattern matcher» développé pour la désambiguïsation lexicale (voir 5.3.2, 5.3.3 et annexe B). Des descriptions linguistiques qui s'appliquent sur les séquences à plat correspondant aux arbres d'analyses permettent ainsi de placer dans les arbres un trait `Coupe` qui signifie qu'il faut couper à cet endroit (ou ne pas aller plus avant dans la lecture de l'arbre pour reprendre éventuellement après).

Constitution d'un profil de filtrage Les profils sont le résultat d'une procédure d'apprentissage qui prend en entrée : du texte, des groupes nominaux, ou des dépendances élémentaires identifiés comme pertinents ou non pertinents. Le texte ou les groupes nominaux sont renvoyés à l'étape d'analyse par *AlethIP* si ceux-ci ne sont pas déjà enrichis. Sinon, ils sont ré-introduits à l'étape 5. Le but est en effet de les représenter sous forme de dépendances syntaxiques élémentaires. Un environnement spécifique (décrit en annexe C) permet de construire des échantillons d'apprentissage à partir de ces dépendances. Le profil de filtrage constitué à partir d'un échantillon est finalement stocké dans une base de données simulant la structure d'un arbre de décision.

3.2 Définition du corpus de travail

Le choix du corpus n'est pas neutre ; c'est à partir du corpus que nous mettons au point et testons nos traitements pour valider nos hypothèses. Nous aurons à soumettre ce corpus à diverses analyses et enrichissements linguistiques et nous utiliserons certaines parties de ce corpus pour constituer des échantillons d'apprentissage. Le corpus de par sa taille et son contenu doit pouvoir s'accorder avec nos objectifs de recherches, aussi bien pour les traitements que pour les résultats, qui doivent être interprétés et évalués en prenant en considération les spécificités du corpus. C'est pourquoi nous précisons maintenant ses caractéristiques.

3.2.1 Hypothèses de construction du corpus

Les SNP peuvent être caractérisés par des dépendances lexico-syntaxiques

Cette hypothèse est fondée sur la conception harrissienne des sous-langages (exposée en 4.1) qui affirme qu'un sous-langage (une langue de spécialité), peut être caractérisé par des couples (prédicat, argument) et (modifié, modifieur) spécifiques. Cela suppose qu'il existe une langue de spécialité propre à un domaine d'activité, et

que ce dernier détermine des restrictions de sélection entre les opérateurs et leurs arguments dans cette langue – comme cela a été démontré dans [Sag87] pour la langue médicale. Ainsi on déduit l'hypothèse selon laquelle la terminologie d'un domaine d'activité est circonscrite dans une langue de spécialité. Le problème des SNP est différent de celui de la terminologie d'un domaine de spécialité (un SNP n'est pas nécessairement un terme et des SNP peuvent être définis à cheval sur deux domaines de spécialité) mais peut être traité de manière analogue, dans la mesure ou l'on donne dans les deux cas une grande importance au lexique et aux relations de dépendances syntaxiques qu'il manifeste. Voici un exemple pour un même nom recteur de deux paradigmes de modifieurs adjectivaux différents dans deux domaines différents :

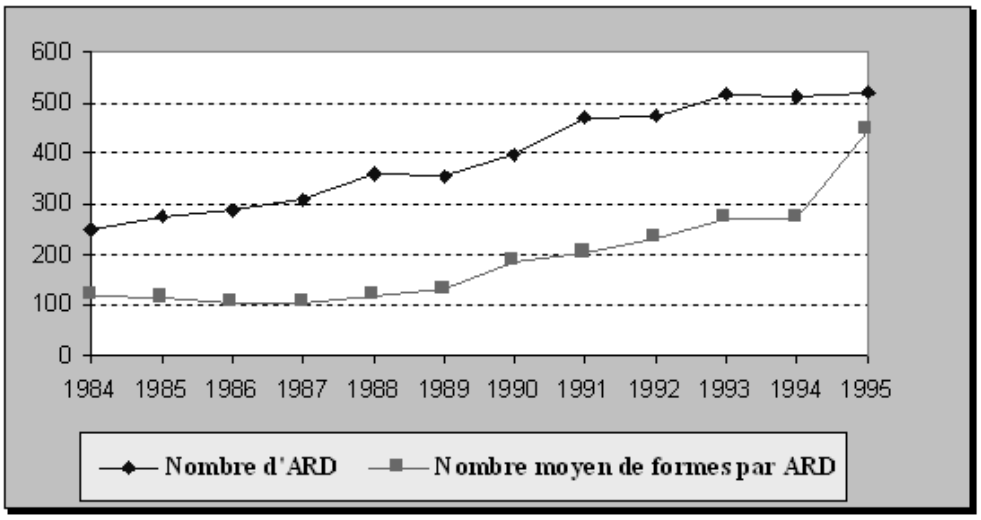
NEUTRONIQUE	BIOLOGIE
activation neutronique	activation cytologique
activation nucléaire	activation chimique
activation gamma	activation enzymatique

Par conséquent dans un souci de réalisme, notre corpus devra couvrir plusieurs domaines d'activité afin d'être en mesure de vérifier si un profil (le profil II) est capable de retrouver de manière sélective des SNP relatifs à un domaine.

Evolution des dépendances lexico-syntaxiques - mesure de la variation

L'évolution du lexique est un indicateur certain de l'évolution d'un domaine d'actvité. Mais l'évolution des dépendances lexico-syntaxiques constitue également un indicateur de l'évolution du domaine; outre que cette évolution est dépendante de celle du lexique, elle permet d'expliciter les relations entre les lexèmes : elle renseigne sur les compatibilités lexicales et les restrictions de sélection pratiquées. Cette forme d'évolution est en étroite connexion avec les phénomènes de variation et notamment de variation terminologique. Comme le montrent les travaux résumés dans [Jac97], la variation terminologique n'est pas un phénomène mineur et doit être prise en considération si l'on veut suivre l'évolution d'un domaine déterminé. En effet un domaine d'activité productif voit toujours ses termes évoluer par rapport à un certain existant; certains disparaissent, d'autres changent morphologiquement et syntaxiquement. Il est intéressant d'étudier diachroniquement l'évolution de la terminologie d'un domaine, du point de vue de l'apparition de nouveaux termes et de la variabilité des termes existants. Ainsi, disposer d'un corpus composé de textes écrits sur plusieurs années est avantageux. Sans toutefois entrer dans le champ d'étude de la variation, qui nécessite comme le montre C. Jacquemin des outils et des traitements élaborés. Nous présenterons une approche globale du phénomène par l'observation de l'évolution des dépendances lexicales sur des sous-corpus datant d'années différentes. (voir chapitre 8, 8.4).

FIG. 3.3 – *Evolution du nombres de formes dans les ARD*



3.2.2 Le corpus EDF-ARD

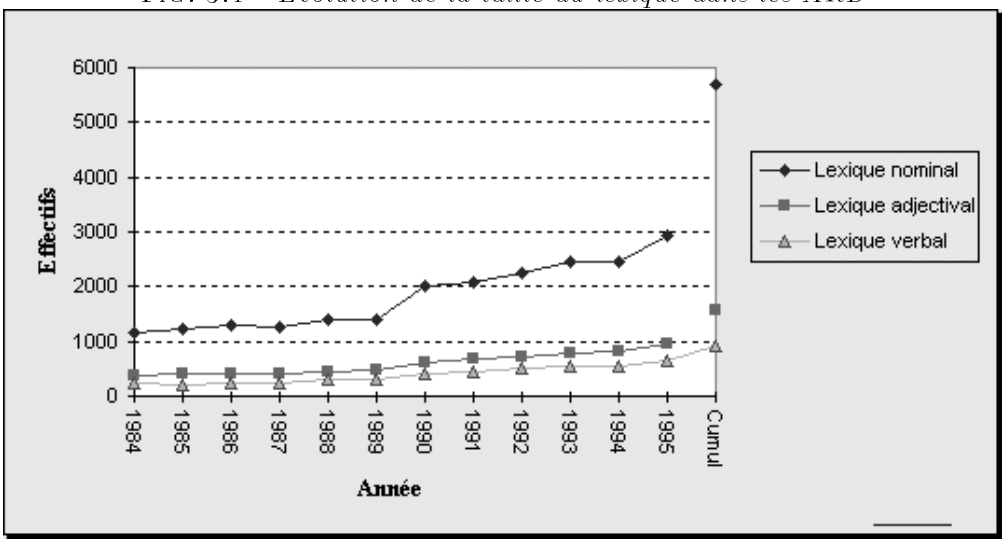
Nature du corpus

Nous avons choisi comme corpus l'ensemble des ARD (Actions de Recherche Développement) de l'axe de recherche numéro 2, en rapport avec le matériel nucléaire, pour les années 1984 à 1995. Une ARD est un court texte, d'une ou deux pages, écrit par un chercheur de la DER, et qui présente l'état d'avancement d'une action de recherche. Ce texte est structuré en 4 parties : contexte de l'action, but de l'action, état d'avancement, et objectifs. Le nombre d'ARD sélectionnées est de 1781.

Taille du corpus L'ensemble des ARD de l'axe 2, pour les 11 ans couverts, représente 998 613 formes graphiques (mots et ponctuations), soit environ 850 000 mots sans la ponctuation (soit 6 mégaoctets). Comme le montre la figure 3.3, le nombre moyen de formes par ARD a notablement crû ces dernières années (d'un quart de page dactylographiée en 1984 à une page et demie voire deux en 1995), de même que le nombre d'ARD par année. L'une des raisons est qu'il a été demandé aux chercheurs d'être plus exhaustifs dans la rédaction des ARD.

Evolution du lexique nominal, adjectival et verbal Ces 850 000 mots constituent un lexique de 5677 noms, 1581 adjectifs et 952 verbes. La figure 3.4 montre que les tailles des lexiques nominal et adjectival ont crû lentement jusqu'en 1989. Ensuite, elles croîssent plus rapidement. Le cumul correspond à la taille du lexique pour toutes les ARD, toutes années confondues. Le lexique verbal a crû très faiblement.

FIG. 3.4 – *Evolution de la taille du lexique dans les ARD*



Phénomènes perturbateurs naturels et artificiels

Connexité avec d'autres domaines L'axe de recherche numéro 2 est en rapport avec le matériel nucléaire. Ainsi, le contenu des ARD de l'axe 2 traite principalement de matériel nucléaire ; mais pas seulement : de nombreux domaines connexes apparaissent inévitablement (informatique, modélisation, environnement, matériaux, traitement des déchets, sécurité ...). On peut donc s'attendre à la présence simultanée de formes issues de diverses langues de spécialité. C'est un phénomène naturel qui rend plus difficile l'identification de paradigmes opérateurs/arguments propres à une seule langue de spécialité. La figure 3.5 illustre ce phénomène.

Qualité typographique du corpus Le corpus avant analyse présente de nombreuses erreurs typographiques. Par exemple, l'espace après le point de fin de phrase ou la virgule de respiration sont absents. Cela entraîne des erreurs à la lemmatisation/catégorisation puis à l'analyse syntaxique. En effet dans ces cas, le point et la virgule ne seront pas considérés comme des séparateurs. Il n'est pas envisageable de corriger simplement (au moyen d'outils Unix comme `sed` ou `awk`) ces erreurs sans intervention manuelle. En effet le corpus contient de nombreux sigles et acronymes utilisant les points (X.Y.Z) et de nombreuses valeurs numériques écrites avec des virgules (4,3). On pourrait éventuellement envisager une correction après une première analyse, puis effectuer un nouveau passage (lemmatisation/catégorisation/analyse syntaxique). Nous avons laissé toutes ces erreurs par souci de réalisme et d'économie de temps.

Par ailleurs, les ARD comportent parfois des fautes de frappes qui sont passées au travers des relectures.

FIG. 3.5 – *Exemple de connexité dans un paragraphe d'une ARD de 1985.*

« . . . Afin d'apporter une plus grande précision dans la définition des *zones inondables*[d] lors de la propagation d'une *onde de submersion*[d] à l'aval d'un *barrage*[a] (pour les besoins de la *protection civile*[c] ou pour l'implantation d'*installations de production électrique*[a] telles que les *centrales nucléaires*[a]), les *codes bidimensionnels d'écoulement*[b] sont adaptés au cas d'une *vallée large*[e].
Objectif 1984 : le *code Cythère ES1*[b] de calcul des *courants*[d], développé pour les usages maritimes et spécialement adapté aux faibles *hauteurs d'eau*[d] . . . »

[a] Production d'énergie.
[b] Recherche opérationnelle.
[c] Droit civil.
[d] Océanographie, Hydrologie.
[e] Géomorphologie.

FIG. 3.6 – *Exemple d'analyse morpho-syntaxique erronée causée par des problèmes typographiques*

"(...) Fonctionnement des vannes et soupapes en vrai grandeur.elle est composée d'une chaudière supercritique avec stockage d'énergie (...)"

```
(..) /gNP [/frNom{GEm,NBs} FONCTIONNEMENT/gPP [/frPrep DE/gNP [/gDet
[/frArt LES]/frNom{GEf,NBp} VANNE]]]]]]]]] /frConj ET /gNP [/gCompNom
[/frNom{GEf,NBp} SOUPAPE /xxSigle PWR]] /frPrep EN /adjStd{GEf,NBs}
VRAI [/gInc [ /xx GRANDEUR.ELLE]] /gVP [/gVtenseC [/gVtenseS [ /frVerbe
{MODEind, TPSpres, PERS3,NBs} ETRE] /partPass{GEf,NBs} COMPOSE] /gF-OBJ1
[/gPP [/frPrep DE /gNP [/gDet [/frArt UNE] /gNP [/gNP [/frNom{GEf,NBs}
CHAUDIERE/adjStd{GEf,NBs} SUPERCRITIQUE]/gPP [/frPrep AVEC/gNP [/gNP
[/frNom{GEm,NBs} STOCKAGE/gPP [/frPrep DE/frNom{GEf,NBs} ENERGIE]] (...)
```

"Réalisation d'études sur les grandes transitoires, la régulation générale, les protections, les incidents dissymétriques,la vérification des options de fonctionnement."

```
/gNP [/frNom{GEf,NBs} REALISATION/gPP [/frPrep DE/frNom{GEf,NBp}
ETUDE]] /frPrep SUR /gDet [/frArt LES] /adjStd{GEm,NBp} GRAND
/adjStd{GEm,NBp} TRANSITOIRE /Caract &comma; /gNP [/gDet [/frArt LA]/gNP
[/frNom{GEf,NBs} REGULATION/adjStd{GEf,NBs} GENERAL]] /Caract &comma;
/gNP [/gDet [/frArt LES]/frNom{GEf,NBp} PROTECTION] /Caract &comma;
/gNP [/gDet [/frArt LES]/gNP [/gCompNom [/gCompNom [/frNom{GEm,NBp}
INCIDENT[/gInc [/xx DISSYMETRIQUES,LA]] /frNom{GEf,NBs} VERIFICATION]/gPP
[/frPrep DE/gNP [/gDet [/frArt LES]/gNP [/frNom{GEf,NBp} OPTION/gPP [/frPrep
DE/frNom{GEm,NBs} FONCTIONNEMENT]]]]]]
```

Dégradation attendue des résultats

Les problèmes typographiques rendent difficile l'identification des mots graphiques dans la chaîne textuelle. Cela entraine des erreurs de lemmatisation et de catégorisation, et produit des entrées lexicales inconnues pour le dictionnaire et pour le lecteur humain. En aval, les analyses syntaxiques (identification des groupes nominaux par exemple) et éventuellement sémantiques, subissent les conséquences de ces erreurs. On dénombre environ 2000 cas de formes soudées du type « aaa.bbb », et environ 1000 cas de formes soudées de type « aaa,bbb ». La figure 3.6 montre un exemple d'analyse morpho-syntaxique erronée produite par l'analyseur AlethIP. La première séquence est le cas de deux phrases qui ont été fusionnées en une seule à cause d'un espace non inséré après le point de fin de phrase. Il en résulte la création d'un lexème `grandeur.elle` catégorisé comme un inconnu et perturbant clairement l'analyse syntaxique. La seconde séquence montre le cas de deux groupes nominaux fusionnés en plein milieu d'une énumération.

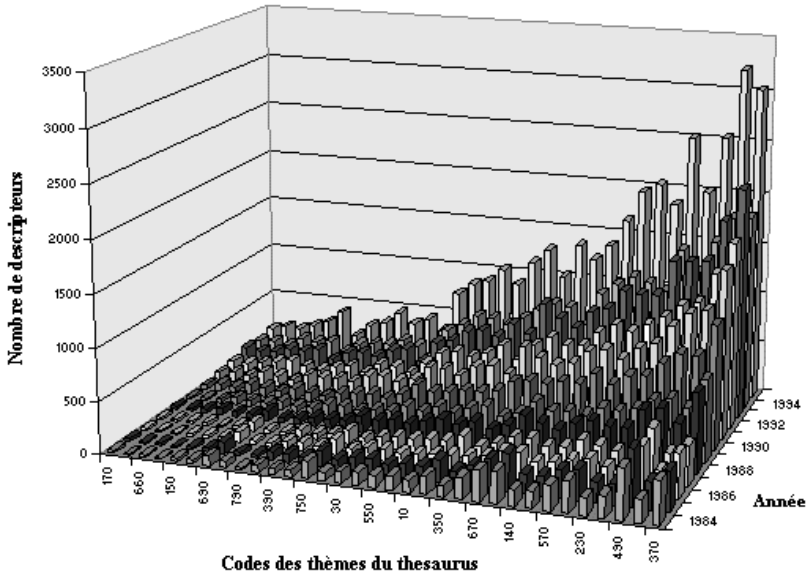
Quant à l'hétérogénéité et la connexité des domaines d'activité abordés dans le corpus, qui sont des phénomènes naturels, elles complexifieront la tâche d'apprentissage des filtres de sélection des SNP, notamment en multipliant le nombre de configurations linguistiques à prendre en compte et en multipliant les règles de filtrage induites. Ce phénomène atténuera certainement le pouvoir séparateur des profils de pertinence. On peut s'y adapter en resserrant la couverture du profil, en construisant des profils très spécifiques (comme le profil II).

Couverture du corpus EDF-ARD par le thesaurus EDF

Pour estimer les différents domaines abordés dans le corpus ARD de 1984 à 1995, nous avons procédé à l'indexation automatique de ce corpus avec les outils disponibles à EDF. Il s'agit d'une application d'indexation automatique contrôlée qui repose sur le moteur d'analyse AlethIP (société Gsi-Erli). Le contrôle s'est effectué à partir du thesaurus EDF. Ce thesaurus est organisé en une cinquantaine de thèmes généraux et chaque thème se subdivise en champs sémantiques dont le nombre total est de 330. Par exemple le thème général de l'informatique se subdivise en 12 champs sémantiques dont «système d'exploitation», «intelligence artificielle», «réseaux informatiques», «stockage d'information», etc. Au total, on dénombre un peu plus de 20.000 entrées (dont environ 13.000 descripteurs et 7.000 synonymes) dans ce thesaurus. Le résultat d'indexation liste pour chaque phrase les descripteurs du thesaurus retrouvés. Les résultats que nous présentons sont à prendre avec beaucoup de précautions. En effet dans son état actuel, le thesaurus est incomplet et de nombreuses erreurs d'indexation sont commises par l'application, notamment à cause de l'indexation d'unitermes non désambiguïsés. En voici un exemple :

(...) un système d'acquisition des mesures nettement plus performant (...)
DOM=090, CS=098, Descripteur = « achat »

FIG. 3.7 – *Couverture du corpus ARD84-95 par le thesaurus EDF. Ce graphique représente le nombre de descripteurs reconnus par année par l'application d'indexation automatique appliquée à notre corpus. Ces descripteurs sont répartis par domaines d'activité (encore appelé thèmes du thesaurus), indiqués en abscisse, et dont la signification est donnée en table 3.1.*



Le descripteur « achat » du champ sémantique 098 (« type de contrat ») appartenant au thème 090 (Droit) a été retenu à tort, puisqu'il s'agit de l'acquisition de mesures et non de l'acquisition d'un objet concret qui a une valeur marchande.

La figure 3.7 montre les effectifs des descripteurs reconnus par domaine et pour les ARD des années 1984 à 1995 (la taille de ces sous-corpus annuels augmentant chaque année, voir figure 1). Les 45 thèmes du thesaurus EDF sont représentés dans les résultats. Dans le graphique, nous n'avons introduit que les 36 thèmes les mieux représentés, les 10 thèmes les moins représentés obtenant des effectifs inférieurs à la centaine. Les thèmes sont introduits sur l'axe des abscisses par ordre croissant des effectifs de descripteurs toutes années confondues. Nous avons volontairement omis dans le graphique les effectifs du thème 880 dit des listes annexes (Il est constitué de listes d'unitermes très généraux: *étude, demande*, qui viennent perturber les résultats d'indexation) qui sont très importants (presque 74000 descripteurs reconnus toutes années confondues, dont 16895 pour la seule année 1995). On observe que les domaines les plus représentés sont les mêmes quelle que soit l'année (les effectifs croissants avec les années et donc avec la taille des corpus annuels). Il semblerait donc que malgré les biais introduits par l'application d'indexation, ce soient toujours

TAB. 3.1 – *Signification des codes des thèmes du thesaurus EDF*

Codes	Libellés des thèmes du thesaurus
430	PHYSIQUE DES REACTEURS NUCLEAIRES
170	AMENAGEMENT DU TERRITOIRE
210	ACTIVITE COMMERCIALE
660	SECURITE
290	REGLEMENTS TECHNIQUES-NORMALISATION
150	ENVIRONNEMENT SOCIAL
530	PROPRIETE DES MATERIAUX
690	MACHINE NON ELECTRIQUE
410	PHYSIQUE DE LA MATIERE
790	CONSTRUCTION
710	MATERIEL ELECTRIQUE-APPAREILLAGE ELECTRIQUE
390	MECANIQUE DES FLUIDES
610	UTILISATION DES EQUIPEMENTS
750	PRODUCTION D'ENERGIE
070	ECONOMIE
030	SCIENCES DE LA TERRE
450	THERMIQUE
550	ETUDE DES MATERIAUX
590	MANOEUVRE ET OPERATION
010	BIOLOGIE
510	PRODUIT DE BASE
350	INFORMATIQUE
630	UTILISATION DE SYSTEME
670	APPAREILLAGE MECANIQUE
470	CHIMIE
140	SCIENCES HUMAINES
090	DROIT (JURISPRUDENCE)
570	TRAITEMENT DES MATERIAUX
250	GESTION DU PERSONNEL
230	SCIENCES DE L'INFORMATION
190	ENTREPRISE
490	METROLOGIE
310	MATHEMATIQUES
370	SCIENCES PHYSIQUES

les mêmes domaines qui émergent. La table 3.1 explicite les codes de thèmes utilisés pour les comptages. Les thèmes sont classés de haut en bas par ordre croissant d'effectifs, toutes années confondues. Les thèmes correspondent ainsi à ceux qui ont été reportés en abscisse dans la figure 3.7, dans le même ordre.

Les derniers thèmes se dégagent de l'ensemble (voir la figure 3.7); cependant la pluralité des thèmes abordés montre la forte connexité des domaines d'activité. On ne peut pas dire que le corpus soit thématiquement homogène (par rapport à l'axe de recherche numéro 2 sur le matériel nucléaire, bien que l'activité autour du matériel nucléaire soit certainement reflétée par ces différents thèmes[2].

3.3 Conclusion

Nous avons présenté dans ce chapitre les outils d'analyses linguistiques utilisés au département SID. Nous récupérons leurs sorties d'analyse pour les introduire à

2. Une analyse plus fine de la couverture du thesaurus est nécessaire, notamment, l'évaluation de l'efficacité du contrôle par le thesaurus dans son état actuel dont on sait qu'il est incomplet. Ainsi une distinction devrait être faite entre formes vedettes et synonymes, entre les unitermes et les termes complexes (moins sensibles aux polysémies), entre le nombre de descripteurs existants par thèmes, le nombre de descripteurs reconnus par thème et le nombre d'occurrences de descripteurs comptabilisés par thèmes.

l'entrée de notre chaîne de traitement. Notons que nous aurions pu tout aussi bien exploiter les sorties d'autres analyseurs syntaxiques, pourvu que ceux-ci fournissent des groupes nominaux dont l'analyse mette en évidence les différents constituants. C'est le cas par exemple de *Lexter* qui peut produire des syntagmes nominaux maximaux structurés en termes de tête-expansion, ou de *Sylex* [Con91] qui produit des analyses des dépendances entre constituants. Nous aurons l'occasion de présenter plus en détail certaines étapes de la chaîne de traitement dans les chapitres suivants. La désambiguïsation sera abordée au chapitre 5 (5.3.2). Le filtrage sera expliqué au chapitre 6. Enfin la constitution de profils de filtrage par apprentissage sera présentée au chapitre 7.

Nous avons également présenté le corpus EDF sur lequel nous avons testé notre chaîne de traitement. Ce corpus est très hétérogène du point de vue des notions qu'il aborde bien que les textes qui le composent appartiennent au thème spécifique du «matériel nucléaire». Il présente aussi des erreurs typographiques. En ce sens c'est un corpus conforme à la réalité de textes tout-venants, qui doivent être traités au jour le jour sans investissement lourd de correction et de nettoyage. Une autre caractéristique est son étalement dans le temps. Nous avons fait au chapitre 8 une brève étude diachronique qui montre combien le renouvellement des combinaisons lexicales est important.

Le prochain chapitre exposera les choix syntaxiques et sémantiques que nous avons faits pour identifier en corpus la notion de syntagme nominal pertinent.

Chapitre 4

Choix d'une approche syntaxique et sémantique

Nous avons défini au chapitre 2 l'orientation que nous voulions donner au repérage des SNP. Nous allons maintenant présenter l'approche linguistique et informatique nécessaire pour l'implémentation de la méthode. Comme cela est énoncé au chapitre 2, ce repérage repose sur l'utilisation d'informations linguistiques et notamment l'exploitation d'une analyse syntaxique des groupes nominaux en terme de dépendances lexicales.

4.1 Les possibilités offertes par la syntaxe

4.1.1 L'hypothèse distributionnelle

La conception de Z. Harris du rapport entre la syntaxe et la sémantique se fonde sur la distribution des mots dans les phrases. D'après Harris [Har71], l'hypothèse distributionnelle met en relation la distribution syntaxique des mots (appelés *unités*) avec leur contenu informationnel : «*la signification des unités et de leur relations grammaticales est liée à la restriction imposée sur les combinaisons de ces unités avec d'autres*» (p. 14). La deuxième facette de l'hypothèse distributionnelle est le concept de sous-langage. Selon Harris, la grammaire d'un sous-langage partage sa syntaxe avec celle de la langue générale et se distingue de celle-ci par des coocurrences spécifiques de classes de mots (compatibilités lexicales) qui sont différentes de la langue générale. Ainsi, la sémantique peut être déduite par une étude des mots qui apparaissent dans des contextes syntaxiques donnés (tel nom accepte tel modifieur adjectival, tel verbe accepte tel argument). La syntaxe est indépendante du domaine, mais les compatibilités de combinaison entre les *unités*, c'est à dire les restrictions de sélection, sont dépendantes du domaine. Le sens (la sémantique) serait ainsi construit à travers une organisation structurée (la syntaxe) et morphologique (le lexique).[1].

1. Cette construction est le processus interprétatif même alors que l'interprétation informatique de la conception distributionnelle tend à en faire le résultat d'une combinatoire. C'est ce qui explique

A partir de l'hypothèse distributionnelle – qui se résume ainsi : «le sens des mots se déduit des constructions dans lesquelles il figurent» [HN96] – une autre hypothèse se déduit : si deux formes différentes partagent des contextes identiques, leurs sens sont proches. Une telle hypothèse, couplée au concept de sous-langage, permet d'envisager des applications de traitement du langage. Ainsi, en constituant des paradigmes d'opérateurs et d'arguments, il est possible de constituer des classes sémantiques propres au domaine (comme le font [Gre94] et [HN96]), de même qu'il est possible, en mettant en évidence les restrictions de sélection du texte, de généraliser des patrons syntaxico-sémantiques propres au sous-langage du domaine de spécialité. Ces patrons peuvent alors servir à faire de la désambiguïsation syntaxique ([BPV93c, BPV93a], [Res93]).

4.1.2 La syntaxe permet une analyse distributionnelle fine

L'hypothèse distributionnelle est à la base des travaux de N. Sager sur le discours médical [SFe87]. Sager et son équipe ont développé une méthodologie de mise en évidence de catégories sémantiques propres au sous-langage du domaine médical. La méthode est manuelle et repose sur la normalisation syntaxique des phrases du corpus. Par normalisation syntaxique, il faut entendre un certain nombre de transformations des phrases pour les ramener à des schémas syntaxiques plus simples et plus réguliers (par exemple, suppression de la diathèse, distribution des groupes coordonnés, etc.) [Dac94]. Les phrases normalisées sont appelées phrases élémentaires. L'alignement vertical et le tri de ces phrases élémentaires fait apparaître les opérateurs du sous-langage et leurs arguments, c'est à dire les classes de mots utilisées avec ces opérateurs. Un travail d'interprétation final permet de faire une correspondance entre ces classes de mots et des classes sémantiques fortement liées au domaine médical. Le nom donné aux classes sémantiques ainsi constituées est également le résultat d'une l'interprétation.

De nombreux travaux ont cherché à valider l'hypothèse distributionnelle avec l'aide des outils informatiques. L'interprétation qui est alors donnée de l'environnement syntaxique, du contexte, donne lieu à deux types d'analyse distributionnelle : l'analyse distributionnelle basée sur un contexte graphique, et celle basée sur un contexte syntaxique. Pour donner un exemple et comparer les deux types de modélisation du contexte, nous prenons la phrase suivante extraite du corpus ARD : «*la plupart des études font intervenir des phénomènes physiques complexes plus ou moins bien connus et modélisables physiquement ou mathématiquement.* », et considérons que l'automate pointe sur le mot *physiques*.

Le premier type de modélisation du contexte considère une fenêtre graphique appelée n-grammes qui est définie en général sur n mots à gauche et à droite. En voici un exemple, toujours avec le mot *physiques*:

---

la nécessité du travail interprétatif pour la validation des résultats de traitements automatiques basés sur ce principe, voir paragraphe suivant : 4.1.2

intervenir	des	phénomènes	physiques	complexes	plus	ou
–3m	-2m	-1m	physiques	+1m	+2m	+3m

Un traitement statistique (information mutuelle) de ces contextes permet alors de faire ressortir des collocations [CH90], définies comme des combinaisons de mots dépendantes du domaine, récurrentes et qui constituent des syntagmes dotés d'une certaine cohésion. Dans ce type de traitement, le corpus utilisé doit nécessairement être de taille importante pour que les mots apparaissent fréquemment avec leurs contextes privilégiés.

Le système *Xtract* [Sma93a, Sma93b] utilise de tels n-grammes (-5; +5) pour trouver des collocations en ajoutant de l'information linguistique. Une catégorisation grammaticale des mots trouvés dans les n-grammes est effectuée sur les collocations les plus significatives et permet d'identifier les relations verbes–arguments et les groupes nominaux. En revanche, la méthode d'extraction automatique de terminologie de B. Daille [Dai94] commence par l'étiquetage du texte et procède ensuite au calcul des n-grammes.

Le deuxième type de modélisation du contexte repose sur une analyse syntaxique des relations de dépendance entre les mots. Par exemple, toujours à partir de l'exemple précédent et si l'on s'intéresse à la relation de type nom–adjectif, on obtient les relations suivantes :

phénomènes	physiques
phénomènes	complexes
phénomènes	connus
phénomènes	modélisables

L'utilisation d'une fenêtre syntaxique est plus lourde à mettre en oeuvre qu'une fenêtre graphique, puisqu'elle demande une analyse linguistique. Mais l'avantage de la fenêtre syntaxique est qu'elle fournit des relations syntaxiques entre les mots qui sont avérées (si tant est que l'analyseur fournisse des résultats corrects), ce qui n'est pas le cas avec les autres types de fenêtres. Il en résulte que la taille du corpus nécessaire pour trouver des collocations ou constituer des classes sémantiques ne doit pas être aussi importante que dans le cas des fenêtres graphiques, comme le montre les travaux de [Gre94] et [HNN96]. Dans le cas des fenêtres graphiques, seuls de très grands nombres d'occurrences de mots et de contextes sont capables de gommer les variations passagères de formes (flexions différentes, insertions de modifieurs au sein des n-grammes).

A partir de telles dépendances lexicales syntaxiquement attestées, le système *SEXTANT* [Gre94] calcule des classes sémantiques sur la base d'une mesure de similarité (indice de Jacquard) qui prend en compte les relations lexico-syntaxiques avec leur fonction (sujet, objet, objet indirect, attribut). L'hypothèse distributionnaliste est partiellement vérifiée : les classes construites sont souvent pertinentes, mais l'ambiguïté des mots simples qui les constituent en altère l'homogénéité.

Le logiciel *ZELLIG* [HNN96, HN96, BHNZ97] construit également des classes sémantiques à partir de dépendances élémentaires, mais contrairement à [Gre94] la méthode utilisée est symbolique et permet de conserver et d'exploiter les critères de construction des classes lors de l'interprétation finale. Cela n'est pas possible avec les mesures statistiques de similarité [Nau96] qui synthétisent ces informations sous la forme d'indicateurs numériques.

Conclusion Le recours à l'analyse syntaxique autorise une analyse distributionnelle plus fine sur des corpus de taille moins importante que lorsque l'on recourt à des n-grammes. Elle permet la construction de classes sémantiques et le calcul de contraintes de sélection. Pour le filtrage de syntagmes pertinents , nous nous inspirons partiellement de la méthodologie distributionelle de Sager : nous nous restreignons à l'étude des syntagmes nominaux dont l'analyse aura été réalisée automatiquement, et nous normalisons les syntagmes à l'aide de dépendances élémentaires (voir plus loin 5.2.1). Nous pourrons alors dégager des paradigmes d'opérateurs–arguments ou modifié–modifieurs. Des informations linguistiques étendues – morphologiques, syntaxiques, sémantiques – permettront d'affiner l'analyse distributionnelle.

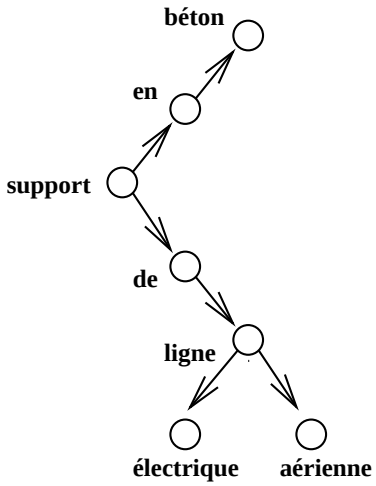
Notons enfin que le recours à la syntaxe rend possible un calcul de la variation syntaxique. C'est une piste qui montre déjà son efficacité et dont les applications pour la recherche documentaire paraissent prometteuses. La prise en compte des phénomènes de variation syntaxique permet en effet d'améliorer la reconnaissance terminologique [Jac97],[HBGN+97] et de mettre en évidence des relations notionnelles entre les syntagmes nominaux [Jac96].

4.1.3 Grammaire de dépendance et grammaire de constituants

Nous avons présenté au chapitre 3 le format de sortie des analyses syntaxiques d'*AlethIP*. Ce format de sortie s'exprime dans une formalisme de grammaire de constituants. I. Mel'cuk [Mel88a] contraste les grammaires de constituants (*Phrasal Structure trees*) et les grammaires de dépendances (*Dependancy-trees*). Dans les grammaires de constituants, les unités lexicales sont regroupées ensemble pour en former d'autres d'un niveau syntaxique supérieur (par exemple : groupe nominal, groupe verbal, phrase). Ceci introduit la notion de structure syntaxique. Tête et modifieur peuvent ainsi être distingués. Ces informations figurent sur des noeuds non-terminaux. Seules les unités lexicales réalisées figurent sur des noeuds terminaux. Enfin, les grammaires de constituants sont linéaires, la structure et l'ordre des éléments peut se déduire de l'ordre des fomes lexicales de l'énoncé.

Les grammaires de dépendances, quant à elles, n'introduisent pas la notion de structure syntaxique. L'axe syntagmatique linéaire n'est pas déductible de la représentation : l'ordre des mots n'est pas représenté. En revanche, les unités sont mises en relation en termes de liens hiérarchiques : le formalisme des dépendances exprime directement le rapport naturel entre les formes lexicales. Enfin, dans une représentation en termes de dépendances, il n'y a que des noeuds terminaux. Ces principales caractéristiques sont illustrées par la figure 4.1 qui donne une représentation du syntagme

FIG. 4.1 – *Représentation en termes de dépendances du syntagmes* support de ligne électrique aérienne en béton



support de ligne électrique aérienne en béton: on pourrait également lire *support de ligne aérienne en béton* ou encore *support en béton de ligne aérienne électrique*, etc.

Nous optons pour le formalisme des dépendances pour décrire les éléments qui composent les syntagmes nominaux. Ceci nous permet de représenter directement les relations qu'ils entretiennent et de normaliser le syntagme nominal comme des combinaisons élémentaires. Mais nous conservons le formalisme des constituants (utilisé par *AlethIP*) pour la représentation globale des syntagmes nominaux, qui conservent l'ordre syntagmatique des formes lexicales. De cette manière, nous serons en mesure de reconstituer un groupe nominal à partir de ses dépendances élémentaires. Cette possibilité sera exploitée lors du filtrage des groupes nominaux (voir chapitre 6).

4.1.4 Dépendances syntaxiques élémentaires : réalité linguistique et interprétabilité

Comme le montre la figure 4.1, dans une représentation de type *D-trees*, les unités lexicales sont mises en relation deux à deux. Le lien de dépendance est un lien binaire. Ainsi, si l'on décompose les couples de la figure 4.1, on obtiendra des dépendances comme {support→en}, {en→béton}. Nous appellerons dépendances élémentaires[2] les dépendances binaires (combinées si nécessaire) exprimant la relation entre une tête et un modifieur (ou un prédicat et un argument) dans leur ordre d'apparition

2. Nous utiliserons également le terme d'arbre élémentaire, étant donné que la contrainte exprimée (un tête et un modifieur dans l'ordre d'apparition) définit un syntagme minimal, même si ce dernier n'a pas de réalisation effective en corpus

dans l'énoncé[3]. Ainsi, toujours à partir de la figure 4.1, les dépendances syntaxiques élémentaires seront : {support→en→béton}, {support→de→ligne}, {ligne→aérienne} et {ligne→électrique} (voir également figure 5.6 page 77).

Les dépendances élémentaires n'ont pas forcément de réalisation effective dans le corpus, elle peuvent représenter des syntagmes discontinus. Elles ne sont donc pas interprétables au même titre que des syntagmes. Par exemple la dépendance *valeur à gamme* (`/frNom{Csem=4;Morpho=2} VALEUR/frPrep A/frNom{Csem=6;Morpho=2;D=0} GAMME`) n'a pas de sens considérée isolément, elle doit être replacée dans son contexte : «*valeur moyenne quadratique à large gamme de la tension des chambres neutroniques*». Il peut aussi arriver que des syntagmes soient constitués d'une unique dépendance, ainsi le descripteur du thesaurus EDF *générateur à induction*; dans ce cas la dépendance est interprétable comme un syntagme (désormais *dépendance* aura la signification de *dépendance élémentaire*).

D'autres dépendances sont issues d'une analyse syntaxique erronnée. Par exemple la dépendance *âge de fréquence* (`/frNom{Csem=4;Morpho=4} AGE/frPrep DE/frNom{Xcons=5;-Morpho=2;Csem=4;D=d} FREQUENCE`) est étrange. Lorsque l'on considère le syntagme d'origine «*évolution en fonction du temps et de l'âge de la fréquence des intervention en fonctionnement et en entretien des différents matériels*» on voit qu'il s'agit ici d'une erreur de l'analyseur *AlethIP*, *fréquence* est en réalité argument du nom prédicatif *évolution* alors qu'il a été analysé à tort comme modifieur du nom *âge*.

4.2 Le problème de l'étiquetage sémantique

L'approche distributionnelle permet de faire dériver le composant sémantique de la syntaxe; d'autres approches développent un composant sémantique indépendant de la syntaxe. Paradoxalement, alors que nous avons choisi d'exploiter la syntaxe comme le suggère la méthodologie distributionnelle, nous avons choisi une sémantique indépendante de la syntaxe qui emprunte sa forme à des catégories sémantiques préexistantes au texte. La principale raison est que les classes sémantiques produites avec la méthode distributionnelle souffrent d'un certain nombre de faiblesses : leur cohérence n'est pas toujours garantie si elles sont produites automatiquement, et il faut les nommer; or un traitement automatique ne permet pas de nommer de telles classes, l'intervention humaine est requise et demande un travail d'interprétation coûteux en temps. Une autre raison est que les corpus EDF ne se limitent pas à un ou deux domaines de spécialité; comme nous l'avons esquissé au chapitre 3.2, ces domaines sont très nombreux et couvrent de très vastes champs disciplinaires. Il ne pouvait donc être question d'induire des classes sémantiques propres à ces nombreux domaines. Nous avons donc choisi de faire appel à des classes sémantiques indépendantes de tout domaine de spécialité. En récupérant la couche sémantique du dictionnaire AlethDic (voir en annexe A), nous avons pu tirer profit d'un lexique existant et économiser de longues heures de travail. Après une simplification des classes sémantiques du dictionnaire AlethDic (voir en annexe A.2.1) nécessaire pour

3. Cette dernière contrainte permet de distinguer les adjectifs antéposés des adjectifs postposés.

faciliter le processus de désambiguïsation, nous avons disposé d'un lexique sémantique assez large, projetable sur le corpus ARD.

Mais pourquoi appliquer l'analyse distributionnelle – qui a fait ses preuves sur l'analyse de sous-langages (langues de spécialité) – au problème documentaire de la détection de pertinence? Pour répondre à cette question nous faisons l'hypothèse que la frontière entre deux domaines de spécialité (tel syntagme est du ressort de la chimie, tel autre de la médecine) n'est pas moins ténue que celle qu'il y a entre deux points de vue documentaires (tel syntagme est pertinent, tel autre ne l'est pas)? Il faudrait aussi souligner que la définition de ce qu'est un domaine[4] et l'établissement d'une frontière entre les domaines est bien souvent problématique étant donné parfois l'intrication et la dépendance des disciplines entre elles. En revanche, la définition d'un critère de pertinence s'appuie simplement sur le jugement, l'intérêt et la connaissance qu'a un individu pour une question. Dans ces deux cas, nous faisons l'hypothèse qu'il est possible de s'appuyer sur la distribution du lexique dans les syntagmes.

4.2.1 Les possibilités offertes par un étiquetage sémantique

Contraindre des descriptions syntaxiques et simuler des phénomènes de restrictions de sélection sont les deux possibilités offertes par l'utilisation d'étiquettes sémantiques. Il faut déterminer pour cela quelle forme d'étiquetage doit être employé, quelle vision du sens doit permettre d'attribuer les étiquettes.

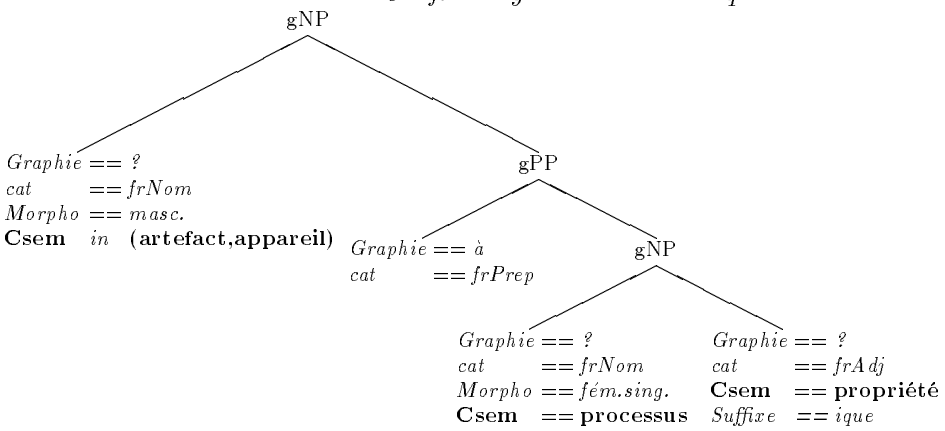
Concevoir des filtres syntaxico-sémantiques

L'assignation de catégories sémantiques aux noms, adjectifs, et adverbes, combinée à une analyse syntaxique, conduit à manipuler des structures syntaxiques enrichies. Ce type de représentation permet aussi bien la description des séquences rencontrées en corpus que la spécification de filtres syntaxico-sémantiques utilisables pour l'extraction de groupes nominaux dans les textes.

Un filtre syntaxico-sémantique comme celui de la figure 4.2 est plus précis que le patron syntaxique correspondant, qu'il soit plat : NOM *à* NOM ADJ ou structuré : [NP [N SP [Prep SN [Nom Adj]]]]. Sur les feuilles de cet arbre, on pose les contraintes de filtrage. Ainsi, ce filtre peut retenir des groupes nominaux comme "chauffage à accumulation dynamique", "câble à isolation synthétique", "contacteur à ouverture automatique". Nous verrons toutefois que de tels arbres de filtrage sont difficiles à mettre en oeuvre puisque leur succès dépend en grande partie de la capacité de l'analyseur syntaxique à faire une analyse parfaite – ce qui est rare avec les analyseurs robustes utilisés en milieu industriel. (c'est pour cette raison que l'on choisira un

4. Peut-être confondons-nous ici les domaines de la terminologie, définis strictement comme des domaines de connaissance circonscrits par les disciplines scientifiques (version européenne), et les domaines d'activité (version anglosaxonne?). Mais peut-être n'y a-t-il pas de différence entre les deux.

FIG. 4.2 – *Un filtre syntaxico-sémantique*



principe de filtrage différent, basé sur les RSE – voir sections 5.2.1 et 6.1 – qui s'adapte mieux aux erreurs d'analyse syntaxique).

«Guider» l'analyse distributionnelle : isoler des paradigmes lexico-sémantiques pour la généralisation de configurations linguistiques

Nous empruntons à [MF95] l'idée de généralisation de configuration linguistique. Leur module de généralisation prend en entrée les sorties d'un module de calcul de cooccurrences et trouve des régularités parmi les cooccurrences fournies. Les régularités sont recherchées à partir de la forme lexicale et du concept associé qui est donné par un thesaurus. Il s'agit donc de considérer parmi un ensemble de configurations linguistiques, celles qui peuvent être regroupées sur la base de traits communs. Par exemple «thèse imprimée» et «note rédigée» sont tous les deux des documents, ou plus généralement des objets sémiotiques.

De ce point de vue, les étiquettes sémantiques permettent de caractériser sémantiquement et de manière très générale un opérateur (ou un nom modifié) et ses arguments (ou modifieurs). Par exemple dans la table 4.1, on peut voir les modifieurs prépositionnels introduits par la préposition *de*, qui modifient le nom *température*. On peut y lire que les modifieurs de température peuvent être au choix : un nom de processus ou d'activité, un nom d'artefact, un nom de substance, un nom de lieu, ou un nom abstrait. La table 4.2 montre un paradigme d'objets sémiotiques et certains de leurs modifieurs. Ces informations permettent d'amorcer la création de catégories sémantiques propres au corpus [Nau96]. Par exemple les noms *note* et *thèse* pourront être placés dans une même catégorie bien qu'ils ne soient pas nécessairement modifiés par les mêmes adjectifs (*rédigé, publié*). De manière plus générale, ce type de description permet d'établir des restrictions de sélection propres à un corpus, un domaine ou une langue de spécialité. Cela prend tout son intérêt si on est en mesure de vérifier l'hypothèse harissienne, c'est-à-dire si des combinaisons opérateur/argument propres

TAB. 4.1 – *Exemple d'un paradigme de modifieurs pour le nom* temperature

/frNom{Csem=4 [a]} TEMPERATURE	/frPrep DE	/frNom{Csem=1;D=0 [b]}	AMBIANCE
/frNom{Csem=4} TEMPERATURE	/frPrep DE	/frNom{Csem=6;D=0}	REFERENCE
/frNom{Csem=4} TEMPERATURE	/frPrep DE	/frNom{Csem=24;D=0}	PAROI
/frNom{Csem=4} TEMPERATURE	/frPrep DE	/frNom{Csem=25;D=0}	ARC
/frNom{Csem=4} TEMPERATURE	/frPrep DE	/frNom{Csem=25;D=0}	TUYAUTERIE
/frNom{Csem=4} TEMPERATURE	/frPrep DE	/frNom{Csem=25;Xcons=3;D=0}	ADAPTATEUR
/frNom{Csem=4} TEMPERATURE	/frPrep DE	/frNom{Csem=43;D=0}	AIR
/frNom{Csem=4} TEMPERATURE	/frPrep DE	/frNom{Xcons=3 [c];Csem=47;D=0}	ENTREE
/frNom{Csem=4} TEMPERATURE	/frPrep DE	/frNom{Csem=47;D=0}	SURFACE
/frNom{Csem=4} TEMPERATURE	/frPrep DE	/frNom{Xcons=1;Csem=72;D=0}	DEMARRAGE
/frNom{Csem=4} TEMPERATURE	/frPrep DE	/frNom{Xcons=1;Csem=72;D=0}	ELABORATION
/frNom{Csem=4} TEMPERATURE	/frPrep DE	/frNom{Xcons=1;Csem=72;D=0}	REFRIGERATION
/frNom{Csem=4} TEMPERATURE	/frPrep DE	/frNom{Xcons=1;Csem=72;D=0}	SOUFFLAGE
/frNom{Csem=4} TEMPERATURE	/frPrep DE	/frNom{Xcons=2;Csem=72;D=0}	TRANSITION
/frNom{Csem=4} TEMPERATURE	/frPrep DE	/frNom{Xcons=3;Csem=72;D=0}	CONTROLE
/frNom{Csem=4} TEMPERATURE	/frPrep DE	/frNom{Xcons=3;Csem=72;D=0}	FONCTIONNEMENT

[a] Pour les significations du trait `Csem`, voir A.2 en annexe A

[b] absence de déterminant. Les autres valeurs sont données en A.5.1

[c] `Xcons=3` : nom non prédicatif mais admettant un attachement prépositionnel. Les valeurs de ce trait sont définies en table 5.3

TAB. 4.2 – *Exemple d'un paradigme de modifieurs adjectivaux pour des noms d'*OBJETS SÉMIOTIQUES *(Csem=37)*

/frNom{Csem=37;Morpho=1}	FICHE	/partPass{Morpho=1}	APPROUVE
/frNom{Csem=37;Morpho=1}	FORMULE	/partPass{Xcons=6;Morpho=1}	FOURNI
/frNom{Csem=37;Morpho=1}	GAMME	/adjStd{Morpho=1;Csem=148}	LARGE
/frNom{Csem=37;Morpho=1}	LISTE	/adjStd{Morpho=1}	OPTIONNEL
/frNom{Csem=37;Morpho=1}	NOTE	/partPass{Morpho=1}	PARU
/frNom{Csem=37;Morpho=1}	THESE	/partPass{Xcons=6;Morpho=1}	CONSACRE
/frNom{Csem=37;Morpho=2}	BANQUE	/adjStd{Morpho=2;Csem=142}	COMPLET
/frNom{Csem=37;Morpho=2}	BASE	/adjStd{Morpho=2;Csem=142}	RELATIONNEL
/frNom{Csem=37;Morpho=2}	GRILLE	/partPass{Morpho=2}	DETAILLE
/frNom{Csem=37;Morpho=2}	PHOTO	/adjStd{Morpho=1;Csem=142}	THERMIQUE
/frNom{Csem=37;Morpho=4}	MANUEL	/adjStd{Morpho=4;Csem=142}	INFORMATIQUE
/frNom{Csem=37;Morpho=4}	PAPIER	/partPass{Morpho=1}	DETAILLE

à une langue de spécialité sont spécifiques au domaine. Par un raisonnement inductif on peut alors glisser vers la généralisation puis la prédiction : certaines combinaisons lexicales et restrictions de sélection, constituent alors une signature du domaine pour lesquelles elles sont spécifiques.

4.2.2 Définir un jeu d'étiquettes sémantiques

Pour généraliser ces combinaisons lexicales sous la forme de restrictions de sélection, il faut que nous soyons en mesure d'affecter aux unités linguistiques (en particulier les noms) des étiquettes qui rendent compte de leur signification. Mais la caractérisation sémantique des segments de langue que rend possible un étiquetage sémantique dépend de la réalité linguistique des étiquettes. D'un point de vue linguistique général, la question de la signification d'une unité linguistique n'est pas résolue : elle fait l'objet de plusieurs hypothèses. La difficulté vient de l'articulation du couple signifiant-signifié avec le concept, ou plutôt les différentes acceptions du concept. L'antique modèle aristotélicien de la triade Mot/Concept/Chose toujours exploité dans les recherches linguistiques et cognitives [RCA94] ne fait pas l'unanimité. Il ne permet pas d'automatiser l'interprétation sémantique : il se heurte aux problèmes de la polysémie et du calcul du sens d'énoncés composés de multiples unités linguistiques. D'autres modèles de signification sont élaborés (par exemple les modèles inférenciels [Kay97], différenciels, ou leur version unifiée exposée dans [RCA94]) qui visent à dépasser ces limitations.

Mais la question de la connaissance n'est jamais résolue et se pose dans le champ de la linguistique de la manière suivante : quels rapports y a-t-il entre le sens linguistique et le sens conceptuel ? Quelle réalité accorder au concept ?

Si l'on admet que :

1. Le sens linguistique propre à chaque langue est lié à ses formes linguistiques,

2. Le sens conceptuel commun à toutes les langues est libre de toute forme linguistique et reflète la pensée humaine appréhendée dans un certain état de son évolution,

alors le rôle de la sémantique est d'étudier les modalités du va-et-vient entre le sens conceptuel et la forme linguistique à laquelle il se mêle pour créer un sens spécifiquement linguistique. Elle doit donc aborder les deux aspects linguistique et conceptuel. Mais le sémanticien est confronté à une difficulté d'obervation : il ne peut observer directement les pensées qu'il est en train de formuler. La nature du concept et les lois de formulation et d'interprétation linguistiques lui sont cachées par sa propre activité de pensée. Il est donc conduit à formuler des hypothèses sur des réalités qu'il ne peut observer directement. Ainsi chez Saussure, il est question du signifié et du concept, mais le rapport entre les deux n'est pas clairement établi [Ras91]. F. Rastier y voit la cause de la dualité entre le concept logique et le concept psychologique : « *La réduction du signifié au concept logique reste à la base de la sémantique vériconditionnelle. La réduction du signifié au concept psychologique est*

à la base de la sémantique "psychologique" ou cognitive » [Ras91] p. 74. Les six acceptions différentes du concept que F. Rastier énumère (p.125) révèlent des sensibilités phénoménologiques différentes chez les sémanticiens. Dans ces conditions, est-il possible de donner un caractère véritablement scientifique à une sémantique, est-il possible de fonder une sémantique sur une conception qui emprunte son mode d'observation à l'esprit de l'observateur plutôt qu'à la nature de l'objet observé[5]?

Cette question se pose systématiquement dès que nous cherchons à caractériser sémantiquement un mot, en lui assignant une catégorie sémantique par exemple. Mais d'un point de vue linguistique informatique, nous verrons que nous sommes forcés d'adopter une position beaucoup plus pragmatique pour l'étiquetage sémantique. Nous exposons les principales approches du sens linguistique afin de situer notre choix en matière de catégorisation sémantique.

Différentes approches du sens linguistique

Récanati dans [Ré97] fait un panorama des principales conceptions en matière d'approche du sens linguistique. Il dégage quatre principales tendances – le fixisme, le ségrégationnisme, le contextualisme modéré, et enfin le contextualisme radical – que nous exposons ci-après d'après l'article [Ré97].

Le fixisme La tradition fixiste définit le sens de chaque mot comme un noyau vériconditionnel, c'est à dire que l'usage d'un mot se fait en accord avec des conditions d'applications bien déterminées dans le monde. La prise en compte de plusieurs noyaux dans un énoncé conduit au calcul de sa valeur de vérité, au sens de Frege. Cette conception rencontre des difficultés dans au moins deux situations. La première est le traitement des expressions indexicales, dont le sens varie contextuellement. Dans ces cas, les conventions de langage ne fixent pas le sens, mais une signification linguistique «intermédiaire», qui détermine le sens en contexte. Ce phénomène est traité par le fixisme comme une exception. La deuxième situation concerne les phénomènes de polysémie. L'ambiguïté d'une expression se manifeste lorsque les conventions de langage lui associent plusieurs ensembles de conditions de satisfactions. *Les différentes acceptions d'une expression ambiguë sont chacune «fixes» et indépendantes du contexte. C'est la sélection d'une acception ou d'une autre qui est contextuelle* [Ré97]. Il n'est donc pas question de calcul du sens d'un mot à partir du contexte, le sens du mot est déjà préexistant, le contexte ne fait que marquer sa présence. Cette conception limite le processus d'interprétation à un parcours de possibilités, mais ne laisse pas de place à une élaboration progressive du sens au cours de l'interprétation.

Le ségrégationnisme Cette conception quant à elle repose sur un modèle de génération : le sens n'est pas sélectionné mais élaboré en fonction du contexte. Contrai-

5. Idée tirée de l'ouvrage [Ste23] dans lequel des critères de validité d'une science sont énoncés. La citation originale exprime la nécessité « *d'une conception* [scientifique] *qui n'emprunte pas son mode d'observation à l'esprit de l'observateur mais à la nature de l'objet observé*».

rement aux modèles de sélection de sens, un nombre indéfinis de sens peut être engendré. L'approche ségrégationniste, qui distingue deux niveaux : le sens linguistique littéral et les facteurs extralinguistiques, asserte que le sens est engendré par l'interaction sémantique des éléments de la phrase entre eux. Le contexte est un co-texte. Un exemple est l'approche de C. Fuchs et B. Victorri. Elle est définie dans [FV92] : «*(...) on peut associer à toute expression polysémique un noyau de sens, sous-déterminé, à partir duquel peuvent se déployer une pluralité de significations construites de manière dynamique au cours du processus d'interprétation, en fonction d'indices contextuels* [linguistiques]. *Il ne s'agit donc pas d'un processus d'élimination ou de choix forcé à partir de significations pré-déterminées, mais d'un processus de construction dans lequel l'interprétation d'une expression s'affine au fur et à mesure de l'analyse de l'ensemble de l'énoncé*».

Le contextualisme modéré Toujours dans [Ré97] : «*Le contextualisme rejette l'idée que les énoncés ont des conditions de satisfaction en vertu purement de leur signification linguistique (fixiste ou générative). Le sens – les conditions de satisfaction varient systèmatiquement avec le contexte, car les mécanismes de génération de sens sont, d'emblée, des mécanismes pragmatiques faisant appel et au savoir encyclopédique. Le contextualisme modéré attribue aux mots des significations fixes, mais qui ne se confondent pas avec la «valeur sémantique» qu'ils endossent contextuellement, et qui est engendrée en contexte. A partir de cette signification linguistique fixe, le contexte tant linguistique qu'extra-linguistique permet d'engendrer une valeur sémantique.* » Le contexte est ici défini au sens large et comprend le contexte linguistique et le contexte extra-linguistique.

Le contextualisme radical *Le contextualisme radical est aux antipodes du fixisme : il abandonne l'idée que les mots possèdent une signification linguistique fixe. Certes les mots – indépendamment du contexte – doivent bien apporter quelque chose qui les distingue les uns des autres et dont l'interaction avec le contexte engendre la valeur sémantique. Cette contribution peut être très différente d'une signification linguistique au sens traditionnel*» [Ré97]. La conception harrissienne du sens se rapproche certainement du contextualisme radical.

Les limitations du fixisme

Du point de vue des faits linguistiques, nous rejetons la position fixiste comme la définit [Ré97]. Nous pensons qu'elle ne permet pas de refléter toute la réalité des faits linguistiques.

Double articulation du langage et économie linguistique La théorie du fixisme ne prend pas en compte le principe d'économie linguistique formulé par A. Martinet, ainsi la polysémie n'y est pas traitée comme un phénomène fréquent et naturel, mais comme une exception. D'après [Mar70], la polysémie n'est pas une exception mais une caractéristique générale du signe linguistique. Et c'est en vertu

du principe d'économie linguistique que des unités existantes sont réemployées pour d'autres significations. En effet le seul principe de la double articulation du langage [Mar70], grâce aux possibilités infinies qu'il offre pour la création d'unités significatives, pourrait venir à bout des besoins communicatifs. Une telle économie «du nombre de signes» à mémoriser est réalisée grâce au phénomène de la polysémie et la possibilité de composer du sens en contexte.

Une conception du sens marquée par la logique Le fixisme est fortement marqué par l'influence des logiciens-positivistes. Il fait du sens d'un mot un noyau vériconditionnel qui se distingue par ses conditions d'applications. Le sens est ainsi formalisable et peut être exploité par la logique (calcul des prédicats) qui attribue à chaque énoncé une valeur de vérité. Cette conception calculatoire objective le sens et le fige dans les conventions de langues. Elle définit un univers sémantique statique, prédéterminé par des significations fixées. Or la langue est vivante et ne reste pas fixée dans ses conventions, elle évolue sans cesse : chacun est libre de redéfinir, pour un usage personnel ou communautaire (terminologies) la signification de n'importe quel signe ou son. C'est cela qui fait qu'une langue est vivante[6] [Mar70]. On ne peut rendre compte d'une telle créativité potentielle que par des processus dynamiques comme le principe de compositionnalité.

Fixisme et compositionnalité Une acceptation de la thèse fixiste implique un rejet du principe de compositionnalité. Les expressions polylexicales comme les noms composés (on exclut ici les expressions figées), seraient considérées comme des blocs dont les composants ne contribuent pas à l'interprétation de l'ensemble.

Le fixisme réduit le contexte linguistique au contexte sémiotique On ne peut attribuer à un signe un sens par défaut et parler ensuite d'exceptions pour rendre compte d'autres usages. Par contre on peut penser que la signification retenue pour un signe est celle qui est la plus courante ou la plus naturelle dans une communauté donnée, un contexte donné. Par exemple si l'on demande à plusieurs personnes à quoi leur fait penser le mot *base*, elles répondront différemment selon leurs préoccupations du moment ou les représentations et les réseaux connotatifs qu'elles se sont constitués au fil de leurs expériences. Par exemple, le mot *base* renverra entre autres à :

6. Les dictionnaires accusent toujours un certain retard. Adopter une position fixiste, c'est un peu comme si l'on justifiait le sens par l'existence des définitions. Or le sens ne tient pas à l'existence des définitions, mais à la permanence des représentations qu'évoquent les sons et les signes chez les individus. Cette permanence est maintenue pour rendre possible la communication linguistique. Le dictionnaire qui objective le sens et référence les usages les plus conventionnels joue un rôle de repère caduc.

une notion de contenu	*base de donnée*
une notion d'agent chimique	*base faible*
une notion de lieu qui concentre une activité	*base aérienne, base militaire*
une notion de figuration géométrique spatiale	*base du triangle*
une notion de localisation spatiale	*base du fauteuil*
une notion de point de départ, d'ingrédient fondamental.	*«la base de la sauce béchamelle est le roux blanc»*

La notion de contexte linguistique doit ainsi être étendue très largement. On pourrait le définir comme l'ensemble des perceptions et des connaissances d'un individu à un instant donné. Dans [FV92], le contexte est conçu comme *indissociablement sémantique et pragmatique et son traitement relevant de l'extralinguistique*. La notion de contexte doit intègrer la notion d'individu puisque tout énoncé oral ou écrit est interprété par un ou plusieurs individus. L'interprétation automatique a finalement peu de choses en commun avec l'interprétation humaine car la machine ne peut que simuler des scénarios à partir d'un contexte réduit à une suite de signes.

Contextualisme modéré ou radical

La polysémie est un phénomène propre aux langues naturelles, ce n'est pas une exception comme le considère le fixisme : un nombre indéfini de sens peuvent être associés à une forme. Un argument en faveur du contextualisme radical est le rapport entre le sens d'un signe et son contexte : le contexte a un rôle définitoire plutôt que seulement désambiguïsant. Cela est montré par les exemples de P. Cadiot dans [PC97] autour du mot *client* que nous reproduisons ci-dessous :

- Un cavalier s'adresse à un autre cavalier qui s'apprête à monter tel cheval : «Tu te méfieras, c'est un **client** un peu vicieux parfois»
- Un tueur à gages demande à son commanditaire : «Qui est mon **client** cette fois-ci ?».
- Un commentateur sportif à propos du prochain adversaire d'une équipe de football «Le prochain **client** d'Auxerre en Championnat d'Europe sera d'une autre trempe».
- Un journaliste à propos d'un homme politique : «C'est un **client** plutôt facile».
- Une mère qui vient chercher un de ses enfants à l'école : «Bon, je file, j'ai un autre **client** à prendre à la maison qui risque de se réveiller».
- Un déménageur à ses collègues : à propos d'un meuble : «Va falloir faire très gaffe, le prochain **client** coûte la peau des fesses».
- Un astronome à un de ses collègues dans le cadre d'un travail collectif : «Ton **client** à toi c'est Jupiter».

Dans ces exemples, il n'y aucune occurrence de *client* qui renvoie à une relation

marchande entre deux individus. Il faut bien alors invoquer le contexte linguistique au sens large pour expliquer de telles acceptions. De ce point de vue, qui donne au contexte un pouvoir de détermination sémantique supérieur à la signification isolée d'un signe, les rapports entre le signe et son contexte deviennent complexes. D'un point de vue méthodologique, on a coutume d'isoler le signe de son contexte, mais en réalité, il se fond dans le contexte. Il n'y a pas de différence entre ce que l'on appelle du point de vue analytique le signe et son contexte. Les deux sont une réalité indissociable. Cela est souligné dans [RCA94] : « *Le signe isolé hors contexte est en lui-même un artefact (...).* ». L'ambiguïté naît lorsque le(s) signe(s) est(sont) extrait(s) du contexte, ou lorsque le contexte est trop peu déterminant. Un signe linguistique est une information est saillante dans un contexte, à l'instar d'une tâche de couleur dans un paysage. Ce n'est qu'après considération de ses rapports avec l'ensemble qu'une signification peut lui être attribuée.

Peut-on affirmer pour autant que les unités linguistiques n'ont absolument aucune signification hors-contexte ? Dans les exemples de *client*, il y a bien un point de départ, un noyau sémantique peut être isolé : le client est «quelque chose à quoi l'on a affaire». Mais cet noyau n'est isolé qu'après une analyse des exemples. Il n'y a aucune garantie pour que celui soit fixe, d'autres exemples pourraient mettre en évidence d'autres noyaux sémantiques, aujourd'hui ou demain, l'évolution de la langue ne pouvant être prédite.

L'abîme entre les conceptions du sens et leurs implémentations possibles

Ce que l'on cherche à faire réaliser aux machines, c'est passer du signe au concept sans la participation de l'individu au moment où cette équivalence est réalisée. Il en résulte un abîme entre les positions que l'on défend sur une conception du sens (comme le contextualisme) et l'implémentation possible de cette conception. En effet le processus d'automatisation est réducteur et ne peut rendre compte de la richesse de l'expérience réelle du sens :

- Tout d'abord, la numérisation implique, au niveau élémentaire du traitement de l'information, la manipulations de signaux. La machine ne traite que des signaux (électriques) binaires. Ainsi tous les signes sont convertis sous cette forme. Du sens, déjà traduit sous forme de signes dans la langue écrite, il ne reste que des signaux en machine. La transformation du sens en signes, puis des signes en signaux [Ras91] est la principale difficulté qui s'érige contre la manipulation du sens par un automate.

- Il en résulte que la notion de contexte au sens large, riche de l'expérience humaine, est purement réduite à un contexte de signes-signaux; ce qui correspond à la réduction d'un objet immatériel (le sens) à une forme matérielle minimale.

- Dans ces conditions à quoi d'autre qu'une adresse mémoire, un signe peut-il renvoyer pour une machine?

    1. Il ne peut renvoyer à des représentations ressenties.

2. Il ne peut renvoyer à des concepts, ces derniers ne pouvant être atteints qu'au moyen de représentations.
3. Il ne peut que renvoyer à un sens artefactuel, issu de manipulations sémiotiques, c'est-à-dire à d'autres ensembles de signes.
   (a) Cela peut être une équivalence de type sémiotique exprimée sous la forme de chaînes de caractères. Par exemple, on décide de mettre en relation la forme *voiture* avec la forme VÉHICULE, en convenant que VÉHICULE représente un concept.
   (b) Une autre solution est de s'appuyer sur la réduction du contexte au contexte des signes linguistiques mémorisés en machine. Pour la machine, les signifiants sont la seule source de connaissance de la langue. La linguistique de corpus exploite cette possibilité lorsqu'elle énonce cette hypothèse distributionnelle : «deux formes qui partagent de nombreux contextes différents ont une proximité sémantique». Définir le sens sur cette base sémiotique-matérielle donne certains résultats et permet de regrouper des signes ayant des conditions d'application communes. Par contre seule l'interprétation humaine est en mesure de juger de l'intérêt et du sens des ensembles de signes ainsi regroupés [HN96].

Il en résulte que les points de vue fixiste et contextualiste modéré sont ceux qui peuvent être implémentés avec le plus de fidélité : le figement du sens dans une convention «tel signe-tel sens», le choix du sens sur la base du contexte linguistique.

4.2.3 Choix d'une catégorisation sémantique

D'un point de vue technique, le choix des catégories sémantiques projetées sur le lexique doit prendre en compte l'utilisation qui est faite de ces catégories. En effet, il n'est nul besoin de catégories fines et subtiles si la tâche ne l'exige pas. D'autre part, le choix des étiquettes est aussi déterminé par les solutions de désambiguïsation disponibles. Le choix de la méthode de désambiguïsation peut en effet avoir un impact sur le nombre des catégories à définir et leur granularité sémantique. Mais tout d'abord, il convient de faire un choix entre deux approches distinctes, auxquelles nous faisions allusion précédemment lorsqu'il était question des deux principaux types de représentation du sens en machine. La première, basée sur une équivalence sémiotique exprimée par des chaînes de caractère est développée ci-après comme une approche dite «fixiste exogène». La seconde, fondée sur le contexte sémiotique de l'unité à renseigner, est dite «contextualiste endogène».

L'approche fixiste exogène

Par approche exogène, il faut entendre que les informations de signification sont recherchées à l'extérieur du corpus. Pratiquement cela signifie que l'on fait appel

à une ressource comme un dictionnaire ou un thesaurus et que l'on projette sur le lexique du corpus les informations sémantiques données par cette ressource. Cette approche est nécessairement fixiste : aux unités linguistiques sont associées une ou plusieurs significations déterminées par des conventions de langue (l'usage référencé dans un dictionnaire ou une autre ressource).

Les catégories sémantiques sont en général définies à partir de la connaissance que l'on a sur le monde. C'est donc le fruit d'un travail subjectif et introspectif. Il en résulte que les choix effectués sont toujours discutables, étant donné qu'il y a un nombre indéfini de solutions pour découper la réalité et faire la typologie de ses objets (artefact, entité abstraite, substance naturelle, ...). La granularité des catégories sémantiques, c'est-à-dire l'échelle de généralité ou spécificité des catégories d'objets désignées par les catégories, est aussi dépendante de choix subjectifs individuels : doit-on distinguer entre les substances naturelles, les substances chimiques produites par l'homme, les substances comestibles, les substances liquides, etc. Ces découpages sont toujours matière à discussion sans que l'on puisse trancher assurément. Un autre point faible de cette approche est que pour être en mesure de traiter des corpus d'actualité, elle nécessite une mise à jour des nouvelles catégories qui peuvent faire leur apparition, ainsi que du lexique correspondant à l'ensemble des catégories définies. Pratiquement cela pose des problèmes de maintenance lexicale. En revanche, cette approche fixiste ne nécessite pas la mise en oeuvre de calculs sémantiques complexes, le sens des unités linguistiques étant fixé.

Exemple d'exploitation d'un système lexical «fixiste» P. Resnik dans [Res93] exploite la taxonomie de *WordNet* (nous faisons en 4.2.4 une brève présentation de la base lexicale *WordNet*). Celle-ci peut être considérée comme relevant d'une approche fixiste. Resnik projette sur le lexique de son corpus les catégories conceptuelles de *WordNet* pour mettre en évidence des restrictions de sélection. De telles restrictions de sélection sont appliquées pour résoudre des attachements prépositionnels et des ambiguïtés de coordination et définit un score de «préférence sélectionnelle» établi sur la probabilité de cooccurrence du verbe et de la classe sémantique dont relève l'objet du verbe. Ce type d'approche s'oppose à d'autres travaux sur le calcul de restrictions de sélection qui s'appuie sur l'approche contextuelle de l'analyse distributionnelle (par exemple ceux de D. Hindle dans [Hin90][7].)

Nous aborderons plus avant les solutions en matière de désambiguïsation.

7. Hindle dans [Hin90] calcule des contraintes de sélection sur les associations binaires de type verbe-sujet ou verbe-objet à partir de leur score d'information mutuelle. Il exploite pour cela un corpus de 6 millions de mots analysé par un analyseur robuste. Des classes sémantiques sont ensuite calculées à partir de ces contraintes de sélection en comparant le score d'association des deux noms avec chacun des verbes du corpus.

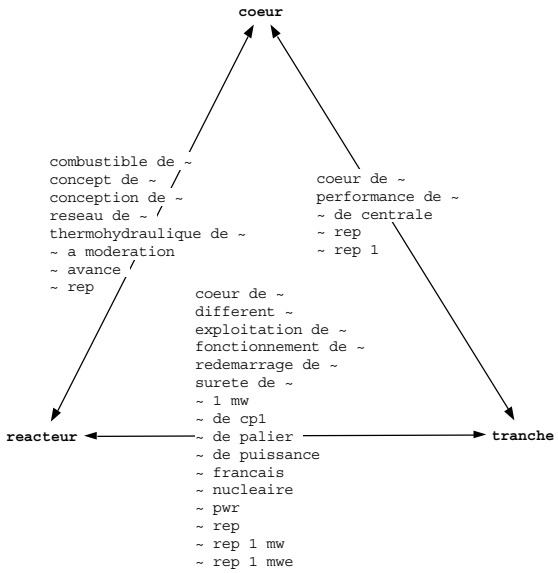
L'approche contextualiste endogène

Cette approche consiste à faire émerger d'un corpus ses propres catégories sémantiques, conformes au découpage que fait la langue de la réalité. Le principe de la méthode a été présenté en 4.1.2 lors de la présentation de l'analyse distributionnelle. Une catégorie sémantique représente un ensemble de contextes partagés par plusieurs unités linguistiques. L'hypothèse contextualiste sous-jacente est que deux formes qui partagent de nombreux contextes différents, ont une proximité sémantique. On obtient ainsi des classes sémantiques constituées à partir de données lexico-syntaxiques où la subjectivité du linguiste n'intervient pas directement. L'approche est ainsi endogène : aucune information sémantique n'est introduite dans le corpus. Les catégories sont alors dépendantes du découpage de la réalité que fait la langue (ou la langue de spécialité), et contrairement à l'approche exogène précédente, ne se veulent pas «universelles».

Il convient de distinguer l'approche endogène sur la langue et sur un corpus. Ainsi les classes d'objet de Gross [Gro94] sont construites à partir d'une démarche endogène sur la langue. La procédure ne peut être automatisée : les classes d'objets assimilables à des catégories sémantiques sont définies à partir des usages syntaxiques. L'intervention du jugement humain est nécessaire pour réaliser des tests linguistiques comme la commutation des arguments d'un opérateur.

Un exemple de système endogène Nous avons déjà évoqué en 4.1.2 les systèmes *SEXTANT* [Gre94] et *ZELLIG* [HNN96, BHNZ97]. Le premier s'appuie sur une classification statistique des distributions lexico-syntaxiques et le second sur une classification symbolique des distributions lexico-syntaxiques. Nous avons testé le système *ZELLIG* (dans un état de développement antérieur à son état actuel) sur notre corpus EDF-ARD [Nau96]. Rappelons que *ZELLIG* prend en entrée des groupes nominaux qui sont décomposés en dépendances syntaxiques élémentaires. A partir de ces dernières sont construites des graphes de mots associés selon leurs contextes syntaxiques. Ces graphes ou composantes connexes sont construites sur un seuil correspondant au nombre de contextes différents que doivent partager deux mots pour être liés. *ZELLIG* construit également des cliques à partir des composantes connexes (c'est-à-dire les sous-graphes dans lesquels chaque noeud est relié à tous les autres noeuds, c'est-à-dire dont tous les couples de noeuds (mots) partagent des contextes syntaxiques communs). De part leurs propriétés formelles les cliques peuvent fournir des agrégats encore plus significatifs que les composantes connexes, ayant une meilleure cohérence conceptuelle. Ainsi, la figure 4.3 montre une clique (calculée pour un seuil de 5 contextes partagés) qui a lié les noms *coeur, tranche* et *réacteur* et dans laquelle l'expert pourra reconnaître une relation de méronymie entre *coeur* et *réacteur*, ainsi qu'entre *réacteur* et *tranche*. Contrairement à ce qui se passe lorsque la méthode est purement statistique (comme avec *SEXTANT*), les résultats présentent aussi les données qui ont permis de les produire. Ainsi sur les arcs de la clique apparaissent les contextes syntaxiques partagés par les mots attachés aux noeuds.

FIG. 4.3 – *Exemple d'une clique produite par* ZELLIG



Comme cela est montré dans [HN96] et [BHNZ97] pour le langage médical, l'interprétation de tels agrégats est toujours nécessaire, ne serait-ce que pour nommer les classes. Par exemple pour la clique déjà mentionnée, on peut avancer une classe spécifique comme «artefact-partie de centrale nucléaire». Il n'est pas possible d'assigner automatiquement à ces paradigmes un statut sémantique. L'étude de [BHNZ97] montre également, en projetant sur les noeuds des composantes connexes les catégories conceptuelles de thesaurus médicaux, que les catégories tirées des deux systèmes, l'un fixiste (thesaurus médicaux), l'autre contextualiste (*ZELLIG*) ne se recouvrent pas toujours. Il y a un décalage entre les catégories conceptuelles et linguistiques; Cela laisse donc supposer une difficulté à faire cohabiter les deux approches par exemple pour mettre à jour une ontologie ou affecter un nom de catégorie à des classes de mots constituées automatiquement.

P. Resnik dans [Res93] estime qu'il est difficile d'évaluer et de comparer les classes produites. Leur coût calculatoire pour la remise à jour des classes est important. Il souligne également le problème de caractérisation sémantique des classes. Celles-ci sont distributionnelles, «*mais dans quelle mesure sont-elles syntaxiques ou sémantiques*»? Il note également que dans les approches distributionnelles tous les sens des mots sont assimilés et la polysémie est difficilement prise en compte.

Stratégies de désambiguïsation

Lorsque l'on dispose d'un lexique sémantique, plusieurs méthodes de désambiguisation sont possibles. Si les catégories sémantiques ont été constituées à partir d'une méthode endogène, la question de la désambiguïsation ne se pose pas : les contextes qui ont permis de définir ces catégories fournissent l'information nécessaire à la désambiguïsation. Mais si les catégories sémantiques utilisées sont exogènes, une méthode de désambiguïsation s'impose.

Nous distinguerons les méthodes sans apprentissage des méthodes avec apprentissage. Les méthodes avec apprentissage requièrent un corpus sémantiquement annoté avec le jeu de catégories sémantiques voulu. Elles dispensent l'opérateur d'ordonner les règles de catégorisation selon un ordre de spécificité décroissante des contextes désambiguïsants.

Méthodes sans apprentissage La désambiguïsation-catégorisation manuelle est la méthode la plus simple à mettre en oeuvre. Pour mettre en évidence des restrictions de sélection, Basili et al. [BPV93b] procèdent à une catégorisation sémantique manuelle de leur corpus de petite taille. Une telle catégorisation sémantique est nécessaire pour préparer des échantillons d'apprentissage correctement annotés. L'entreprise est plus ou moins ardue selon la taille du corpus et le jeu d'étiquettes utilisé. Des catégories très grossières demanderont moins de travail que des catégories subtiles : l'étiqueteur humain se posera beaucoup moins de questions. Une telle tâche est facilitée par l'utilisation d'une interface adaptée, qui permet d'accélérer l'attribution d'étiquettes en visualisant le corpus selon certains points de vue. Par exemple, le logiciel *CorTeCs* [Hei96] est un programme d'aide à la correction de l'étiquetage de textes catégorisés et lemmatisés. Les textes sur lesquels il permet de travailler doivent être préalablement segmentés en phrases et en mots et éventuellement étiquetés au niveau des mots. Le travail de correction est réalisé au moyen d'une interface graphique permettant de parcourir le texte et de corriger les segmentations et l'étiquetage. L'interface permet de manipuler des données textuelles annotées, de visualiser des concordances, et de regrouper tous les mots ou les étiquettes identiques à la correction en cours. Ceci autorise la propagation dans l'ensemble du texte d'une correction réalisée en un point précis du texte. Enfin, plusieurs correcteurs peuvent progresser dans leur travail en plusieurs sessions, les corrections effectuées étant enregistrées au fur et à mesure.

La désambiguïsation manuelle peut être relayée par des règles qui permettent de généraliser les corrections manuelles. Ces règles prennent en compte des contextes spécifiés pour attribuer des catégories sémantiques à des formes. Par exemple le logiciel *SATO* [Dao96] *(Système d'Analyse de Textes par Ordinateur)*[8] combine annotation manuelle et à base de règles. Il est destiné à soutenir une variété d'activités d'analyse de données textuelles (analyses qualitatives ou quantitatives) et ne se li-

8. Ce produit est distribué par le Service d'Analyse de Textes par Ordinateur (ATO) de l'Université du Québec à Montréal

mite pas à l'annotation sémantique[9]. Avec des scénarios, et un système de requêtes, *SATO* peut définir des actions d'annotations qui s'appliquent automatiquement. Les contextes sont repérés au moyen de patrons de concordance. Mais dans ce cas l'ordre d'application des règles est important, les portées des expressions régulières exprimant les contextes pouvant entrer en conflit.

Méthode avec apprentissage Les méthodes avec apprentissage sont soit symboliques, soit stochastiques.

La méthode d'E. Brill [Bri92, Bri94] basée sur le principe de correction d'erreur est symbolique. Initialement développée pour faire de la catégorisation-désambiguïsation de catégories grammaticales, elle a pu être portée avec succès pour des tâches d'étiquetage syntaxique comme la résolution d'attachement prépositionnel (ramené à un problème de catégorisation) dans [BR94], ou le repérage de syntagmes dans [RM95]. La méthode d'E. Brill est en fait utilisable pour tout problème d'étiquetage. Il convient cependant d'être prudent car deux facteurs d'importance interviennent : la taille de l'échantillon d'apprentissage et la cardinalité du jeu d'étiquettes.

Voici comment la méthode devrait être mise en oeuvre pour un étiquetage sémantique : dans un premier temps, il faudrait définir un dictionnaire qui pour chaque nom, adjectif et adverbe donne sa catégorie sémantique la plus fréquente dans les textes à traiter. Ensuite, il faudrait définir un corpus de référence et disposer de deux exemplaires de ce corpus, l'un désambiguïsé à la main, l'autre, catégorisé automatiquement avec le dictionnaire. Bien entendu, on peut s'attendre à ce que la catégorisation du corpus avec le dictionnaire produise un certain nombre d'erreurs. Le processus d'apprentissage consisterait alors à construire des règles de correction d'erreur pour les formes du texte qui ont été mal catégorisées (par rapport au corpus étiqueté à la main, dont on suppose qu'il ne contient pas d'erreur).

Le contexte linguistique (morphologique, lexical, syntaxique voire sémantique) de la forme ambiguë dont l'étiquette doit être corrigée définit alors les conditions d'application de la règle de correction. L'assignation d'une nouvelle étiquette définit l'action de la règle de correction.

Cependant, une règle donnée, qui corrige une forme dans un contexte spécifique peut causer une erreur de catégorisation dans un autre contexte assez similaire. Dans ce cas, la règle de correction choisie est celle qui est la plus efficace, son efficacité étant définie comme le nombre d'erreurs de catégorisation que la règle a corrigé moins le nombre d'erreurs de catégorisation que la règle a causé [dL95]. Cette méthode évite à l'utilisateur d'avoir à ordonner ses règles par ordre décroissant de spécificité des contextes déclencheurs. Le système se stabilise et statue sur le choix des meilleures règles à partir du critère d'efficacité.

9. Tout document est découpé en mots-graphiques et à partir de ces derniers un lexique du document est construit. Le document est ensuite représenté par une séquence d'entiers qui reproduit l'ordre des mots du document à partir de références au lexique. Le texte est traité sous sa forme graphique, aucune lemmatisation ou catégorisation ne précèdent les traitements. Sous forme ASCII, les enrichissements et annotations effectués dans le texte sont visibles grâce à un balisage.

Le point difficile est que le système est alimenté avec un corpus qui a été préalablement désambiguïsé avec le jeu d'étiquettes que l'on souhaite appliquer. Il convient alors de déterminer la taille minimale du corpus pour que l'apprentissage génère des règles dont la couverture est suffisante pour catégoriser d'autres textes. Le corpus Brown utilisé par Brill pour la phase d'apprentissage faisait environ 22000 mots (1000 phrases), pour un nombre de parties du discours n'excèdant pas 15. Mais dans le cas d'un jeu d'étiquettes sémantiques dont on peut supposer la taille plus importante, il faut certainement prévoir un corpus plus volumineux, représentatif des nombreuses configurations contextuelles des unités linguistiques à catégoriser.

Les chaînes de Markov à états cachés peuvent être utilisées pour la catégorisation grammaticale et l'analyse syntaxique probabiliste [Raj95]. Elles sont est également applicables à une tâche de catégorisation sémantique. Le principe, qui consiste à calculer la probabilité conditionnelle d'apparition d'une catégorie.

Enfin, D. Yarowsky dans [Yar92] présente un modèle de catégorisation sémantique particulier. La méthode pourrait être qualifiée d'endogène : elle ne demande pas de corpus d'apprentissage manuellement étiqueté. Ceci est rendu possible grâce à l'exploitation du thesaurus Roget. La méthode désambiguïse les mots de textes anglais en définissant un modèle statistique des catégories du thesaurus Roget qui servent d'approximation à des classes conceptuelles. Chaque catégorie du thesaurus est associée à une définition. L'évaluation du système est faite sur douze mots polysémiques et donne un taux de réussite de 92%. Voici comment D. Yarowsky procède :

1. La première phase est de rassembler des contextes qui sont représentatifs des catégories du thesaurus Roget. Le corpus utilisé est l'encyclopédie Grolier (10 millions de mots). La procédure consiste à prendre chaque catégorie du thesaurus Roget en considérant sa définition. Ensuite sont extraites du corpus toutes les concordances de 100 mots autour de chaque mot de la définition de la catégorie Roget. A ce stade la polysémie n'est pas prise ne compte. Certaines concordances seront donc parasites car sémantiquement étrangères pour la catégorie Roget utilisée.

2. La seconde phase consiste à identifier parmi les collections de contextes constituées quels sont les mots les plus saillants. Un mot saillant est un mot qui apparaît plus souvent dans le contexte d'une catégorie qu'en d'autres endroits du corpus, en leur affectant à chacun un poids.

3. Enfin, les poids calculés sont utilisés pour prédire la catégorie d'un mot polysémique qui apparaît dans de nouveaux textes.

4.2.4 Un exemple de système fixiste : *WordNet*

WordNet [MBF+90, BFGM90] est une importante ressource lexicale pour l'anglais. Les hypothèses qui ont conduit à architecturer celle-ci sont essentiellement

psycho-linguistiques et les choix effectués sont justifiés par des résultats expérimentaux en psychologie (catégorisation et organisation de lexique dans la mémoire, tests de rapidité d'association pour certaines relations lexicales). Le lexique y est organisé selon les catégories grammaticales (nom, adjectif, verbe, adverbe) et pour chacune de ces catégories l'information lexicale est représentée sous la forme d'un réseau différent. Ainsi les noms sont organisés dans la mémoire lexicale comme des hiérarchies conceptuelles. Les verbes sont organisés selon des relations de nécessité ou de cause à effet. Le point le plus intéressant dans l'architecture de *WordNet* est que ses concepteurs ont essayé d'organiser le lexique d'après le sens des mots plutôt que leur forme. De ce point de vue, *WordNet* ressemble plus à un thesaurus qu'un dictionnaire.

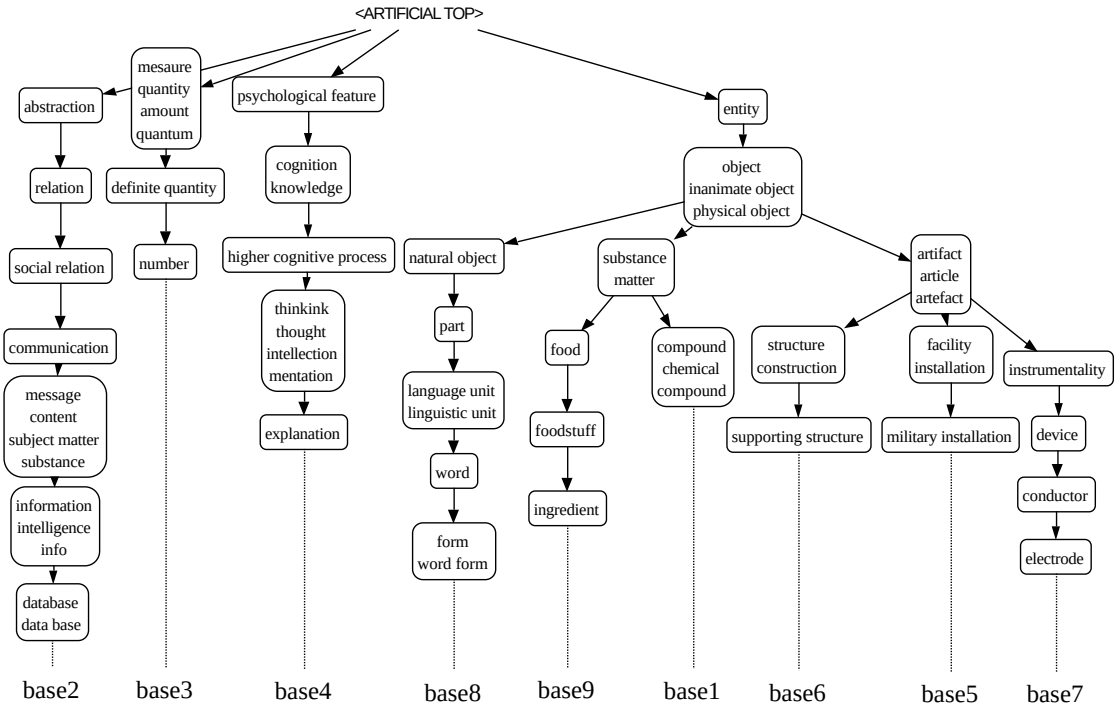
WordNet est conforme à l'approche fixiste. Il fait un inventaire du lexique et des concepts connus. Toutefois chaque sens ne se réduit pas à l'existence d'un forme unique. Les concepts lexicalisés sont représentés par des définitions. De plus chaque sens est représenté par un ensemble de synonymes (*synset*). Ces *synsets* n'expliquent pas le concept mais signalent son existence par ses différentes réalisation linguistiques, ce qui permet de l'identifier. La notion de sens dans *WordNet* repose également sur les relations qu'entretiennent les unités linguistiques entre elles. Ainsi, le réseau de noms déclare des relations de synonymie, d'antonymie, de méronymie, d'hyponymie. Le déploiement de toutes les relations d'hyponymie entre les noms crée une hiérarchie de concepts nominaux. Par exemple le nom *electrode* peut être considéré comme un nom d'*inanimate object* car il dominé dans la hiérarchie par *inanimate object*, les concepts nominaux intermédiaires étant : *conductor, device, instrumentality, artefact* (voir figure 4.4.). Un tel réseau permet de représenter les différentes acceptions d'une unité polysémique. Considérons par exemple le nom français *base* et neuf de ses acceptions données en table 4.3, celles-ci peuvent être localisées dans la sous-partie de la taxonomie de *WordNet* représentée en figure 4.4.

TAB. 4.3 – *Les différents sens de base, principalement d'après le Petit Robert*

base1	corps capable de neutraliser les acides en se combinant à eux
base2	ensemble de données, base de données
base3	nombre d'unités d'un certain ordre pour former un système de numération
base4	principe fondamental sur lequel repose un raisonnement
base5	zone militaire
base6	partie inférieure d'un objet sur laquelle il repose
base7	électrode de commande d'un transistor
base8	partie importante d'un mot : racine, radical
base9	ingrédient fondamental dans une préparation culinaire

Désambiguïsation sémantique avec *WordNet* La ressource *WordNet* est le point de départ de nombreux travaux de recherche sur la désambiguïsation de l'anglais. Cela a débuté avec l'annotation sémantique d'un corpus [GMS+94] et le développement d'une application *SemCor* permettant de visualiser des concordances

FIG. 4.4 – *Mise en relation des différents sens du mot français* base *(voir Tab. 4.3) dans la hiérarchie des concepts nominaux de WordNet*



d'étiquettes sémantiques. E. Voorhees [Voo93] et P. Resnik [Res95] l'utilisent pour faire de la désambiguïsation lexicale et améliorer la recherche documentaire. Dans [LTV95] il est exploité pour un apprentissage de contextes désambiguïsants. Dans [CGE96] il est utilisé pour construire un modèle statistique de désambiguïsation lexicale.

4.3 Conclusion

Choix d'une approche syntaxique De l'approche distributionnelle, nous n'exploitons que la méthode syntaxique qui apporte l'idée de normalisation syntaxique (mise en application dans [Sag87]). Nous l'adaptons au groupe nominal pour le normaliser sous la forme de dépendances élémentaires. Cela nous permet de définir des micro-patrons syntaxico-sémantiques et d'envisager le filtrage de syntagmes nominaux à partir de ces patrons. Nous laissons donc de côté tout le travail de constitution de classes sémantiques de la méthode distributionnelle.

Choix d'une approche sémantique En effet, contre nos convictions, et en raison des problèmes de faisabilité évoqués, nous n'adoptons pas le point de vue contex-

tualiste qui aurait pu s'appuyer sur une approche distributionnelle pour définir des catégories sémantiques. Nos catégories sémantiques proviendront donc d'une source extérieure au corpus. Parmi les stratégies de désambiguïsation que nous avons évoquées, notre choix s'est porté sur une méthode de désambiguïsation à base de règles symboliques, malgré l'intérêt évident des méthodes à base d'apprentissage. Nous expliquons ce choix dans le chapitre suivant.

Science vs. technologie Nous avons également donné notre point de vue sur la possibilité d'implémenter une théorie du sens. Etant donné le réductionnisme qu'impose le traitement informatique, il semblerait qu'à l'heure actuelle, la possiblité d'implémenter une telle théorie soit une preuve de son inadéquation à la langue. Les expériences de modélisation en linguistique informatique ne se conçoivent que comme des approximations de la réalité linguistique pour répondre à des besoins techniques. On se situe dans le domaine de la technologie.

Au stade de développement actuel des machines, la connaissance de la nature des concepts et de leurs rapports au langage n'apporterait pas de solutions techniques en linguistique informatique. Comme en IA, pour atteindre un objectif, on ne cherche plus à simuler exactement de ce qui se passe dans la tête de l'expert en terme de processus mentaux, mais on modélise une approximation en tenant compte des possibilités de la machine; l'essentiel étant le résultat. Il ne faut donc pas s'attendre à ce que, lorsque l'on introduit des catégories sémantiques, on puisse manipuler du sens. Ce ne sont que des catégories outils, de simples artefacts linguistiques, aussi bien dans leur version fixiste que contextualiste. Il est pertinent de s'interroger pour savoir si ces catégories – que nous présentons dans le chapitre qui suit – sont adéquates et permettent de résoudre le problème posé, mais nous pensons qu'il n'est pas pertinent de s'interroger sur leur réalité ou identité sémantique sans confondre moyen technique et réalité de la langue, technologie et science du langage.

Nous présentons à présent les traitements informatiques implémentés conformément aux choix syntaxiques et sémantiques effectués. Le principe du système de désambiguïsation y sera également présenté.

Chapitre 5

Manipuler du texte enrichi

Les phases de transformation de la séquence textuelle qui aboutissent à son enrichissement linguistique permettent, à partir de chaînes de caractères, de représenter du texte sous la forme de structures arbitrairement complexes auxquelles sont attachées des informations morphologiques, syntaxiques ou sémantiques. Cela offre de nouvelles possibilités de traitement – nous en avons évoqué certaines au chapitre précédent – qui vont nous permettre de développer notre système de filtrage. Dans le présent chapitre nous présentons les traitements d'enrichissement que nous effectuons, après récupération des sorties des outils d'analyse existants sur lesquels nous nous appuyons et qui ont été présentés au chapitre 3.

5.1 Apporter de la valeur ajoutée aux chaînes de caractères

TAB. 5.1 – *Exemple de valeurs sémantiques associées à des suffixes nominaux*

Suffixe	Base	Exemple	*Valeur*
–acées	nom	rosacées, cucurbitacées	*famille de plantes*
–ade	verbe	glissade, baignade	*action ou résultat de l'action*
–ade	nom	citronnade	*collectif*
–age	verbe	assemblage, serrage	*action*
–ail(le)	verbe	gouvernail, tenaille	*instrument*
–aire	nom	actionnaire, disquaire	*agent, fonction, métier*
–ateur	verbe	calculateur, utilisateur	*machine, agent*
–icien	nom	technicien, informaticien	*spécialiste de*

Aux noms et aux adjectifs est attachée une information de suffixe. Par exemple *-age* dans «surmen-*age*» et *-ique* dans «linguist-*ique*». Ces derniers sont déterminés à partir des données fournies dans [Gui70]. Ils sont intéressants car porteurs de valeurs sémantiques.

TAB. 5.2 – *Exemple de valeurs sémantiques associées à des suffixes d'adjectifs*

Suffixe	Base	Exemple	*Valeur*
–ain(e)	nom propre	romain, africain	*habitant de*
–ais(e)	mom	anglais, bordelais	*habitant de*
–al(e)	nom	racial, théatral	*qui appartient à*
–ant(e)	verbe	mutant, existant	*qui fait l'action du verbe*
–(at)ique	nom	géométrique, dogmatique, informatique	*relatif à*
–escent(e)	nom	fluorescent, luminescent	*qui a la qualité de < nom >*
–eur	verbe	échangeur, conducteur	*qui fait l'action du verbe*
–ible	verbe	flexible, traduisible	*qui peut être < verbe > , dont on peut < verbe >*
–oire	verbe	exploratoire, vibratoire, préparatoire	*qui participe à l'action du verbe*

Pour minimiser les temps de traitements et assurer une certaine robustesse, la reconnaissance des suffixes est faite sans grande précision par une simple recherche de sous-chaînes de caractères. Elle n'a rien a voir avec un processus de reconnaissance morphologique sophistiqué [CDGK94] et ne s'appuie pas sur une modélisation des suffixes comme par exemple celle de [CLB94]. Les tableaux 5.1 et 5.2 montrent des exemples de valeurs sémantiques associées à des suffixes respectivement de noms et d'adjectifs. La première colonne donne le suffixe, la seconde colonne indique à partir de quelle base le nom ou l'adjectif est dérivé. La troisième donne des exemples et la quatrième les valeurs sémantique fournies par L. Guilbert [Gui70]. Ce qui compte pour nous, ce n'est pas d'identifier strictement une valeur pour un suffixe donné (cela est problèmatique en soi, puisque certains suffixes sont ambigus) mais c'est que les suffixes aient une valeur, quelle qu'elle soit. Par ailleurs, les suffixes sont utilisés seulement dans le cas où aucune catégorie sémantique n'a pu être assignée à un nom ou un adjectif.

Des informations lexico-syntaxiques sont ajoutées : une information de prédicativité est attachée aux noms. La notion de nom prédicatif n'étant pas définie dans le dictionnaire *AlethDic*, une approximation a été définie en combinant certaines informations de ce dictionnaire. Le trait `Xcons` et ses différentes valeurs rendent compte de la prédicativité du nom, qui peut être considérée comme effective pour les valeurs `1`, `2` et `3`. La table 5.3 définit les différentes valeurs du trait `Xcons`.

Des étiquettes sémantiques sont attachées aux noms, aux adjectifs et aux adverbes. Le choix des étiquettes pour les noms et les adjectifs est discuté plus loin (section 4.2.2). Leurs listes exhaustives sont données en annexe A (A.2, A.3 et A.4).

TAB. 5.3 – *Définition des différentes valeurs du trait* `Xcons`

`Xcons=1`	le nom accepte des arguments introduits par des prépositions, suffixe en *-tion* ou *-age*, il existe une relation nom–verbe (exemple : *dérivation–dériver*); ainsi les noms : *collaboration, piquage*
`Xcons=2`	le nom accepte des arguments introduits par des prépositions, suffixe en *-tion* ou *-age*, pas de relation nom–verbe dans le dictionnaire; ainsi les noms : *prestation, adéquation.*
`Xcons=3`	le nom accepte des arguments introduits par des prépositions, il existe une relation nom–verbe; ainsi les noms : *analyse, collecteur*
`Xcons=5`	le nom accepte des arguments introduits par des prépositions; ainsi les noms : *forum sur (l'emploi), crue de (du fleuve), but de (l'action).*
`Xcons=6`	S'applique aux adjectifs, participes passés ou participes présents qui entrent dans des constructions du type *exempt de, accordé par, conduisant à*

5.2 Les traitements syntaxiques

5.2.1 Normalisation des groupes nominaux : décomposition en dépendances lexico-syntaxiques élémentaires

TAB. 5.4 – *Relations lexico-syntaxiques élémentaires générées à partir de l'arbre de la figure 5.1*

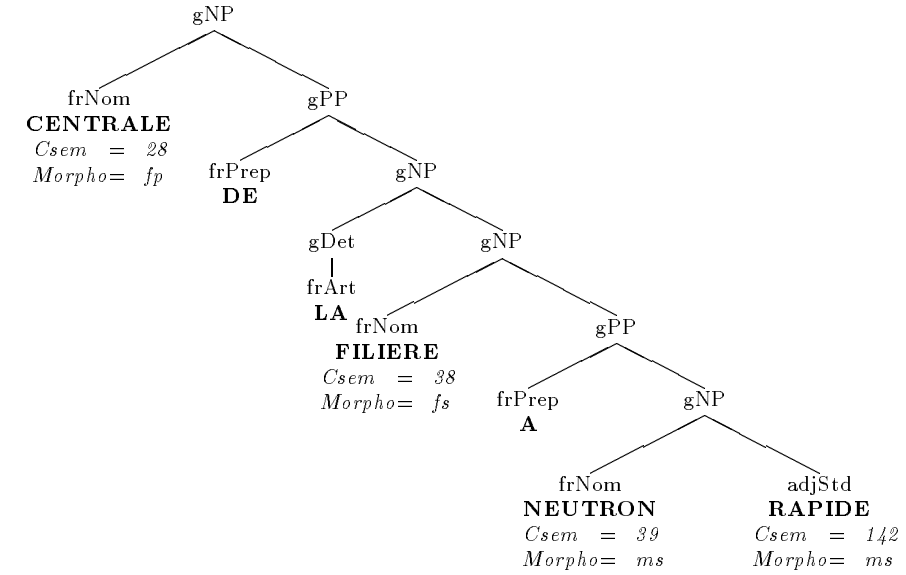
centrale	de	(la) filiere
filiere	à	neutron
neutron	rapide	

Une fois l'analyse du texte effectuée par *AlethIP*, l'extraction des dépendances lexicales attestées par la syntaxe devient possible. Etant donné que nous nous focalisons sur les groupes nominaux, nous nous intéresserons avant tout aux relations entre les noms et les adjectifs[1]. Les dépendances concernées correspondent donc principalement à des schémas du type NOM-ADJECTIF, NOM-PRÉPOSITION-NOM et NOM-NOM. Ces schémas, ici mis à plat, sont en fait des arbres élémentaires et prennent place dans la structure de l'arbre correspondant à l'analyse du groupe nominal. On donne en table 5.4 un exemple de dépendances produites à partir du syntagme «*centrales de la filière à neutrons rapides*» dont l'analyse est visible en figure 5.1. Les adjonctions de ces arbres élémentaires construisent un arbre syntaxique nominal. Cependant, ces arbres élémentaires ne sont pas équivalents à ceux des grammaires d'arbres adjoints (TAG). En effet, ils ne constituent pas des entrées dans un lexique grammaire. Ils ne sont pas définis a priori, ni porteurs de conditions d'adjonctions. Ils ne sont que le

1. Nous nous somme limités dans un premier temps à ces relations. Il serait nécessaire de prendre en compte également les adverbes

FIG. 5.1 – *Analyse syntaxique du syntagme «Centrales de la filière à neutrons rapides»*

.



résultat d'une décomposition d'un arbre syntaxique en dépendances binaires. En revanche l'analyse de ces arbres élémentaires dans un corpus donné doit être en mesure de faire ressortir des contraintes d'adjonctions empiriques. C'est dans cette optique qu'ils sont utilisés pour induire des grammaires de langue de spécialité dans [GH97].

Obtention des dépendances élémentaires

Etant donné la structure des arbres d'analyses d'*AlethIP*, nous avons utilisé une méthode qui n'est pas générale mais dépendante de la structure de ces arbres pour en extraire les dépendances.

La table 5.5 décrit l'algorithme d'extraction des dépendances. La fonction Explore(Arbre, Direction, Noeud de départ) parcourt l'arbre dans la direction demandée. Le parcours stoppe sur un noeud dont la catégorie est autorisée. Les catégories autorisées (nom, adjectif, inconnu, ...) sont déduites de la catégorie du noeud d'origine *T[X]* . Les dépendances sont extraites et enregistrées avec leur position dans l'arbre. Le booléen *T[X].deja* est mis à jour pour éviter la création de doublons sur la suite du parcours. Le recours à un tableau de pointeurs qui pointent sur les noeuds de l'arbre évite l'écriture d'une fonction à appel récursif pour le parcours de l'arbre. L'algorithme d'extraction des dépendances élémentaires s'en trouve simplfié et optimisé. L'algorithme n'est pas générique car la fonction d'exploration des relations de dominance ne fonctionne qu'avec des arbres générés par *AlethIP* (ce qui implique la prise en compte de la spécificité des étiquettes et des niveaux de profondeur propre à l'analyse d'*AlethIP*, pour le repérage des prépositions par rapport au nom modifié

TAB. 5.5 – *Algorithme d'extraction des dépendances élémentaires*

Soit *A*, l'arbre syntaxique du syntagme nominal représenté par une structure de type arbre n-aire.
Soit *T*, un tableau de pointeurs et de booléens. Le pointeur pointe vers les *Nmax* noeuds terminaux de l'arbre *A*. Et pour chacun d'entre eux, un drapeau *deja* marque s'il a déjà été mis en dépendance.
Chaque noeud *N* de l'arbre *A* est identifié par une constante arbitraire *ident* qui correspond à l'élément du tableau *T* qui pointe sur ce même noeud (*T[N→ident].noeud* == *N*).

```
Pour X de 0 à Nmax Faire

Pour tous les MODE de {Pere, Frere, Fils} Faire {

N = Explore (A, mode, T[X].noeud);

Si (N n'est pas nul) et (T[N→ident].deja est Faux)

Alors {
    ExtraireLaDépendance(T[X], N);
    T[X].deja = Vrai;
    }
}
```

ou modifiant par exemple).

La fonction ExtraireLaDépendance(*Noeud1, Noeud2*) imprime seulement les informations attachées aux deux noeuds de l'arbre passés en paramètres. Si les noeuds n'ont pas la même profondeur dans l'arbre, on imprime également la préposition qui introduit le noeud le plus profond et qui dépend du noeud le moins profond.

TAB. 5.6 – *Relations lexico-syntaxiques élémentaires générées à partir du syntagme : «support de ligne électrique aérienne en béton».*

support	en	béton
support	de	ligne
ligne	électrique	
ligne	aérienne	

Par exemple, si l'on considère l'arbre d'analyse représenté en figure 5.1. Le tableau à parcourir sera constitué des entrées suivantes : {*centrale,filière,neutron,rapide*}. Considérons en premier lieu *centrale* : il dépend d'un gNP (groupe nominal) qui domine un gPP (groupe prépositionnel) dont la tête nominale est *filière*. La première dépendance élémentaire sera donc *centrale* – *filière*, soit «$\mathbf{n}_1$=*centrale* **prep**=*de* **dét**=*défini* $\mathbf{n}_2$=*filière*» en conservant l'information de la préposition et du déterminant. Ensuite vient *filière* : il dépend d'un gNP qui domine un gPP introduit par la préposition *à* et dont la tête est *neutron*. Il dépend également de *centrale* mais ce dernier nom a déjà été mis en dépendance. La deuxième dépendance sera donc «$\mathbf{n}_1$=*filière* **prep**=*à* **dét**=∅ $\mathbf{n}_2$=*neutron*». Ensuite vient *neutron* : il dépend d'un

gNP qui domine également un adjectif *rapide*, ce qui donnera une nouvelle dépendance : «**nom=***neutron* **adj=***rapide*». La mise en dépendance avec *filière* sera écartée car elle a déjà été faite. Enfin vient *rapide* qui est déjà utilisé dans la dépendance précédente et qui n'est pas en relation avec d'autres lexèmes, il est laissé de côté; le parcours du tableau est terminé. Trois dépendances élémentaires ont ainsi été extraites (voir table 5.4).

Une méthode générique La méthode d'obtention de dépendances élémentaires décrite dans [HF96, HBDJ95] est générique et basée sur une simplification progressive de l'arbre. Elle est utilisée par le système *ZELLIG* [HNN96] pour construire des graphes de mots associés d'après leurs contextes syntaxiques. Les dépendances élémentaires sont identifiées lors du processus de simplification. Ces simplifications sont soit des «dé-concaténations» (*chacun des sous-arbres dominés par la racine de l'arbre examiné est considéré à tour de rôle* [HN96]) soit des désadjonctions (*dans l'arbre entier, tous les constituants qui modifient un autre constituant sont éliminés* [HN96]). La simplication prend fin lorsque l'arbre est considéré comme élémentaire. Les arbres dits élémentaires, ceux sur lesquels la simplification doit s'arrêter, sont déclarés à part et fournis au programme comme paramètres. Le point fort de cette méthode est qu'elle permet de construire un historique des opérations de simplification qui ont été nécessaires pour obtenir l'arbre élémentaire, et donc qu'à partir d'un arbre élémentaire, on peut connaître les opérations de dérivations qui sont nécessaires pour obtenir tout ou partie de l'arbre final. La méthode exige toutefois que l'arbre soit représenté en respectant certaines contraintes formelles [HN96] : «*cela suppose de représenter un modifieur comme un (sous-)arbre dont un des fils porte la même étiquette que la racine de l'arbre et qui comporte plusieurs fils*».

Utilité d'un transducteur d'arbre Les récents travaux de B. Habert [HHPB+97] ont conduit à la réalisation d'un transducteur d'arbres. L'outil permet de normaliser les structures des arbres d'analyse produits par différents analyseurs. Il est paramétré avec des méta-règles qui décrivent quelles transformations effectuer dans l'arbre source pour obtenir l'arbre transformé. Les arbres d'*AlethIP-AlethGram* peuvent donc, après reformulation de leur description, être soumis à l'algorithme de déconstruction en dépendances élémentaires de *ZELLIG*. L'écriture des méta-règles nécessaires à la reformulation a du reste déjà été réalisée pour la comparaison des extracteurs *AlethIP* et *Lexter* [HHPB+97].

Les types de Dépendances Syntaxiques Elémentaires définies

La spécificité des arbres d'analyse d'*AlethIP-AlethGram* nous a conduit à distinguer plusieurs types de dépendances élémentaires. Les distinctions faites prennent en compte leurs structures et les éventuelles graphies inconnues pour l'analyseur. Ces catégories sont assignées par le catégoriseur, ainsi : nom inconnu (`xxNom`), adjectif inconnu (`xxAdj`) ou simplement d'inconnu (`xx`). Les distinctions faites entre les types de dépendances visent essentiellement à faciliter par la suite les opérations de comptage

et de filtrage. Nous distinguons les dépendances à deux positions des dépendances à trois positions. Les dépendances à deux positions sont : NOM ADJECTIF, ADJECTIF NOM, NOM_1 NOM_2, NOM_1 XX, XX NOM_2, XX_1 XX_2. Les dépendances à trois positions sont : NOM_1 PRÉPOSITION NOM_2, NOM_1 PRÉPOSITION XX, XX PRÉPOSITION NOM_2 et XX_1 PRÉPOSITION XX_2.

Voici quelques exemples de dépendances élémentaires extraites du corpus ARD-EDF pour les différents types définis. Elles sont présentées sous leur forme texte. La catégorie assignée par le catégoriseur précède une liste de traits entre accollades laquelle précède la forme lexicale (pour la signification des traits se reporter en annexe A). Ainsi la dépendances *fissure représentative* de type NOM ADJECTIF est décrite comme suit :

```
/frNom{Csem=68;Morpho=1} FISSURE /adjStd{Csem=108;Morpho=1} REPRESENTATIF
```

La dépendance *nouveau capteur* de type ADJECTIF NOM est décrite comme ceci :

```
/adjStd{Csem=104;Morpho=3} NOUVEAU /frNom{Csem=26;Morpho=3} CAPTEUR
```

La dépendance *activité réacteur* de type NOM_1 NOM_2 est décrite comme ceci :

```
/frNom{Csem=1;Morpho=1} ACTIVITE /frNom{Csem=26;Morpho=1} REACTEUR
```

Les dépendances de type NOM_1 XX, XX NOM_2, XX_1 XX_2 (une ou plusieurs formes inconnues pour le système) sont par exemples décrites comme ceci :

```
/frNom{Csem=55;Morpho=4} COMITE /xx REP 2000
/xxNom GRANITUREDE /frNom{Morpho=4} PRESSE-ETOUPE
/xxSigle GV /xxSigle PWR
```

On notera que *garniturede presse-étoupe* provient d'un problème typographique (voir section 3.2.2) qui a entraîné une erreur de délimitation du nom *garniture* et de la préposition *de*. La dépendance *étanchéité du joint* de type NOM_1 PRÉPOSITION NOM_2 est décrite comme ceci :

```
/frNom{Csem=2;Morpho=1} ETANCHEITE /frPrep DE /frNom{Csem=25;Morpho=3;D=d} JOINT
```

Enfin, les dépendances du type NOM_1 PREP XX, XX PREP NOM_2 et XX_1 PREP XX_2 peuvent avoir cette allure :

```
/frNom{Csem=29;Moprho=1} CUVE /frPrep DE /xxNomD=N REP
/xxNom FRAGILISATION /frPrep PAR /frNom{Npred=1;Csem=72;Morpho=2;D=0} IRRADIATION
/xxSigle CPP /frPrep DE /xxSigle REP
```

5.3 Les traitements sémantiques : désambiguïsation lexicale

Nous avons choisi une approche fixiste qui définit le sens des mots hors contexte. L'ambiguïté des unités lexicales sera donc représentée par des catégories sémantiques attachées aux noms, aux adjectifs et aux adverbes. Un seul nom pourra renvoyer à des référents multiples. De même, un adjectif sera interprété différemment en fonction du nom qu'il modifie (*longue combustion* vs. *longue digue*). Il faut se donner les moyens de choisir la catégorie sémantique du nom, de l'adjectif. Pour cela, on fait l'hypothèse[2] que le contexte linguistique de la forme à désambiguïser permet de déterminer, dans la plupart des cas, la bonne valeur sémantique. Mais conformément à l'approche fixiste, le contexte ne permet pas de modifier, préciser, calculer un nouveau sens à partir d'un noyau existant. Il permet seulement la sélection d'un sens «statique» identifiable dans certains contextes connus d'avance. Le processus de désambiguïsation sera donc un processus de sélection de sens et non de construction du sens. Nous décrivons ci-après le jeu d'étiquettes sémantiques qui a été projeté sur le lexique.

5.3.1 Description du jeu d'étiquettes utilisé

Nous avons cherché à tirer parti de la couche sémantique du dictionnaire *AlethDic*. Nous avons toutefois simplifié les combinaisons de classes sémantiques et de traits distinctifs utilisées dans *AlethDic*. Le dictionnaire, l'opération de simplification et les catégories sémantiques résultantes sont présentés en annexe A. Seuls les noms et les adverbes ont bénéficié de ce lexique sémantique d'*AlethDic*. Pour les noms, 72 catégories ont été définies. Nous avons défini nous-même les étiquettes sémantiques pour les adjectifs (une cinquantaine au total). Sur le corpus ARD, une trentaine d'étiquettes sont utilisées pour les noms, et une vingtaine pour les adjectifs. La granularité des étiquettes est variable grâce à leur organisation hiérarchique par héritage. La signification précise des étiquettes est décrite en annexe A.

Etiquettes pour les noms

Les étiquettes associées aux noms sont des symboles graphiques qui représentent des types de référents. Ce sont des sortes de catégories conceptuelles. Par exemple, le nom «ordinateur» se voit assigné l'étiquette APPAREIL. Les étiquettes sont organisées en une hiérarchie ontologique : la forme complète de l'étiquette APPAREIL est en fait : ENTITÉ-CONCRET-INANIMÉ-APPAREIL. Ainsi, il n'est pas faux de dire qu'«ordinateur» est INANIMÉ (dans le sens où il ne doit pas son existence matérielle

2. Comme nous l'avons vu au chapitre précédent, il s'agit d'une hypothèse réductionniste. On sait que le seul contexte de formes lexicales n'est pas suffisant pour désambiguïser, qu'il doit être complété du contexte extralinguistique propre aux individus impliqués dans l'acte de communication. Mais à cette réduction du contexte correspond aussi une réduction du nombre de sens possibles. La question est donc : l'imprécision des sens à attribuer lors de la désambiguïsation s'accommode-t-elle de la pauvreté du contexte ?

à des tissus vivants). Les étiquettes ont été définies indépendamment de domaines d'activité. C'est-à-dire que l'on pourra toujours rattacher une notion terminologique d'un domaine donné à cette hiérarchie, à partir des trois types de référents source : entités concrètes (référent matériel), abstraites (référent invisible, immatériel) ou «expérientielles» (référent dynamique : processus, phénomène, ...), ou plus profondément dans la hiérarchie.

Le nombre d'étiquettes retenues (72) est un compromis entre la masse d'ambiguïtés à gérer dans le lexique (celle-ci croissant avec l'augmentation du nombre d'étiquettes, puisqu'on introduit des distinctions sémantiques plus fines et plus nombreuses) et le pouvoir de filtrage de filtres de sélection de candidats termes, basés sur des schémas syntaxico-sémantiques [NHM96b].

Si nous avions pu avoir recours à un équivalent français de la hiérarchie des concepts nominaux de WordNet, nous aurions dû limiter celle-ci à la partie haute, celle des concepts les plus généraux. Car désambiguïser des sens aussi spécifiques que ceux du bas de cette hiérarchie, demanderait un travail considérable d'écriture de règles de désambiguïsation, il est probable que les contextes lexico-syntaxiques autour des noms polysémiques seraient insuffisants pour résoudre la polysémie. Une méthode statistique de désambiguïsation comme celle de [Yar92] paraît plus appropriée[3]. Toutefois, le fait de manipuler une hiérarchie (gigantesque dans le cas de Wordnet : jusu'à 14 niveaux de profondeur) permet de projeter sur le lexique des ensembles de catégories sémantiques de cardinalité variable. Dans un premier temps, nous cherchions à mettre en oeuvre un système de catégories simple afin de déterminer si cela vaut la peine d'utiliser un système de catégories plus fines.

La simplicité et le caractère général des catégories fait gagner un temps certain dans la constitution du lexique et l'écriture des règles de désambiguïsation associées. De plus, nous ne sommes pas sûr que l'usage de catégories sémantiques très précises permette dans le seul contexte phrastique de résoudre les ambiguïtés possibles autour de cette catégorie. En revanche, adopter un ensemble très réduit de 12 étiquettes comme dans [BPV93b, BPV93c][4] simplifierait le processus de désambiguïsation, mais aboutirait certainement à la définition de filtres trop tolérants.

Distinction entre type de référent et notion terminologique

Si les notions terminologiques peuvent être subsumées par des catégories conceptuelles plus générales, elles peuvent être considérées comme des catégories conceptuelles spécialisées, et il faut leur réserver un traitement distinct lors de la désambiguïsation. Mais nous ne voulons pas confondre l'affectation d'une notion terminologique (ce qui revient à faire de l'indexation) et l'affectation d'une catégorie sémantique, comme le type de référent, qui est plus générale que la notion termino-

3. Nous l'aurions mise en pratique si un équivalent du thesaurus Roget avait été disponible pour le français. Le nombre de catégories du thesaurus Roget (environ 1800) est inférieur au nombre de concepts nominaux dans WordNet (plus de 13 000 dans la version 1.5)

4. C'est-à-dire les douze catégories suivantes : ACT, HUMAN_ENTITY, ANIMAL, VEGETABLE, MATERIAL, BUILDING, BY_PRODUCT, ARTIFACT, MACHINE, PLACE, QUALITY, MANNER

logique

Soit par exemple le nom *centrale*, codé monosémique par le dictionnaire, comme un BÂTIMENT PROFESSIONNEL. Sur le plan linguistique, il n'y a pas ambiguïté et cela est utilisable pour la sélection de candidats termes dans un patron syntaxico-sémantique. Mais sur le plan terminologique, c'est ambigu : CENTRALE NUCLEAIRE vs. CENTRALE D'ACHAT, bien que dans les deux cas *centrale* puisse toujours être considéré comme un bâtiment professionnel. Cette ambiguïté relève d'un problème d'indexation automatique.

Il y a ainsi deux sortes de désambiguïsation : la première, linguistique, pour permettre à des patrons syntaxico-sémantiques d'être efficaces. La seconde, qui fait partie du processus d'indexation, et qui consiste à associer une notion terminologique à un groupe nominal ou à une unité lexicale simple (par exemple, si *centrale* apparaît seul, sans modifieur).

Considérons par exemple trois des multiples sens du nom *base* :

Sens : type de référent	Exemple de termes
base1 : ARTEFACT	base de donnée, base de connaissance, base de faits
base2 : SUBSTANCE	base forte, base faible
base3 : LIEU	base navale, base militaire

Si le nom *base* apparaît dans le texte, la désambiguïsation linguistique consistera à lui associer un type de référent parmi ceux connus du dictionnaire. Ce type de référent peut être soit considéré comme un trait linguistique, soit comme une notion générale. Ainsi on ne confond pas désambiguïsation lexicale linguistique et attribution d'une notion terminologique à une forme eliptique. Dans les deux cas c'est le même contexte de signes qui intervient, mais ce n'est certainement pas le même contexte métalinguistique et extralinguistique : dans le cas de la désambiguïsation linguistique, c'est une connaissance métalinguistique de la langue et du lexique qui intervient, dans le second cas, c'est une connaissance de type extralinguistique, encyclopédique, une connaissance du domaine d'activité[5].

Pour ce que nous cherchons à faire, la première désambiguïsation est la seule nécessaire. En effet, nous cherchons à définir des patrons syntaxico-sémantiques suffisamment précis pour qu'ils soient en mesure de décrire la forme linguistique que peuvent prendre des SNP. Il s'agit avant tout de la description d'une forme. Une représentation linguistique non ambiguë permet de manipuler une forme plus précise.

Etiquettes pour les adjectifs

Les étiquettes associées aux adjectifs cherchent à exprimer une plus ou moins grande potentialité de modification du référent nominal. Elles ont été définies empiriquement par l'observation de la relation nom—adjectif dans des termes certifiés et

5. On retrouve aussi dans cette dualité linguistique/conceptuel les problèmes que pose la terminologie (voir introduction), notamment la dualité entre concept général/concept spécialisé

des groupes nominaux non dénominatifs [NHM96a]. Le but est de distinguer les adjectifs qui enrichissent ou modifient le référent nominal pour former une dénomination (échange économique, processus chimique) des adjectifs qui ont un pouvoir descriptif trop faible, et qui ne peuvent altérer de manière significative le référent nominal (récent échange, processus similaire). Certains adjectifs ont plusieurs étiquettes, selon la valeur sémantique du nom qu'ils modifient ; par exemple, *long* quantifie le temps (LOC-TEMPS-ASPECT-DURÉE) avec un nom d'OPÉRATION ou d'ACTIVITÉ, alors qu'il quantifie la taille (QUANT-TAILLE) avec un nom d'artefact.

Nous avons rapidement abandonné cette entreprise de codage étant donné la très forte polysémie des adjectifs. Toutefois nous utilisons ce jeu d'étiquettes dans nos traitement (celui-ci est combiné avec les suffixes d'adjectifs) car un certain nombre d'entre elles désignent des adjectifs dont le sens est rarement ou jamais ambigu. Ainsi *similaire* est toujours un adjectif de comparaison, de même *différent* lorsqu'il est postposé. Nous avons donc conservé le codage dans son état inachevé. Le nombre d'adjectifs codés est d'environ 2000. Seule une dizaine de règles de désambiguïsation pour les adjectifs les plus fréquents (comme *différent*: comparaison, énumération) a été écrite.

Etiquettes pour les adverbes et les prépositions

Le dictionnaire *AlethDic* fournit des valeurs sémantiques pour les adverbes. On trouve : AFFIRMATION, CHRONOLOGIE, FRÉQUENCE, HABITUDE, INTENSITÉ, NÉGATION, QUANTITÉ, TEMPS. Par exemple : hautement : INTENSITÉ, immédiatement : TEMPS, moins : QUANTITÉ, non : NÉGATION, précédemment : CHRONOLOGIE.

Pour le moment, aucune étiquette pour les prépositions n'a été définie. L'idée est de donner une valeur sémantique à la préposition lorsque cela est possible, notamment dans certaines configurations «N1 prep N2» ou les catégories sémantiques de N1 et N2 sont connues. Par exemple dans les configurations «N1 à N2» où «à N2» peut introduire une propriété si N1 est un nom concret et N2 un nom de matière. Si la valeur de la préposition peut être déduite de ses contextes droit et gauche immédiats, il n'est pas nécessaire de coder cette valeur, puisque si le contexte est codé, la valeur de la préposition devient implicite.

5.3.2 Désambiguïsation basée sur le contexte

Le contexte de la forme ambiguë fournit des éléments essentiels pour la précision de son sens. Ainsi : *parc (nucléaire + naturel + floral + à huîtres + à jouer + automobile)*. A un schéma sémantique abstrait et non saturé - ici, *parc* : espace délimité – sont associés de nouveaux attributs ou de nouvelles propriétés (fonction, usage). Cela correspond à l'introduction de nouveaux éléments linguistiques (adjectifs, syntagme prépositionnel, noms) dans le contexte de la forme nominale. La représentation initiale se trouve réexprimée, précisée, modifiée. Avec l'appoche fixiste, cette vision générative d'un noyau de sens modifié dynamiquement, est remplacée par ce type de désambiguïsation : on ne connaît pas le sens du nom *parc*. Mais une règle énonce

que si *parc* est modifié par l'adjectif *naturel* alors il désigne un ESPACE GÉOGRAPHIQUE. Une autre règle énonce que s'il est modifié par l'adjectif *nucléaire*, il désigne un BÂTIMENT INDUSTRIEL, etc.

Dépendance des contextes-solutions vis-à-vis du corpus de mise au point

S'appuyer sur le contexte lexico-syntaxique de la forme pour identifier son acception implique inévitablement une étape de recensement de ces contextes-solutions. Ceux-ci, une fois recensés et stockés, sont utilisés ultérieurement pour résoudre des formes ambiguës. Mais ils sont dépendants du texte à partir duquel ils ont été mis au point. Pour désambiguïser un nouveau texte, il faut donc vérifier que les contextes-solutions sont toujours efficaces et le cas échéant, déterminer de nouveaux contextes-solutions. C'est dans ce but que nous avons développé un environnement d'écritures de règles de désambiguïsations, accélérant l'identification des contextes-solutions (s'ils existent) (voir annexe B).

Le cas des formes elliptiques

Introduit dans le texte pour la première fois dans sa forme complète, un syntagme nominal peut être ensuite réintroduit sous une forme elliptique, simplifiée ou ayant subi une légère variation. Lorsque l'on tombe sur un nom polysémique et que son contexte immédiat ne fournit pas d'indication fiable pour statuer sur le type de référent du nom, on fait l'hypothèse qu'il s'agit d'une référence anaphorique (ou cataphorique) elliptique. Dans ce cas, il faut retrouver la forme à laquelle l'ellipse fait référence, ce qui est difficilement automatisable. Nous avons implémenté les routines pour rechercher l'occurrence de la dernière ou de la prochaine forme non elliptique. Toutefois nous n'avons pas encore vérifié si une telle recherche donne des résultats cohérents.

5.3.3 Les règles de désambiguïsation

La vocation première des règles est la reconnaissance de formes linguistiques. Ceci est nécessaire lors de l'identification des contextes des lexèmes ambigus à traiter. Une fois la forme et son contexte repérés, des actions peuvent être envisagées. La but de la démarche est de gagner du temps par rapport à une désambiguïsation manuelle, par l'application automatique de règles, mais aussi de disposer de principes de désambiguïsation applicables sur différents corpus, par le stockage et la réutilisation de ces règles.

Principe de fonctionnement

Les règles de désambiguïsation s'appuient sur un traitement symbolique de l'information. Chaque règle est définie par un identifieur qui correspond à la forme graphique du lexème à désambiguïser. On peut aussi préciser la catégorie lexicale de la forme. Lorsque le programme lit le texte, il déclenche les règles lexicales présentes

dans le dictionnaire de désambiguïsation. Ainsi, la règle ayant pour identifieur *sortie* se déclenchera pour toute les formes *sortie* dans le texte, toutes catégories lexicales confondues; alors que la règle dont l'identifieur est *sortie*.NOM ne se déclenchera que pour les formes *sortie* qui ont été catégorisées comme des noms. Les règles doivent être déclarées selon une syntaxe particulière. Elles sont ensuite compilées pour être exécutées par le moteur de désambiguïsation.

La syntaxe des règles et des exemples de règles sont visibles en annexe B.

Reconnaissance de formes linguistiques

Les formes linguistiques sont reconnues par un système de mise en correspondance (*pattern matching* d'un schéma descriptif avec du texte. Le texte est alors représenté sous une forme séquentielle linéaire, chaque unité lexicale étant enrichie des informations linguistiques que nous avons définies au début de ce chapitre. Un schéma descriptif est une expression régulière sur des mots. L'unité de base n'est pas le caractère mais la forme lexicale. Les opérateurs de Kleene décrivent la présence optionnelle ou obligatoire des formes lexicales :

!	Une et une seule unité lexicale
?	Zéro ou une unité lexicale
$\star$	Zéro ou n unité(s) lexicale(s)

L'opérateur «Un ou n» n'a pas été implémenté étant donné son peu d'utilité pour décrire des formes linguistiques est équivalent à un ! suivit d'un $\star$. L'algorithme de mise en correspondance, ne traite pas l'ambiguïté. Si plusieurs mises en correspondance sont possibles dans le même énoncé, seule la première solution est prise en compte. Les autres ne sont pas calculées. La table 5.7 résume le principe de la mise en correspondance. La première solution adoptée est celle qui suit les contraintes des étapes 1 et 2. Les éléments obligatoires sont identifiés dans la phrase en progressant de la gauche vers la droite. Cette approche qui ne prend que la première solution minimise le temps de calcul.

Description des unités Les unités linguistiques associées aux opérateurs obligatoires de Kleene doivent être décrites par des propriétés linguistiques. Les éléments optionnels peuvent rester non contraints. Pour des questions de lisibilité, l'attachement des contraintes linguistiques à un opérateur se fait par l'intermédiaire d'un

TAB. 5.7 – *Principe de la mise en correspondance de formes linguistiques sans traitement de l'ambiguïté*

En entrée : une séquence SEQ, un schéma descriptif DES.

Etape 1 Prendre connaissance dans DES des éléments optionnels qui jouxtent les éléments obligatoires.

Etape 2 Identifier les éléments obligatoires dans SEQ en respectant les éventuels éléments optionnels qui les jouxtent, **de la gauche vers la droite**.

Etape 3 Identifier les éléments optionnels dans SEQ

En sortie (si succès) : le schéma DES dont les opérateurs sont associés à des segments de la séquence SEQ.

registre. En voici un exemple :

```
⋆   !∼1   ?∼2   !∼3   !∼4   ?∼5   ⋆
cat(∼1) == nom
cat(∼2) == adverbe
cat(∼3) == adjectif
cat(∼4) == preposition
cat(∼5) == article

...
suffixe(∼1) in [-ation, -ateur]
...
```

Dans cet exemple, on cherche à identifier dans une séquence linguistique un syntagme qui commence par un nom, suivi éventuellement d'un adverbe, immédiatement suivi d'un adjectif puis d'une préposition, cette dernière étant éventuellement suivie d'un article. Une dernière contrainte spécifie que le premier nom doit se terminer en *-ateur* ou *-ation*.

Limitations - Evolution possible

La principale limitation de telles descriptions est qu'elles doivent être ordonnées de la plus spécifique à la plus générale avant d'être appliquées. Lorsque le nombre de règles est important cela pose des problèmes de maintenance. Nous y avons été peu confronté étant donné qu'à une entrée lexicale, nous n'avons jamais associé plus de cinq ou six descriptions. Mais ne serait-ce que pour une question de performance, il serait souhaitable de convertir de telles descriptions en automates à états finis. L. Karttunnen [Kar91, KCGS97] montre que des expressions régulières comme celles que nous utilisons peuvent être représentées sous la forme d'automates à états finis. Le système *INTEX* utilise également des automates à états finis pour faire de la reconnaissance morpho-syntaxique [Sil93].

5.3.4 Désambiguïsation du corpus EDF-ARD

L'objectif initial était d'assigner à chaque nom, adjectif et adverbe une étiquette sémantique. Etant donnée l'ampleur de la tâche, nous avons revu cet objectif à la baisse. Nous avons limité l'étiquetage sémantique des adjectifs au strict minimum [6]. Nous avons également limité l'écriture de règles de désambiguïsation aux formes ambiguës les plus fréquentes.

Nous sommes partis de l'existant trouvé dans le dictionnaire *AlethDic* [7] :

2163 noms identifiés comme monosémiques dans AlethDic
1079 noms reconnus polysémiques dans AlethDic
2435 noms inconnus du dictionnaire AlethDic

Soit 3514 noms pour lesquels il faut écrire une règle d'affectation ou de désambiguïsation sémantique (sur un total de 5677).

649 adjectifs sont renseignés dans AlethDic (adjectifs géographiques et de couleur)
932 adjectifs sont sans valeur sémantique

Soit 932 adjectifs à renseigner sur un total de 1581 adjectifs.

L'écriture des règles de désambiguïsation a été faite avec l'outil développé à cet effet (voir en annexe B.2). La relative unité thématique du corpus a permis de minimiser le travail de désambiguïsation lexicale (en réduisant les polysèmes et les homonymes à prendre en compte).

Résultats de la catégorisation sémantique des noms

Au total, 725 règles de catégorisation sémantique de noms ont été écrites, soit un total de 905 contextes-solutions définis. C'est-à-dire que pour un nom, 1,25 règles ont été écrites en moyenne. Seul le sous-corpus *ard95* (textes d'ARD de l'année 1995) a été utilisé pour définir ces règles de désambiguïsation. On a considéré que ce corpus était représentatif en terme de couverture lexicale et de diversité de contextes des formes ambiguës, notamment parce qu'il offre le lexique nominal le plus large (voir figure 3.4) et parce qu'il est le plus important de par sa taille. Ensuite, ces mêmes règles construites sur *ard95* ont été appliquées à tout le reste du corpus. La table 5.8 montre les résultats de catégorisation en terme de noms étiquetés/non étiquetés.

Lecture des résultats La table 5.8 présente les résultats de catégorisation par sous-corpus. Dans la première colonne est indiqué le sous-corpus et sa taille en nombre de mots. La seconde colonne fait état du nombre de mots catégorisés comme

6. Etant donné que la modélisation de valeurs sémantiques pour les adjectifs est une tâche complexe (ils sont très polysémiques), nous n'affecterons de catégories sémantiques qu'aux adjectifs qui discréditent le plus souvent la pertinence des syntagmes

7. Pour les classes sémantiques dans AlethDic et de leur mise en correspondance avec notre système d'étiquette, voir annexe A

TAB. 5.8 – *Résultats d'étiquetage du corpus EDF-ARD par sous-corpus. Les règles de désambiguïsation ont été définies à partir d'un échantillon constitué des noms les plus fréquents dans le sous-corpus* ard95

corpus/mots[a]	*occ-N*[b]	*occ-Nc*[c]	*occ-Nnc*[d]	*Nc*[e]	*Nnc*[f]	(*connus*[g]	*inconnus*[h])
ard84/29430	7895	6812	1083	950	547	189	358
	26.8%	86.3%	13.7%	63.46%	36.53%	34.5%	65.44%
ard85/31483	8504	7381	1123	999	522	207	315
	27%	86.8%	13.2%	65.68%	34.32%	39.6%	60.3
ard86/30554	8172	7023	1149	1006	598	214	384
	26.7%	86%	14%	62.7%	37.3%	35.7%	64.3%
ard87/32968	8841	7700	1141	1117	610	201	409
	26.8%	87%	13%	64.7%	35.3%	32.9%	67.1%
ard88/42599	11580	10207	1373	1118	599	244	355
	27.1%	88.14%	11.86%	65.1%	34.9%	40.7%	59.3%
ard89/45996	12020	10604	1416	1104	587	233	354
	26.1%	88.2%	11.78%	65.2%	34.8%	39.7%	60.3%
ard90/73231	19723	17131	2592	1364	1160	335	825
	26.9%	86.8%	13.2%	54%	46%	28.8%	71.2%
ard91/96042	25857	22455	3402	1510	1257	406	851
	26.9%	86.8%	13.2%	54.5%	45.5%	32.2%	67.8%
ard92/109879	28798	25253	3545	1612	1416	429	987
	26.2%	87.7%	12.3%	53.2%	46.8%	30.2%	69.8%
ard93/138468	36049	25253	4438	1742	1788	472	1316
	26%	87.6%	12.4%	49.3%	50.7%	26.3%	73.7%
ard94/137789	35907	31411	4496	1766	1916	430	1486
	26%	87.4%	12.6%	47.9%	52.1%	22.4%	77.6%
ard95/230174	60022	52464	7558	2116	3158	492	2666
	26%	**87.4%**	**12.6%**	**40.1%**	**59.9%**	**15.5%**	**84.5%**
Moyenne	26.3%	*87.2%*	*12.8%*	*57.2%*	*42.8%*	*32.4%*	*67.6%*

[a] Identifie le sous-corpus et le nombre de mots qu'il contient.

[b] Nombre de formes catégorisées comme des noms dans le sous-corpus.

[c] Nombre d'occurrences de noms qui ont été catégorisés sémantiquement.

[d] Nombre d'occurrences de noms qui n'ont pas été catégorisés sémantiquement.

[e] Nombre de noms différents qui ont été catégorisés sémantiquement.

[f] Nombre de noms différents qui n'ont pas été catégorisés sémantiquement.

[g] Proportion de noms connus du dictionnaire qui n'ont pas été catégorisés sémantiquement.

[h] Proportion de noms inconnus du dictionnaire qui n'ont pas été catégorisés sémantiquement. Il s'agit de fomes graphiques inconnues auxquelles le lemmatiseur/catégoriseur a affecté une étiquette de nom hypothétique.

des noms dans chaque sous-corpus. Sous le chiffre du nombre d'occurrences est indiquée la proportion des noms. Sur tous les sous-corpus, la proportion de noms oscille entre 26 et 27%. En troisième et quatrième colonnes sont indiqués les nombres d'occurrences des noms étiquetés et des noms qui n'ont pas été étiquetés. Sous chaque effectif est indiquée la proportion correspondante. Pour interpréter ces chiffres, il faut rappeler que l'écriture de règles de désambiguïsation a été faite à partir du sous-corpus *ard95* sur les formes ambiguës les plus fréquentes de ce sous-corpus. Le taux de noms catégorisés oscille ainsi entre 86 et 87%, pour une moyenne de 87.2% sur tout le corpus. C'est un résultat encourageant : si nous avions écrit suffisamment de règles de désambiguïsation sur *ard95* pour augmenter ce taux de catégorisation, nous aurions certainement augmenté alors le taux de catégorisation sur le reste du corpus. Les cinquième et sixième colonnes donnent un éclairage différent et relativisent fortement les chiffres des deux colonnes précédentes. Elles indiquent respectivement la taille et proportion du lexique des noms catégorisés et du lexique des noms non catégorisés dans chaque sous-corpus. Pour le sous-corpus *ard95*, ces chiffres sont respectivement de 40% de noms différents catégorisés et 60% de noms différents non catégorisés. Ces 60% du lexique nominal non catégorisé se répartissent en 15.5% de noms connus du dictionnaire AlethDic (colonne 7) et 84.5% de noms inconnus du dictionnaire (dernière colonne). Ces 15% de noms connus du dictionnaire mais qui n'ont pas été catégorisés correspondent aux noms les moins fréquents dans le corpus, pour lesquels n'avait pas été écrit de règle de désambiguïsation. Quant aux 84.5% de noms inconnus, ils correspondent à des formes inconnues du dictionnaire qui ont été hypothétiquement catégorisées comme des noms par le lemmatiseur-catégoriseur d'AlethIP. Les taux d'étiquetage sémantique du lexique nominal connu s'améliore progressivement pour les sous-corpus précédents, de *ard94* à *ard84*. On passe ainsi de 47.9% à 63.4%. La raison principale est que la taille de ces sous-corpus décroît progressivement de 1994 à 1984, ce qui diminue la proportion de noms inconnus, et assure une meilleure couverture du lexique de noms pour lesquels il existe des règles de désambiguïsation sur le lexique du sous-corpus.

Evaluation difficile sans corpus de référence Etant donné le temps qui nous était imparti, nous ne présentons pas d'évaluation de la qualité des étiquettes sémantiques attribuées aux noms. Une évaluation manuelle serait nécessaire pour cela. Il faudrait vérifier pour chaque nom si sa catégorisation est correcte, et le cas échéant inspecter quelle règle d'affectation/désambiguïsation a été appliquée. Nous sommes cependant en mesure de donner une appréciation qualitative des résultats. Il est raisonnable d'affirmer que les étiquettes attribuées aux noms du sous-corpus *ard95* sont quasiment toutes correctes, puisque les règles de désambiguïsation qui ont décidé de l'affectation ont été pistées manuellement et leurs résultats contrôlés, grâce à une interface dédiée. La lecture des sorties catégorisées sur lesquelles nous avons travaillé montre que les catégories affectées sont satifaisantes. Les erreurs identifiées pour l'ensemble du corpus ont deux causes. La première est que la règle de désambiguïsation a été confrontée à un contexte qu'elle n'a pas su traiter. Dans ce cas d'indécision,

une catégorie sémantique par défaut est tout de même affectée à la forme lexicale. La seconde cause est à attribuer aux insuffisances du dictionnaire *AlethDic* dans sa version 1.5.5 [8]. Précisément, il est un certain nombre de formes que le dictionnaire déclare monosémiques (une seule catégorie sémantique associée à la forme) alors qu'en réalité elles peuvent être polysémiques. Nous n'avons pas remis en question le caractère monosémique des formes déclarées par AlethDic.

5.4 Conclusion

Dans ce chapitre, nous avons présenté les traitements d'enrichissement linguistique que nous faisions subir au texte analysé par *AlethIP*. Des informations morphologiques, syntaxiques et sémantiques viennent augmenter la représentation du texte. L'extraction des dépendances élémentaires des groupes nominaux permet de construire un certain point de vue sur les syntagmes nominaux qui nous permet de les filtrer, comme nous allons l'expliquer dans le chapitre suivant.

Enfin, nous avons présenté notre système de désambiguïsation qui permet d'assigner des catégories aux unités lexicales. Nous sommes conscient que la méthode de désambiguïsation choisie souffre d'un certain nombre de limitations. Le coût pour définir les règles est important. Cela a nécessité en premier lieu le développement d'un environnement *ad hoc* (voir annexe B). Et il a fallu ensuite mettre au point les règles (nous avons passé plus d'une centaine d'heures sur ce point). Chaque règle est certes réutilisable, mais à la condition de traiter des corpus abordant les mêmes domaines d'activité que le corpus à partir duquel elles ont été élaborées (et dans notre cas, ils étaient nombreux : voir paragraphe 3.2.2).

Tout en conservant une approche sémantique fixiste, l'évolution souhaitable du système de désambiguïsation serait d'utiliser une technique d'apprentissage. Nous aurions pu faire appel à une méthode statistique ou symbolique à correction d'erreur mais ce qui est sûr, c'est que dans ces deux derniers cas, nous n'aurions pas pu faire l'économie d'un codage manuel des catégories sémantiques, pour définir un corpus de référence. Les outils qui ont été développés sont adaptés à la mise au point d'un tel corpus [9].

8. Rapelons que ce dictionnaire évolue et que nous l'avons exploité dans un certain état de développement.

9. Si nous avions travaillé sur l'anglais, nous aurions pu bénéficier de l'ontologie existante de *WordNet* pour la définition des catégories sémantiques, et des travaux de désambiguïsation réalisés sur cette base lexicale publique.

Chapitre 6

Filtrer des syntagmes nominaux

Nous présentons dans ce chapitre la méthode utilisée pour filtrer des groupes nominaux. Etant donné que les filtres sont les résultats d'une procédure d'apprentissage, ce chapitre et le suivant (Apprentissage de filtres) sont fortement dépendants.

6.1 Un filtrage basé sur les dépendances syntaxiques élémentaires

Contrairement à des systèmes qui misent tout sur des règles purement syntaxiques (que ce soit par l'usage de patrons syntaxiques d'extraction ou le repérage des syntagmes à l'aide de frontières ou de marqueurs) et qui sont opérationnels immédiatement sur les corpus (*AlethIP* (Erli), *Lexter* [Bou94b]) nous avons choisi de prendre appui sur les compatibilités lexicales pour exprimer la pertinence des syntagmes. Cela implique que nous ne pouvons pas filtrer des noms simples, mais seulement des syntagmes polylexicaux.

Le principe du filtrage est d'appréhender les syntagmes non pas comme des arbres d'analyses complets dont la structure est parfois complexe, mais comme des ensembles de dépendances lexico-syntaxiques élémentaires. Le filtrage s'en trouve simplifié [1]. Le procesus de filtrage se décompose en 2 phases : la première est l'application de schémas ou règles de filtrage construits grâce à la procédure d'apprentissage [2] aux dépendances élémentaires du groupe nominal à filtrer. Cela est présenté plus bas en 6.2.1. La seconde est la reconstitution du syntagme originel en prenant en compte les conséquences du filtrage. Cette phase est décrite en 6.2.2.

1. Cette simplification montre toutefois des limites. Voir 6.1.2
2. Celle-ci est présentée au chapitre 7

FIG. 6.1 – *Analyse faite par le système* Lexter

```
amelioration de la connaissance des phenomenes de la zone d'assechement des tubes de GV
        T : amelioration
        E : connaissance des phenomenes de la zone d'assechement des tubes de GV
                T : connaissance
                E : phenomenes de la zone d'assechement des tubes de GV
                        T : phenomenes
                        E : zone d'assechement des tubes de GV
                                T : zone d'assechement
                                        T : zone
                                        E : assechement
                                E : tubes de GV
                                        T : tubes
                                        E : GV
sodium
```

6.1.1 Caractéristiques du filtrage

Le filtrage ne s'appuie pas sur l'axe syntagmatique

Le filtrage des groupes nominaux que nous proposons ne s'appuie pas sur l'axe syntagmatique, c'est-à-dire qu'il ne cherche pas à reconnaître dans des syntagmes des formes *a priori* intéressantes, comme des patrons syntaxiques. Il cherche à tirer parti de la structure syntaxique du syntagme exprimé en termes de dépendances lexicales.

Considérons le syntagme suivant extrait du corpus ARD : *amélioration de la connaissance des phénomènes de la zone d'assèchement des tubes de GV chauffés au sodium*. L'extracteur de groupes nominaux de l'application *AlethIP* (Erli) recherche les groupes nominaux pertinents par extraction de sous-syntagmes à l'aide de patrons syntaxiques. Par exemple si l'unique patron «NOM PRÉPOSITION NOM» est appliqué sur notre exemple, le sous-syntagme *tubes de GV* sera extrait. Pour le système *Lexter* [Bou94b], les choses se présentent différemment : les sous-syntagmes retenus sont ceux qui ont été isolés entre des frontières. L'identification des frontières, qui s'appuie entre autres sur la distinction entre certains types de prépositions, repose sur une heuristique dont le but est le repérage d'un certain modèle de groupes nominaux : les candidats termes. Il s'agit donc de proposer des candidats termes plutôt que de filtrer des syntagmes. La seule marque de filtrage est l'effacement de frontières : *Lexter* fait des coupures au sein du syntagme. Les sous-syntagmes retenus entre les frontières sont ensuite décomposés selon une analyse en terme de tête (T) – expansion (E). L'analyse de notre exemple est présentée en figure 6.1; *chauffés à (le)* (participe passé + préposition) a été considéré comme une frontière. Il reste donc deux syntagmes. Le premier est *amelioration de la connaissance des phenomenes de la zone d'assechement des tubes de GV*. Le second est le mot simple *sodium*.

La solution que nous avons retenue cherche à contourner le problème de la décomposition du syntagme en sous-syntagmes en ne s'appuyant ni sur des frontières ni sur des patrons d'extraction. Pour filtrer le syntagme cité en exemple, on s'appuiera sur les dépendances entre les constituants : On cherchera à évaluer en fonction de critères

stockés dans un profil si les dépendances suivantes sont pertinentes : `amélioration de (la) connaissance, connaissance des phénomènes, phénomènes de (la) zone, zone d'assèchement, assèchement de (les) tubes, tubes de GV, tubes chauffés, tubes à (le) sodium`[3]. Les résultats du filtrage pour cet exemple sont montrés plus loin en 6.2.2.

Ainsi aucun patron syntaxique n'est bon ou mauvais a priori. La longueur et la structure du syntagme ne sont absolument pas contraints. En revanche, le matériel lexical entrant dans la composition des syntagmes joue le rôle contraignant. Cette contrainte est exprimée par la relation régisseur (prédicat ou nom recteur) – régis (argument ou modifieur).

Difficulté de filtrer des arbres d'analyses complets

A l'heure actuelle, les analyseurs syntaxiques et en particulier les analyseurs robustes – ceux qui traitent de grandes quantités de texte tout-venant – produisent des analyses dans lesquelles subsistent de fréquentes erreurs d'attachements prépositionnels et adjectivaux. C'est le cas d'*AlethIP* ou *Lexter* (Un travail de comparaison de ces deux extracteurs [HHPB+97] montre que de telles erreurs sont moins fréquentes avec *Lexter*. *Lexter* les minimise en effet par un apprentissage endogène des attachements prépositionnels[4] [Bou94b]).

Etant donné donc l'incertitude quant à la consistance et la régularité des structures construites par ces analyseurs à grammaires robustes, deux groupes nominaux identiques peuvent avoir des analyses distinctes dans des contextes différents. L'usage de filtres syntaxico-sémantiques – comme celui présenté en figure 4.2. – perd alors de son intérêt. C'est pourquoi nous avons préféré filtrer sur les dépendances trouvées dans l'arbre syntaxique. Les dépendances se présentent comme des relations directes entre les unités linguistiques. Cette représentation permet de ne pas s'encombrer de structures d'arbre dont la profondeur est variable et qui de ce fait introduisent des aléas dans la structure des filtres lorsque les arbres (ou les forêts d'arbres) sont produites par une grammaire robuste.

Notons par ailleurs que le filtrage sur les dépendances permet dans certains cas de ne pas être trop pénalisé par les erreurs d'attachements de modifieurs ou d'arguments, notamment lorsque des dépendances à trois positions sont supprimées dans l'arbre originel (voir 6.2). Mais il est indéniable que les dépendances étant extraites à partir d'une analyse en constituants, leur qualité est tributaire de cette analyse faite par *AlethIP*.

Vers un système de filtrage indépendant de l'analyseur syntaxique

Un filtrage sur les dépendances syntaxiques permet d'envisager une certaine indépendance vis à vis de l'analyseur et du formalisme d'expression des résultats

3. Cette dernière dépendance est incorrecte en raison d'une erreur d'analyse syntaxique qui fait dépendre *sodium* de *tubes*. Voir figure 6.7.

4. Cela consiste à trancher une ambiguïté d'attachement prépositionnel, après avoir vérifié s'il existe dans le corpus d'autres relations opérateur–argument identiques, auquel cas, l'attachement est réalisé d'après l'information fournie par le corpus.

syntaxiques. En effet, bien que tous les analyseurs ne produisent pas des sorties conformes à un formalisme de dépendances, il est possible d'en extraire les dépendances par un parcours de l'arbre construit. Par exemple cela est possible avec des analyseurs comme *AlethIP*, *Lexter*. Des analyseurs comme *Sylex* [Con91], *SEXTANT* [Gre94] pour l'anglais fournissent directement leurs résultats sous la forme de dépendances. Avec un transducteur d'arbres comme *FRT* [HHPB+97], les sorties des analyseurs qui produisent des arbres peuvent être normalisées pour être exportées vers un module de filtrage.

Apprendre automatiquement des filtres

Un dernier avantage que nous voyons au filtrage de dépendances syntaxiques est la possibilité de construire automatiquement des filtres par apprentissage. En effet, ces objets étant élémentaires, il peuvent être facilement manipulés et décrits, contrairement à des structures arborescentes complexes qui sont beaucoup plus lourdes à gérer.

6.1.2 Les limitations d'une évaluation isolée des dépendances élémentaires

Lors de la constitution des échantillons d'apprentissage, il est apparu que certaines dépendances pouvaient être pertinentes dans certains contextes et non pertinentes dans d'autres. Par exemple pour le profil I la dépendance `longue durée` avait toujours été choisie non pertinente sauf dans le syntagme : *essai de longue durée*. Dans certains cas peu fréquents, la prise en compte du contexte (arborescent) des dépendances lexicales est nécessaire pour faire des distinctions plus fine comme distinguer une dépendance identique trouvée dans deux contextes arborescents différents. Un formalisme de description d'arbre est nécessaire pour décrire systématiquement les contraintes à l'intérieur et aux environs de la dépendance.

Le formalisme des quasi-arbres employé dans [FH96] et qui s'appuie sur les travaux de [MHF83] et [VS92] exploite la *D-theory*. La *D-theory* définit un formalisme logique qui manipule des descriptions d'arbres plutôt que des arbres. Les arbres sont décrits à l'aide d'opérateurs de dominance et de précédence. Ces opérateurs prennent en argument les noeuds de l'arbre représentés par une constante arbitraire. Le formalisme des quasi-arbres introduit de la souplesse dans les descriptions en distinguant opérateur de dominance et opérateur de dominance stricte (père immédiat dans l'arbre). L'avantage d'un tel formalisme est double. D'une part il permet de décrire des contraintes autour d'une «zone» spécifique (par exemple une dépendance) dans une structure syntaxique globale. D'autre part la forme des descriptions (énumération de contraintes positionnelles sous la forme de prédicats à deux places) autorise une représentation normalisée des contraintes syntaxiques, éventuellement augmentée d'associations *trait–valeur*. Celle-ci peut alors être facilement exploitée par des méthodes statistiques (classification automatique, réseaux de neurones) sous la forme d'un vecteur, ou pour faire de l'apprentissage automatique.

Si nous décidions d'étendre la description des dépendances à leur contexte, il nous faudrait adopter le formalisme des quasi-arbres. Il en résulterait cependant une plus grande dépendance vis à vis de l'analyse globale de l'arbre. Si l'analyseur utilisé est robuste, cela risque d'introduire beaucoup de bruit dans les descriptions[5]. Cela ne nous paraît pas intéressant pour le moment, étant donné les résultats de l'analyseur utilisé qui sont bien trop bruités. Par contre cela est envisageable sur un corpus dont l'analyse syntaxique aurait été corrigée à la main.

6.2 La méthode de filtrage

6.2.1 Filtrage des dépendances syntaxiques élémentaires

Il ne s'agit pas seulement de retenir ou de conserver les dépendances. Il faut que les dépendances puissent être retenues tout en étant modifiées. Des coupures et effacements dans les dépendances sont donc possibles. Ces actions se répercutent ensuite dans l'arbre syntaxique originel duquel les dépendances ont été extraites. Voici les entrées/sorties du processus de filtrage :

Entrées	**Sortie**
Syntagmes nominaux à filtrer Un profil de filtrage	Syntagmes nominaux retenus

Les syntagmes nominaux à filtrer proviennent de la procédure d'enrichissement linguistique (voir paragraphe 3.1.2 et chapitre 5). Le profil provient de la procédure d'apprentissage des filtres (voir chapitre 7). Il contient, sous une forme équivalente à un arbre de décision, des combinaisons d'attributs linguistiques associées à des actions de filtrage.

Le filtrage considère chaque dépendance lexico-syntaxique du groupe nominal à filtrer. Chaque dépendance est traduite sous la forme d'une combinaison d'attributs linguistiques (ceci est expliqué au chapitre suivant). Le filtrage s'effectue en comparant ces combinaisons d'attributs à celles qui sont emmagasinées dans le profil de filtrage. La comparaison se fait à partir des combinaisons les moins permissives, (c'est-à-dire les descriptions linguistiques les plus déterminées) vers les combinaisons les plus permissives (c'est-à-dire des descriptions linguistiques sous-déterminées).

Exemple Considérons les dépendances `tarif excessif` (extraite du syntagme *effet négatif des tarifs excessifs*) et `tarif bleu` (extraite du syntagme *négociation sur la base du tarif bleu*). Considérons également un profil minimal dans lequel figurent les combinaisons de propriétés linguistiques suivantes :

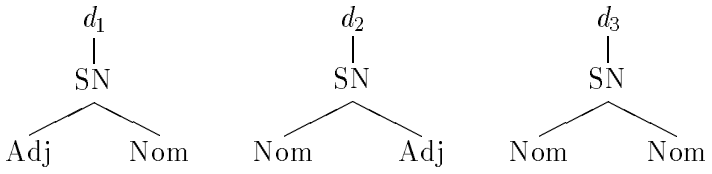
5. Cela entrainerait aussi une complexification de la construction des échantillons d'apprentissage, en cas d'élaboration manuelle de l'échantillon. Chaque dépendance devant être validée en fonction de son contexte. Voir C.1.3

`NOM{graphie=tarif} ADJ{graphie=raisonnable}`	Cette combinaison décrit une dépendance de type NOM ADJ. Le nom doit être *tarif* et l'adjectif *raisonnable*
`NOM{graphie=tarif} ADJ{Csem=127}`	Même type de dépendance que précedemment, mais l'adjectif est moins contraint. Il doit avoir une valeur de qualification appréciative (`Csem=127`)
`NOM{Csem=4} tarif ADJ{Csem=127}`	Ici le nom est doit être un nom d'attribut mesurable (tarif, température, voltage). L'adjectif doit avoir une valeur de qualification appréciative.
`NOM ADJ{Csem=127}`	Pas de contrainte sur le nom. L'adjectif doit avoir une valeur de qualification appréciative.

S'il y a identité entre une description de combinaisons du profil et la description de la dépendance à filtrer, l'action de filtrage associée à la description du profil est appliquée. Ainsi `tarif excessif` ne sera pas retenu par la première description. En revanche la description `NOM{graphie=tarif} ADJ{Csem=127}` permettra de le retenir, avant les deux dernières, les plus générales. Quant à `tarif bleu`, aucune des descriptions ne lui correspond. Les actions associées aux descriptions peuvent entraîner une suppression partielle ou totale de la dépendance : suppression du modifieur, coupure du syntagme, ou simple effacement d'adjectif par exemple. Les actions d'élagage des dépendances et par conséquent des arbres dans lesquelles elles prennent place sont définies par défaut en fonction du type de la dépendance. Toutefois elle peuvent être affinées et contrôlées depuis l'interface de mise au point des échantillons d'apprentissage (voir annexe C). Nous présentons maintenant les différents types d'actions d'élagage.

Cas des dépendances à deux éléments

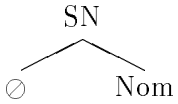
Chaque action est définie pour un type de dépendance particulier. Considérons ainsi, pour les dépendances à deux éléments, les trois principaux types de configurations possibles. Celles-ci sont représentées selon le formalisme d'une grammaire de constituants afin de mieux saisir les conséquences de leur simplification ou suppression dans l'arbre originel :



Elles correspondent aux schémas du type ADJ NOM, NOM ADJ et NOM NOM. Ces schémas sont très généraux, en réalité, l'adjectif peut être un participe passé ou un participe présent, ou un adjectif inconnu. Le nom peut être un sigle, ou un nom inconnu. Soit $simp_2$ la procédure de simplification/filtrage des dépendances à deux positions. Elle prend en paramètre le type de dépendance et l'action à appliquer.

L'action A1 efface le premier membre de la dépendance. Pour le type d_1 cela efface l'adjectif antéposé :

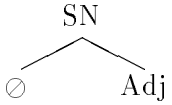
$simp_2(A1,d_1)$ ⇒



nouveau capteur ⇒ capteur

Pour le type d_2, l'action A2 efface le nom tête, il en résulte la suppression de la dépendance (équivalent à $simp_2(A3,d_2)$) :

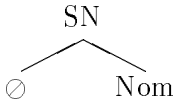
$simp_2(A1,d_2)$ ⇒



problème actuel ⇒ ∅

Pour le type d_3, l'action A1 efface le premier nom du composé binominal :

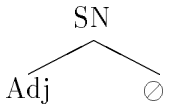
$simp_2(A1,d_3)$ ⇒



aspect sûreté ⇒ sûreté

L'action A2 efface le deuxième membre de la dépendance. Pour le type d_1, cela supprime toute la dépendance (équivalent à $simp_2(A3,d_1)$) :

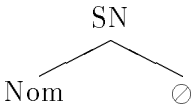
$simp_2(A2,d_1)$ ⇒



nouveau capteur ⇒ ∅

Pour le type d_2, l'action A2 supprime le modifieur adjectival :

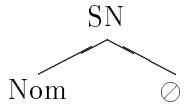
$simp_2(A2,d_2)$ ⇒



vanne actuelle ⇒ vanne

Pour le type d_3, l'action A2 efface le deuxième nom du composé binominal :

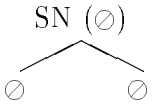
$simp_2(A2,d_3)$ $\Rightarrow$



aspect sûreté $\Rightarrow$ aspect

Enfin l'action A3 appliquée aux dépendances à deux positions supprime toute la dépendance :

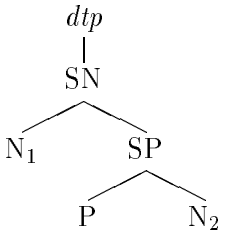
$simp_2(3,d_{\{1,2,3\}})$ $\Rightarrow$



objectif particulier $\Rightarrow$ $\varnothing$

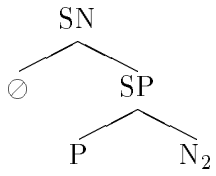
Cas des dépendances à trois éléments

Soit $simp_3$ la procédure de simplification/filtrage des dépendances à trois positions (*dtp*) qui peuvent être représentées de la manière suivante :



L'action A1 efface le premier membre de la dépendance. L'arbre syntaxique d'origine est coupé en deux :

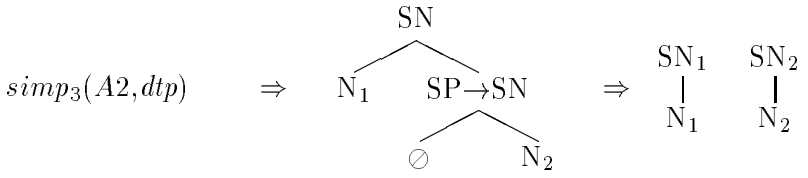
$simp_3(A1,dtp)$ $\Rightarrow$



$\Rightarrow$

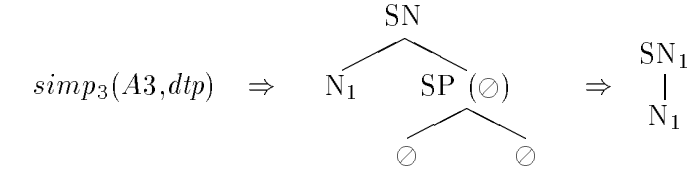
mécanisme d'endommagement $\Rightarrow$ endommagement

L'action A2 coupe au niveau de la préposition. Il en résulte que l'arbre syntaxique construit sur cette dépendance est coupée en deux. Les structures à gauche et à droite de la préposition seront conservées :



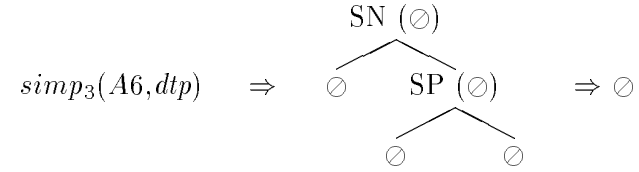
cycle suivant sénario ⇒ cycle, scénario

L'action A3 efface le troisième membre de la dépendance. Cela entraîne également la suppression de la préposition. L'arbre syntaxique d'origine est coupé en deux :



tôle de valeur ⇒ tôle

Enfin, l'action A6 supprime toute la dépendance. L'arbre syntaxique d'origine est à nouveau coupé en deux :



note de synthèse ⇒ ∅

6.2.2 Reconstitution du syntagme nominal

La reconstitution d'un syntagme dont les dépendances ont été filtrées prend en compte les actions d'élagage et les répercute dans la structure du syntagme. Un long groupe nominal pourra ainsi être décomposé en plusieurs, ou réduit à un syntagme plus simple. Les syntagmes nominaux restants seront considérés comme pertinents car nettoyés des combinaisons de propriétés linguistiques jugées non pertinentes.

Contrôle de cohérence Lorsque tout ou partie d'une dépendance est effacée au sein d'un arbre syntaxique, cela produit souvent des incohérences. Ainsi l'arbre originel peut être coupé en deux. Un des arbres restant peut se terminer par une préposition. L'autre peut commencer par une préposition, un adjectif. Il convient donc de contrôler la cohérence des arbres ainsi amputés. Nous décrivons maintenant à partir des figures 6.2, 6.3, 6.4, 6.5 et 6.6 les principales incohérences qui surviennent et comment nous les corrigeons. Les figures représentent les syntagmes sous la forme de dépendances syntaxiques. Une catégorie entourée d'un cercle continu indique qu'elle a

FIG. 6.2 – *Exemple de répercussion pour un nom seul.*

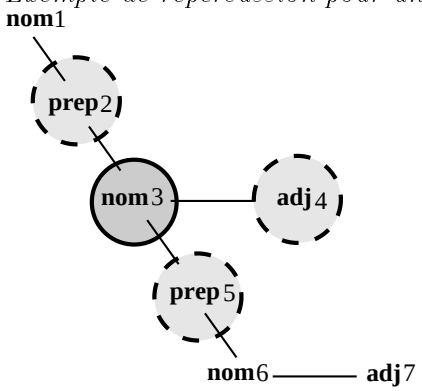
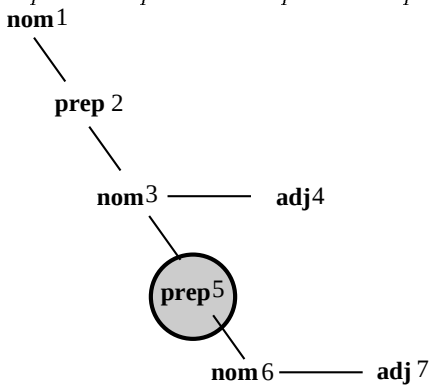


FIG. 6.3 – *Exemple de répercussion pour une préposition seule*



été supprimée par une action d'élagage. Un cercle en pointillé indique la propagation de la suppression autour du point de suppression initial, pour assurer la cohérence du ou des syntagmes finaux.

En figure 6.2, le nom nom_3 effacé est le modifieur (ou l'argument) d'un autre nom nom_1. Il est introduit par une préposition $prep_2$. Ce nom nom_3 effacé est aussi modifié par un adjectif adj_4 et un groupe prépositionnel ($prep_5$–nom_6). Pour conserver la cohérence de l'arbre, il faudra donc supprimer l'adjectif adj_4 et les prépositions $prep_2$ et $prep_5$. Les syntagmes finaux retenus seront donc : nom_1 et nom_6–adj_7 . En figure 6.3, seule la préposition $prep_5$ a été effacée. Les syntagmes restants sont : nom_1–$prep_2$–(nom_3–adj_4) et nom_6–adj_7. En figure 6.4, l'adjectif adj_2 est effacé. Le syntagme final est nom_1–$prep_3$–nom_4. En figure 6.5, le nom nom_1 est effacé. Il est modifié par un groupe prépositionnel introduit par la préposition $prep_2$. Celle-ci est supprimée. Le syntagme conservé est : nom_3–adj_4. En figure 6.5, le nom nom_1 est effacé. Il est modifié par un groupe prépositionnel introduit par la préposition $prep_2$. Celle-ci est supprimée. Le syntagme conservé est : nom_3–adj_4. Enfin, en figure 6.6, la dépendance nom_1-$prep_2$-nom_3 est supprimée. L'adjectif adj_7 est effacé. Le syntagme résultant est le nom simple nom_6.

FIG. 6.4 – *Exemple de répercussion pour un adjectif postposé*

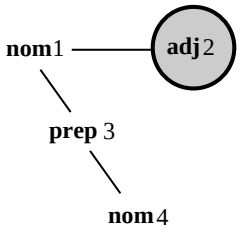


FIG. 6.5 – *Autre exemple de répercussion pour un nom seul*

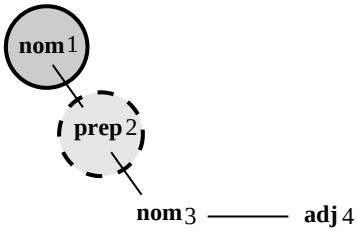
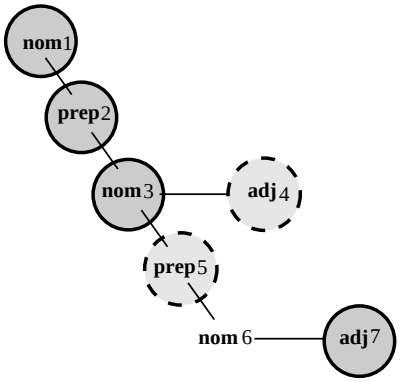


FIG. 6.6 – *Exemple de répercussion après suppression d'une dépendance à trois position et d'un adjectif*



Traitement de la coordination La distribution des éléments coordonnés est implémentée pour les cas les plus simples (adjectifs coordonnés, ou noms coordonnés). Les groupes coordonnés complexes ne sont pas traités). Il y a deux raison à cela. La première est que sorti d'un environnement de développement de grammaire, ce genre de traitement est difficile à implémenter. Il demande des fonctions de haut niveau s'appuyant sur un formalisme grammatical. Or nos traitements s'effectuent hors de l'environnement AlethIP, qui aurait permis de le faire. La seconde raison est que la coordination est souvent mal traitée par AlethIP (pour les cas complexes et les cas de rattachement ambigus). Par exemple, il arrive que l'on obtienne comme résultat d'analyse de la phrase une forêt d'arbres, avec un coordonnant isolé placé au niveau de profondeur zéro dans l'arbre. Dans ce cas, l'analyse présente toujours de multiples incohérences. Distribuer les groupes coordonnés situés de part et d'autre du coordonnant multiplie alors les incohérences.

Attribution d'un score de pertinence Nous n'avons pas encore implémenté la notion de score de pertinence. Nous présentons toutefois plusieurs facteurs qui permettent de définir un tel score. Il pourrait dépendre :

1. D'un score individuel affecté à chaque dépendance qui le compose. En effet, les filtres sont des descriptions associées à une action d'élagage syntaxique. Outre cette action d'élagage est également associée une opération de type $score = constante$ ou $score = score \pm constante$.

2. De la forme du filtre. Les dépendances qui ont été identifiées dans l'échantillon d'apprentissage positif augmentent la pertinence du SNP. Le score attribué doit alors être fonction de la spécificité ou de la généralité de la description du filtre. On peut convenir qu'une dépendance identifiée comme pertinente après un fort relâchement de ses attributs linguistiques (exemple : un nom avec un adjectif de couleur) accroît plus faiblement la pertinence que lorsqu'elle est identifiée avec une description plus précise (exemple : un nom dont la graphie est *tarif* avec un adjectif de couleur, comme `tarif bleu`).

3. Du nombre d'opérations d'effacement ou de suppression effectuées pour aboutir à la forme résiduelle.

Exemples de filtrage

Exemple 1 Considérons le syntagme nominal maximal identifié par l'analyse syntaxique : *amélioration de la connaissance des phénomènes de la zone d'assèchement des tubes de GV chauffés au sodium* (voire figure 6.7). Le filtrage à partir du profil I a permis d'éliminer les dépendances suivantes (voire figure 6.8) : `amélioration de connaissance`, `connaissance de phénomène`, et `phénomène dans zone`. Sur cette dernière dépendance, seule la tête «*phénomène*» est effacée : «zone» a été conservée. Le syntagme pertinent retenu après le filtrage est donc : «*zone d'assèchement des tubes de*

GV chauffés au sodium» (voire figure 6.9). Et s'il est soumis au profil II, plus sévère, il ne reste que deux syntagmes: «*zone d'asséchement*» et «*tubes de GV*».

FIG. 6.7 – *arbre initial*

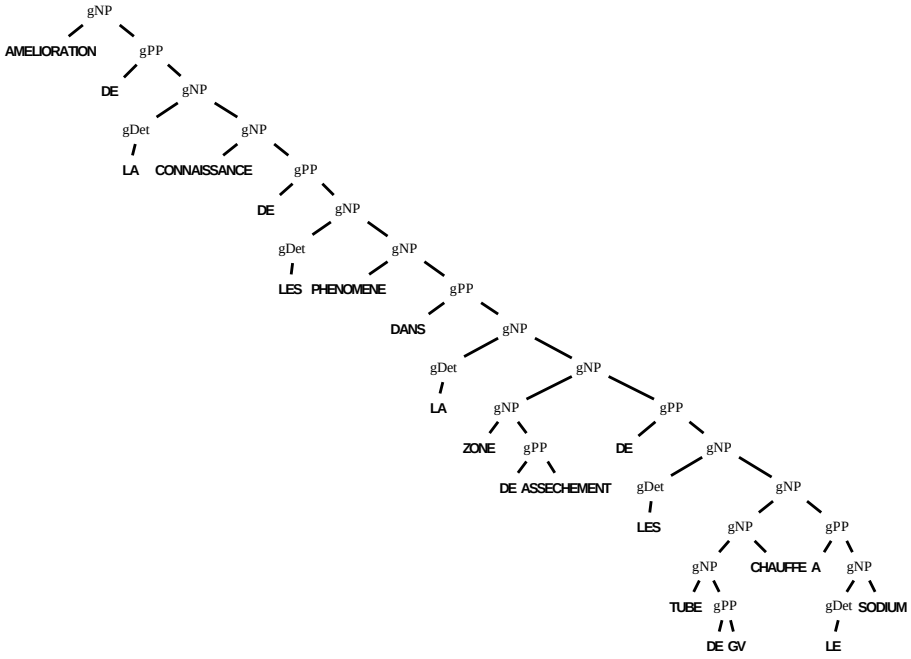
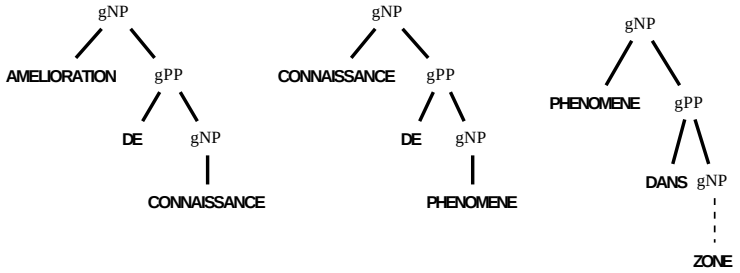
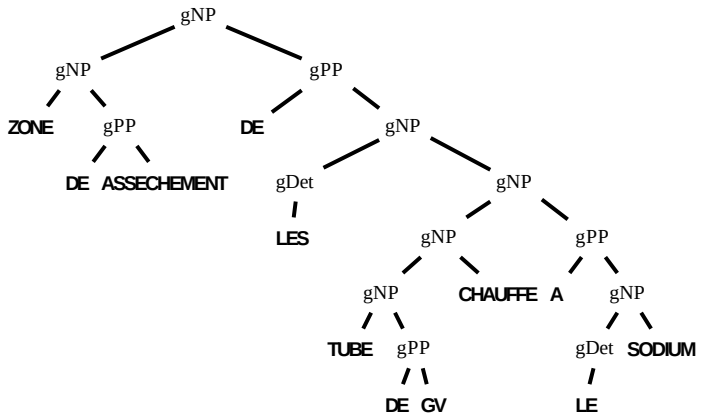


FIG. 6.8 – *Dépendances supprimées*



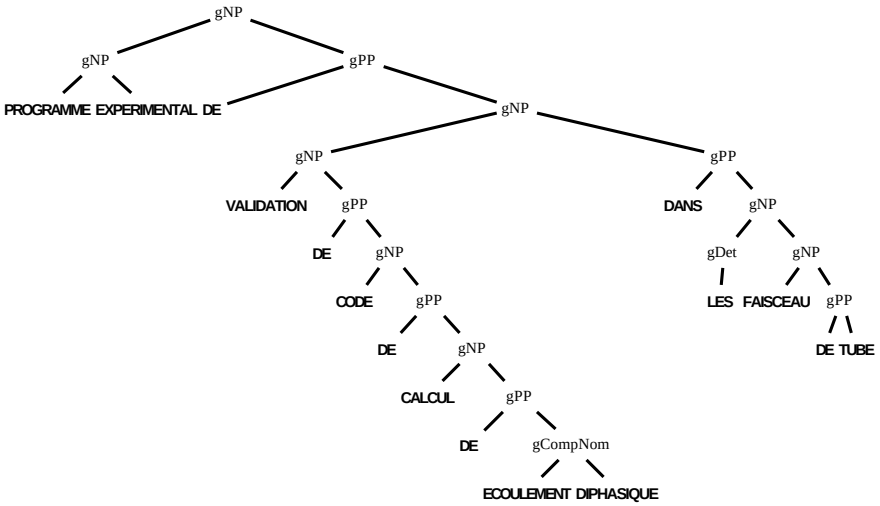
Exemple 2 Considérons le syntagme nominal maximal identifié par l'analyse syntaxique : «*programme expérimental de validation de code de calcul d'écoulement diphasique dans les faisceaux de tubes*» (voire figure 6.10). Le filtrage à partir du profil I a permis d'éliminer les dépendances suivantes (voire figure 6.11) : `programme expérimental`, `programme de validation`. Par ailleurs, une coupure a été réalisée au ni-

FIG. 6.9 – *SNP final*



veau de la préposition *dans*[6]. Les syntagmes pertinents retenus après le filtrage sont donc : «*code de calcul d'écoulement diphasique*» et «*faisceau de tube*».

FIG. 6.10 – *Arbre initial*



6. expliquer pourquoi : possibilité d'activer des règles de coupures au niveaux de certaines prépositions en fonction de leur contexte syntaxique.

FIG. 6.11 – *Dépendances supprimées*

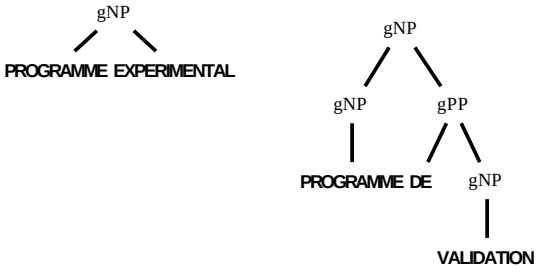
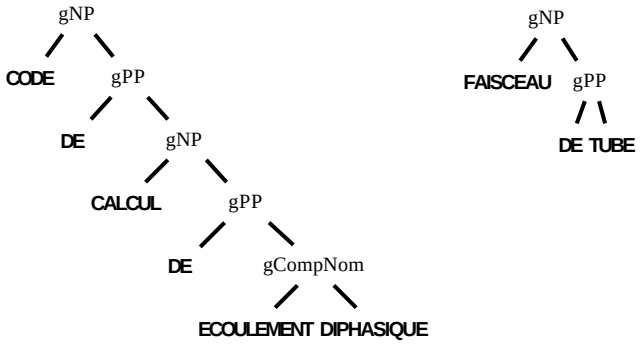


FIG. 6.12 – *SNP finaux*



6.3 Les protocoles de filtrage

6.3.1 Protocole 1 : le filtrage par rapport à un profil complet : négatif–positif

Ce type de filtrage s'applique lorsque le profil a été constitué avec des échantillons négatifs et des échantillons positifs. Dans l'arbre du syntagme nominal à filtrer, les dépendances relevant de descriptions négatives sont élaguées suivant l'action associée au filtre. Les dépendances relevant de descriptions positives sont conservées augmentant ainsi le score du groupe nominal. Enfin, il est des dépendances ne relevant d'aucune description dans le profil. Celles-ci ne peuvent être évaluées. Elles ne subissent aucune action de filtrage. Il y a deux méthodes distinctes pour déterminer si la dépendance.

L'intersection des descriptions de dépendances pertinentes et non pertinentes est vide. Cependant, en générant la combinatoire des différents attributs de la description d'une nouvelle dépendance, c'est-à-dire en sous-déterminant progressivement la description linguistique de la dépendance, il se peut que soit engendrées des combinaisons pertinentes et non pertinentes au regard du contenu du profil. Il font donc établir une priorité de lecture dans le profil : pour une description donnée, commence-t-on par vérifier son statut parmi l'ensemble des descriptions pertinentes ou parmi

celles qui sont non pertinentes?

TAB. 6.1 – *Protocole de filtrage 1 : deux algorithmes possibles*

Soit une dépendance décrite par un ensemble d'attributs linguistiques.
Générer l'ensemble E_{comb} des combinaisons d'attributs linguistiques.
Ordonner les descriptions de E_{comb} des plus riches aux plus pauvres.

MÉTHODE 1

T1 = *T2* = *Faux*
Pour chaque combinaison *Comb* de E_{comb} et Tant que *T1* est *Faux* {
T1 = *T1* ou (la description *Comb* est-elle NON PERTINENTE dans le profil?)
}

Si *T1* est *Vrai* Alors «La dépendance n'est pas pertinente»
Sinon
Si *T1* est *Faux* Alors
Pour chaque combinaison *Comb* de E_{comb} et Tant que *T2* est *Faux* {
T2 = *T2* ou (la description *Comb* est-elle PERTINENTE dans le profil?)
}

Si *T2* est *Vrai* Alors «La dépendance est pertinente»
Sinon
Si *T2* est *Faux* Alors «On ne peux rien dire sur la description *Comb*»

MÉTHODE 2

Pour chaque combinaison *Comb* de E_{comb} Faire {
T1 = (la description *Comb* est-elle NON PERTINENTE dans le profil?)
Si *T1* est *Vrai* Alors «La dépendance n'est pas pertinente», Sortir du parcours.
Sinon
Si *T1* est *Faux* Alors
T2 = (la description *Comb* est-elle PERTINENTE dans le profil?)
Si *T2* est *Vrai* Alors «La dépendance est pertinente», Sortir du parcours.
}

Si *T1* et *T2* sont *Faux* Alors «On ne peux rien dire sur la description *Comb*»

Les algorithmes montrés en table 6.1 proposent deux solutions possibles. Dans les deux cas nous donnons priorité au fltrage négatif. Nous avons choisi la première méthode de la table 6.1. Ainsi on parcourt toutes les combinaisons d'attributs générées, des plus riches au moins riches, pour les comparer aux descriptions non pertinentes. Si ce parcours ne permet pas d'identifier de description, un nouveau parcours est effectué sur les descriptions pertinentes du profil. Si à nouveau aucune description n'est identifiée, alors rien ne peut être dit sur la pertinence de la dépendance à filtrer.

Nous n'avons pas encore testé la seconde méthode. Comme la première, elle donne la priorité au filtrage négatif en comparant d'abord la description à évaluer aux descriptions non pertinentes. Mais si la description à évaluer n'est pas trouvée parmi les descriptions non pertinentes du profil, alors elle est immédiatement recherchée

parmi les descriptions pertinentes du profil, plutôt que d'être simplifiée pour être à nouveau comparée aux descriptions non pertinentes. Certaines erreurs qui nous avons comptabilisées lors de l'étude du comportement de la procédure d'apprentissage (voir chapitre suivant) pourraient trouver leur cause dans ces priorités définies pour le filtrage négatif et positif.

6.3.2 Protocole 2 : le filtrage par rapport à un profil incomplet : positif ou négatif

Si l'on n'a pas la possibilité ou le temps de définir un échantillon d'apprentissage constitué d'exemples négatifs et positifs, il est possible de ne faire appel qu'à des exemples positifs ou négatifs. Par exemple, si l'on souhaite enrichir un thesaurus terminologique, on récupère l'ensemble de ses termes complexes (on écarte les unitermes car ils n'entretiennent pas dépendances avec d'autres unités) pour en faire des exemples positifs. Au préalable, ces syntagmes nominaux auront été analysés syntaxiquement puis enrichis[7] (voir la chaîne de traitement au chapitre 3, en 3.1.2). C'est ce que nous avons fait pour le profil II pour lequel nous n'avions que des exemples positifs (voir chapitre 7, en 7.2.5). La même démarche peut être réalisée avec des exemples négatifs (si l'on ne dispose que de «déchets»). Dans les deux cas, le filtrage est simplifié.

Avec un profil constitué exclusivement d'exemples positifs, le filtrage consiste à supprimer les dépendances du syntagme qui n'appartiennent pas au profil. L'absence d'échantillon négatif radicalise l'action du profil ainsi constitué : aucun exemple négatif ne vient limiter les possibles généralisations abusives des exemples positifs. Après un premier passage sur le corpus, on peut s'aider des résultats pour compléter le profil avec les syntagmes écartés et retenus, et relancer à nouveau le filtrage avec de meilleurs exemples.

Avec un profil constitué exclusivement d'exemples négatifs, le filtrage consiste à supprimer les dépendances du syntagme qui appartiennent au profil. Là encore il y a des risques de généralisations abusives à partir de descriptions négatives dont on ne sait pas si elles pourraient être aussi positives.

6.4 Conclusion

Nous avons présenté dans ce chapitre une méthode de filtrage des syntagmes nominaux. Elle nécessite une représentation du syntagme en termes de dépendances syntaxiques. De ce fait, elle ne se base pas sur l'ordre des mots dans les syntagmes. Elle se situe donc à l'opposé d'une méthode de filtrage à base de patrons syntaxiques. Les dépendances sont évaluées isolément à partir des informations linguistiques qui leur sont attachées. Le filtrage doit ainsi prendre en compte la disparition ou l'altération d'une dépendance au sein d'un arbre syntaxique global, pour être en mesure

7. Dans ce cas on peut s'attendre à des résultats de catégorisation sémantique de moins bonne qualité, étant donné que les syntagmes soumis au système sont dépourvus de contexte.

de générer le ou les syntagmes résultants du filtrage.

Nous présentons dans le chapitre suivant la méthode d'apprentissage pour la construction de filtres de dépendances.

Chapitre 7

Apprentissage de filtres

L'utilisation d'une méthode d'apprentissage pour la construction de filtres a plusieurs avantages. Tout d'abord, cela permet d'adapter le système de filtrage à des textes de domaines d'activités différents, sans pour autant remettre en cause son architecture. Ensuite, cela rend plus aisée la constitution de différents points de vue en fournissant au système des exemples différents. Nous présentons maintenant la méthode d'apprentissage qui a été implémentée.

7.1 Apprentissage automatique

Il y a deux approches en apprentissage automatique. La première s'attache à découvrir les processus mentaux mis en oeuvre dans la cognition puis à les simuler avec des machines. Elle est du ressort de l'IA, branche des sciences cognitives. La seconde approche – celle qui nous concerne – est plus pratique et ne cherche pas à imiter la cognition. Elle est du ressort du traitement automatique de l'information et vise à construire des programmes qui extrapolent des connaissances à partir de données fournies en entrée, et qui éventuellement modifient leur comportement en fonction de ces données [MCM83], [Hut95].

Définition du concept à identifier Le problème soumis à une procédure d'apprentissage est l'identification de «concepts» étant donné leurs descriptions. Soit le concept de pertinence de dépendance syntaxique élémentaire. L'apprentissage doit permettre, à partir des descriptions morphologique, syntaxique et sémantique de dépendances élémentaires pertinentes, de trouver une représentation du concept de pertinence de dépendance élémentaire, afin d'être capable de prédire par la suite la pertinence de nouvelles dépendances[1]. Le résultat de l'apprentissage est l'émergence

1. Ceci peut aussi être modélisé comme un problème de catégorisation des dépendances élémentaires en fonction de leur description. Pour simplifier, on peut se représenter deux catégories possibles *est-pertinente* et *n'est-pas-pertinente*. En réalité, les catégories sont plus nombreuses, puisqu'elles correspondent aux actions d'élagage syntaxique décrites au chapitre 6

d'une représentation conforme au concept. Cette représentation peut être considérée comme une forme de connaissance tirée des exemples fournis au système.

Échantillon d'apprentissage Les données fournies en entrée à la procédure d'apprentissage sont appelées échantillon d'apprentissage. L'échantillon peut être caractérisé de différentes manières [Hut95]. Le domaine d'application définit l'origine des données présentes dans l'échantillon. Le domaine est dit fermé lorsque les paramètres qui conditionnent la solution sont en nombre fini et sont connus. Le domaine est dit ouvert lorsque l'on ne connait pas tous les paramètres qui déterminent la solution. Notre domaine d'application est qualifiable d'ouvert étant donné que l'évaluation de la pertinence d'une forme linguistique dépend de paramètres extralinguistiques que nous ne sommes pas en mesure d'intégrer au système. Dans le meilleur des cas, l'échantillon positif contient uniquement des exemples conformes, de même pour un échantillon d'exemples négatifs. Si l'échantillon contient des exemples inconsistants ou erronés ou en contradiction avec le concept qu'il décrit, l'échantillon est dit bruité. En ce qui nous concerne, les échantillons peuvent être bruités. Cela dépend de la méthode qui a été utilisée pour les constuire. Une sélection manuelle des exemples doit conduire à la constitution d'échantillons non bruités. Une construction automatique d'échantillons (comme la méthode exposée en C.1.1) introduit certainement du bruit. Enfin, l'échantillon peut contenir des exemples positifs et des exemples négatifs ou contre-exemples qui viennent relativiser ou délimiter la capacité à généraliser à partir des exemples positifs.

Description des exemples Le langage de description des exemples (ou *langage des instances* [Mou96]) décrit comment les données sont introduites en machine. Il s'agit d'ensembles d'*attribut=valeur*. Un exemple est équivalent à une conjonction d'*attribut=valeur*. Les attributs peuvent être numériques (par exemple, combien de fois cet adjectif est-il employé avec des noms différents?), et symboliques (par exemple, quelle forme lexicale ce nom prend-il?). Ces deux types d'attributs peuvent prendre des valeurs dans des ensembles ouverts ou fermés. Par exemple, les différentes valeurs de l'attribut «catégorie sémantique d'un nom» sont en nombre fini et correspondent aux catégories sémantiques qui ont été définies. Cette représentation peut être transformée en une représentation conforme à la logique des propositions. A condition que le nombre de valeurs prises par les attributs soit fini, il est alors possible d'appliquer des méthodes connexionnistes ou statistiques sur des données symboliques, en représentant les exemples par des vecteurs.

Généralisation Le mécanisme de la généralisation permet de synthétiser la connaissance présente dans les exemples. Il s'agit de trouver les propriétés communes aux exemples. Cette détection empirique de régularités est effectuée en calculant l'intersection des différents attributs de tous les exemples. La généralisation est définie d'après [Mit82] (cité dans [Mou96]) comme une recherche dans un espace, appelé *espace des généralisations* ou *espace des hypothèses*. Un parcours de cet espace, en-

gendré par les descriptions d'exemples (ou *langage des instances*) est très coûteux. On doit alors contraindre la forme des généralisations possibles, avec un langage de description des généralisations, appelé *langage des hypothèses* dans [Mit82]. La connaissance du domaine d'application permet de définir ce langage des hypothèses et d'introduire ainsi des présupposés dans l'apprentissage. Cela permet de limiter la combinatoire et l'exploration des attributs à généraliser.

7.2 Un apprentissage symbolique

7.2.1 Apprentissage inductif et système automatique

L'apprentissage à partir de données empiriques est nécessairement inductif. A. Chalmers [Cha87] énonce des limites à la méthode inductive, notamment qu'il est possible que les mêmes prémisses (les faits observés) ne conduisent pas toujours à la même conclusion, quand bien même ce rapport de cause à effet aurait été observé de nombreuses fois. Si l'induction présente certains risques pour tirer de l'observation des lois générales, elle paraît en revanche tout à fait appropriée dans le cadre d'une approche informatique.

Si en partant de descriptions empiriques, l'induction donne de mauvais résultats, il faut en attribuer la cause à des descriptions qui étaient incomplètes ou incorrectes. Or les descriptions sont introduites par l'homme et correspondent aux résultats de certaines observations. Il faut donc s'en remettre à la qualité des descriptions entrées en machine par l'homme et supposer qu'il a échoué dans ses observations. Mais pas nécessairement : l'observation précédant la description, il y a une différence substantielle entre les descriptions et les observations. Quand bien même les observations faites par l'homme lui permettraient de saisir une vérité ou le principe d'une loi, les descriptions entrées en machine, ne permettraient pas nécessairement de la modéliser correctement, surtout en matière de phénomènes sémantiques non quantifiables. On voit donc se dessiner une nette frontière entre l'activité du linguiste et celle du linguiste-informaticien. Le linguiste *observe*: il doit saisir l'objet observé dans son principe propre, dans sa réalité, il doit le saisir de l'intérieur. Sa conscience participe activement au processus. En revanche, le linguiste-informaticien *décrit*, il doit se limiter à une forme d'observation purement formelle : l'objet étudié ne peut être saisi que dans sa forme, et pas dans sa nature propre. Et s'il est saisi dans sa nature propre, ce doit nécessairement être exprimé avec des formes. Ce qui est observé doit être objectivé, sans cela, l'implémentation n'est pas envisageable.

Ainsi la méthode inductive ne pose-t-elle pas de problème lorsqu'elle est portée sur un système automatique. Bien au contraire, elle permet de vérifier si les descriptions qui ont été faites des phénomènes à prédire sont pertinentes, et dans le cas présent de constater l'écart entre la réalité linguistique et sa modélisation.

7.2.2 Intérêt de l'apprentissage symbolique

Nous avons opté pour une méthode symbolique plutot que statistique car nous souhaitions que l'apprentissage génère des règles linguistiquement interprétables. En fournissant des données linguistiques sous forme de règles (telle configuration linguistique implique telle action d'élagage dans l'arbre syntaxique d'origine de la dépendance décrite), nous souhaitions obtenir d'autres données linguistiques, présentées sous une forme identique mais plus générale. A. Hutchinson [Hut95] cite entre autres caractéristiques des systèmes à base de règles :

1. Ils permettent de retrouver des exemples dans les échantillons qui sont la cause d'un certain comportement.

2. Les règles produites sont interprétables. Les régularités mises en évidence par le système sont intéressantes en soi d'un point de vue linguistique. Elles représentent certains éléments constitutifs d'une grammaire de corpus spécialisée.

3. On peut étudier l'action d'une seule règle ou d'un petit groupe de règles. Ceci est pratique pour vérifier le bon fonctionnement du système et en faire la mise au point.

4. Ils autorisent la mise à jour incrémentale des échantillons d'apprentissage. Cette possibilité est intéressante car elle permet de mettre à jour les échantillons par un processus en spirale : après définition d'un premier échantillon, apprentissage puis filtrage, les bons et les mauvais résultats peuvent être récupérés pour alimenter la base d'échantillons positifs et négatifs. Cette pratique est avantageuse pour augmenter la taille des échantillons en toute sécurité et calibrer progressivement et précisément le profil de filtrage.

7.2.3 Langage des instances

Chaque exemple de l'échantillon est une dépendance élémentaire définie par son type (voir 5.2.1) et ses caractéristiques linguistiques. Deux types de descriptions sont définis, selon que la dépendance est constituée de trois ou deux éléments lexicaux.

Dans le tableau 7.1 sont déclarés les attributs linguistiques associés aux dépendances de type NOM (première colonne) PRÉPOSITION (deuxième colonne) NOM (troisième colonne). A la place de chaque nom, on peut trouver, une autre catégorie comme un sigle, un inconnu, etc. Tous les attributs descriptifs sont listés dans chaque colonne. On retrouve les traits asociés aux catégories grammaticales : le trait `Cat` (nom, nom propre, nom inconnu, sigle, ...), le trait `Morpho` (genre et nombre), `Xcons` (comportement syntaxique), `Suffix` (type de suffixe), `Csem` (catégorie sémantique) et `Coupe` (effacer et séparer).

Sur le deuxième nom, le trait `Det` précise quel type de déterminant introduisait le nom. Les nombres entre parenthèses indiquent la cardinalité de l'ensemble des valeurs que peuvent prendre les attributs. Pour l'attribut `Graphie`, cela ne peut être déterminé d'avance (cela dépend du lexique dans le corpus). Pour la préposition, le seul attribut

TAB. 7.1 – *Attributs descriptifs (et cardinalité des ensemble de valeurs prises par ces attributs) associés aux dépendances à trois positions*

NOM_1 (XX)	PREP	NOM_2 (XX)	
Graphie (x)	Graphie(750)	Graphie	
Cat (5)		Cat	
Morpho (4)		Morpho	
Xcons (5)		Xcons	, type ⇒ Action d'élagage
Suffix (90)		Suffix	
Csem (72)		Csem	
Coupe (2)		Coupe	
		Det(5)	

TAB. 7.2 – *Attributs descriptifs (et cardinalité des ensemble de valeurs prises par ces attributs) associés aux dépendances à deux positions*

NOM (XX)	ADJ (NOM,XX,PPAS,PPRES)	
Graphie (x)	Graphie	
Cat (5)	Cat	
Morpho (4)	Morpho	
Xcons (5)	Xcons	, type ⇒ Action d'élagage
Suffix (70)	Suffix	
Csem (72)	Csem (50)	
Coupe (2)	Coupe	

est une graphie et on compte 750 prépositions dans le dictionnaire *AlethDic* (prépositions simples et complexes). En pratique seule une vingtaine sont utilisées. A droite de la flêche (⇒) est définie l'action d'élagage dans l'arbre syntaxique dont est issue la dépendance élémentaire. Cela correspond en fait à la définition de la pertinence de la dépendance élémentaire. Si l'action est nulle, la dépendance est considérée comme pertinente. Si l'action efface la dépendance, la dépendance est considérée comme non pertinente. Les actions intermédiaires qui n'effacent qu'une partie de la dépendance signifient que cette partie de la dépendance n'est pas pertinente (le modifieur, la tête, la préposition). Cette partie en question est effacée, ce qui permet de conserver l'autre au sein de l'arbre syntaxique

Le tableau 7.2 donne la description utilisée pour les dépendances de type NOM ADJ. Le nombre de suffixes possibles pour les adjectifs n'est ici que de 70. le nombre de catégories sémantiques est 50. La catégorie grammaticale pour la position modifieur peut être adjectif, nom, inconnu, participe passé, participe présent.

7.2.4 Présupposés d'apprentissage et langage des hypothèses

Un profil de filtrage – construit à partir d'un échantillon – peut être défini comme un ensemble de contraintes linguistiques strictes. Pour augmenter la couverture de ce profil, il faut relâcher, de manière contrôlée, ces contraintes. Ce contrôle dans le relâchement est effectué d'après la connaissance linguistique du domaine d'application. Il permet de minimiser l'exploration de l'espace de généralisation en définissant un langage des hypothèses. C'est à dire que l'on prédétermine en quelque sorte la forme des connaissances résultant de l'apprentissage.

Nous montrons maintenant à partir d'un échantillon minimal, à partir de quels critères nous restreignons la combinatoire des attributs, tout en gardant la possibilité d'étendre la couverture du profil construit sur l'échantillon, pour lui permettre d'extrapoler sur la pertinence de nouvelles formes.

Exemple sur un échantillon minimum

Nous avons choisi comme exemple une dépendance de type NOM ADJECTIF construite sur le nom *robinetterie*. Pour définir ce profil minimal, nous déclarons autour du nom *robinetterie* un ensemble de dépendances élémentaires jugées pertinentes décrites en table 7.3, ou non pertinentes décrites en table 7.4. Le profil ainsi constitué accepte et rejette les seules dépendances déclarées dans les échantillons respectivement positifs et négatifs. Les dépendances élémentaires acceptées et rejetées à partir de ces échantillons sont montrées en table 7.5.

TAB. 7.3 – *Exemple d'échantillon positif pour les modifieurs adjectivaux du nom* robinetterie.

/frNom{Csem=26 [a];Morpho=2}	ROBINETTERIE	/adjStd{Morpho=2;Csem=141}	METALLIQUE
/frNom{Csem=26;Morpho=2}	ROBINETTERIE	/adjStd{Morpho=2;Csem=142}	AUTOMATIQUE
/frNom{Csem=26;Morpho=2}	ROBINETTERIE	/adjStd{Morpho=2;Csem=139}	NUCLEAIRE
/frNom{Csem=26;Morpho=2}	ROBINETTERIE	/adjStd{Morpho=2;Csem=140}	INDUSTRIEL

[a] Signification des traits attachés aux catégories grammaticales

Csem = 26	Nom d'appareil (ENTITÉ-CONCRET-ARTEFACT-APPAREIL)
Morpho = 2	Féminin singulier
Csem =139,142	Adjectifs qui déclarent des propriétés de toutes sortes (travail de codage encore incomplet).
Csem = 141	Adjectif qui déclare des propriétés en relation avec des substances
Csem = 142	Adjectifs déclarant la propriété «relatif au» nom dérivé (ex: industriel: relatif à l'industrie)

2. La valeur des modalités est expliquée dans l'ouvrage [BC90].

TAB. 7.4 – *Exemple d'échantillon négatif pour les modifieurs adjectivaux du nom* robinetterie

/frNom{Csem=26;Morpho=2}	ROBINETTERIE	/adjStd{Morpho=2;Csem=128 [a]}	IMPORTANT
/frNom{Csem=26;Morpho=2}	ROBINETTERIE	/adjStd{Morpho=2;Csem=119}	RECENT
/frNom{Csem=26;Morpho=2}	ROBINETTERIE	/adjStd{Morpho=2;Csem=130}	NECESSAIRE
/frNom{Csem=26;Morpho=2}	ROBINETTERIE	/adjStd{Morpho=2;Xcons=6}	EXEMPT
/frNom{Csem=26;Morpho=2}	ROBINETTERIE	/adjStd{Morpho=2;Xcons=6}	SUJET

[a] Signification des traits attachés aux catégories grammaticales

Csem = 119	Adjectifs produisent une localisation temporelle dans le passé.
Csem = 128	Adjectifs qui qualifient de manière subjective (modalité III de Culioli[2])
Csem = 130	Adjectifs qui ont une valeur déontique (modalité IV de Culioli)
Xcons= 6	Trait affecté à des adjectifs construits avec une préposition (exempt de, nécessaire à, sujet à)

TAB. 7.5 – *Dépendances élémentaires acceptées et* ^{x}*rejetées sans relâchement de contraintes.*

robinetterie métallique	^{x}robinetterie importante
robinetterie automatique	^{x}robinetterie récente
robinetterie nucléaire	^{x}robinetterie nécessaire
robinetterie industrielle	^{x}robinetterie exempte
	^{x}robinetterie sujette

Relâchement des contraintes linguistiques

Différents types de relâchements sont envisageables pour augmenter la couverture du profil. Ils n'ont pas tous le même impact :

1. **Suppression de la contrainte de nombre.** Les mêmes dépendances élémentaires sont acceptées ou rejetées indifféremment au pluriel et au singulier. Par exemples les dépendances acceptées par notre profil minimal figurent en table 7.6.

2. **Remplacement de la graphie par la catégorie sémantique correspondante, sur l'un des deux membres.** Il s'agit du nom ou de l'adjectif pour des dépendances du type NOM ADJ. Il s'agit du premier nom ou du second nom pour des dépendances du type NOM_1 PREP NOM_2. Les dépendances qui seraient acceptées, pour notre exemple, en remplaçant le nom par un nom de même catégorie sémantique (si nous limitons à trois les noms de cette catégorie) sont montrées en table 7.7. La table 7.9 montre des exemples de dépendances élémentaires rejetées en relâchant la contrainte de la forme lexicale de l'adjectif.

3. **Remplacement de la graphie par la catégorie sémantique correspondante sur les deux membres.** Pour se faire une idée des RSE acceptées et rejetées, il faut combiner les résultats des tables 7.7 et 7.8. Ainsi pourront être acceptées des dépendances élémentaires comme `turbine électromagnétique`, `pompe communautaire` et rejetées des dépendances élémentaires comme : `pompe récente`,

TAB. 7.6 – *Dépendances élémentaires acceptées avec relâchement du nombre.*

robinetterie métallique	robinetteries métalliques
robinetterie automatique	robinetteries automatiques
robinetterie nucléaire	robinetteries nucléaires
robinetterie industrielle	robinetteries industrielles

TAB. 7.7 – *Dépendances élémentaires acceptées avec relâchement de la graphie sur le nom. Le nom remplacé a la même catégorie sémantique que* robinetterie.

robinetterie métallique	turbine métallique	pompe métallique	...
robinetterie automatique	turbine automatique	pompe automatique	...
robinetterie nucléaire	turbine nucléaire	pompe nucléaire	...
robinetterie industrielle	turbine industrielle	pompe industrielle	...

turbine indispensable. Un tel relâchement est intéressant si les catégories sémantiques utilisées sont très spécialisées. Sinon, il y a un risque de retenir ou de supprimer un trop grand nombre de dépendances élémentaires. Comme nos catégories sémantiques sont très générales, nous attribuons un score de pertinence moyen aux dépendances élémentaires retenues par ce type de contrainte.

4. **Remplacement de la graphie par son suffixe correspondant sur l'un des deux membres** En applicant une telle substitution, on accorde crédit aux valeurs sémantiques «floues» portées par les suffixes. Ainsi robinetterie électrique pourra être acceptée.

5. **Abandon de la contrainte du déterminant sur les dépendances à trois positions nom$_1$ prep nom$_2$**. Normalement, sur ces dépendances, la présence et le type du déterminant sont mémorisés, ce qui permet de distinguer des dépendances qui diffèrent par la seule présence du déterminant. Par exemple en supprimant cette information, les deux syntagmes *cable à isolation synthétique* et *cable à l'isolation défectueuse* auront la dépendance cable à isolation en commun. Ce relâchement est justifié seulement si l'on souhaite abolir la différence entre des syntagmes hypothétiquement dénominatifs ou non dénominatifs, la présence du déterminant en tête du modifieur prépositionnel plaidant souvent pour un syntagme non dénominatif (surtout s'il s'agit d'un adjectif démonstratif ou possessif).

6. **Relâchement libre de la graphie en conservant sa catégorie grammaticale (nom, adjectif)**. On acceptera alors les dépendances élémentaires du type robinetterie ADJ ou NOM métallique (à partir de robinetterie métallique).

Langage des hypothèses

Si nous ne contraignons pas la combinatoire des attributs des descriptions, l'espace de généralisation devient trop important. Soit V_{Cat} le nombre de valeurs que peut prendre le trait Cat, soit V_{Morpho} le nombre de valeurs que peut prendre le

TAB. 7.8 – *Dépendances élémentaires acceptées avec relâchement de la graphie sur l'adjectif. L'adjectif remplacé a la même catégorie sémantique que l'adjectif substitué.*

robinetterie métallique	robinetterie inoxydable	robinetterie poreuse	...
robinetterie automatique	robinetterie isotherme	robinetterie modulaire	...
robinetterie nucléaire	robinetterie électromagnétique	robinetterie radioactive	...
robinetterie industrielle	robinetterie communautaire	...	

TAB. 7.9 – *Dépendances élémentaires rejetées avec relâchement de la graphie sur l'adjectif. Le nom accepté a la même catégorie sémantique que l'adjectif.*

robinetterie importante	robinetterie conventionelle	robinetterie particulière	...
robinetterie sujette (à)	robinetterie relative (à)	robinetterie envisagée (par)	...
robinetterie récente	robinetterie ancienne	...	
robinetterie nécessaire	robinetterie indispensable	...	

trait `Morpho`, etc. Le nombre maximum théorique de combinaisons d'attributs pour une dépendance élémentaire à deux unités lexicales est de $(V_{Cat} * V_{Morpho} * V_{Xcons} * V_{Suffixe} * V_{Coupe} * V_{Csem} * x_1) * (V_{Cat} * V_{Morpho} * V_{Xcons} * V_{Suffixe} * V_{Coupe} * V_{Csem} * x_2) = 254.016.000.000 * x_1 * x_2$, avec x_1 et x_2 correspondant à la taille du lexique pour la catégorie grammaticale concernée. Ce nombre maximum pour une dépendance élémentaire à trois unités lexicales atteind $167.961.600.000.000 * x_1 * x_2$ combinaisons d'attributs, étant donné qu'il faut prendre en compte l'information de déterminant sur le deuxième nom (5 valeurs possibles) et la préposition (en limitant à 20 le nombre de prépositions possibles). Même en ne prenant pas en compte les formes graphiques du lexique ces chiffres restent trop importants.

Nous avons restreint la combinatoire en prenant en compte les attributs linguistiques «à relâcher» pour étendre la couverture du profil de filtrage. Toutes les informations linguistiques descriptives utilisées pour modéliser la notion de SNP sont prise en compte (voir chapitre 2). Les combinaisons d'attributs qui ne donnent pas lieu a priori à des généralisations intéressantes ne sont pas prises en compte. Par exemple combiner les deux seules catégories (`Cat Prep. Cat`) n'a pas d'intérêt puisque l'on est certain de retrouver ce schéma dans les exemples positifs et négatifs. Les combinaisons retenues sont montrées en tables 7.10 et 7.11. En réalité, celles-ci peuvent être modifiées, la procédure d'apprentissage prenant en paramètre les schémas combinatoires à exploiter. Le nombre de combinaisons d'attributs générées respectivement pour une dépendance à deux unités et à trois unités est donc de 15 (voir table 7.10) et 8 (voir table 7.11).

7.2.5 Construction des profils de filtrage

Le processus d'apprentissage produira des généralisations de niveaux différents, selon le nombre et le type des attributs qui ont été combinés. Il y aura ainsi des schémas assez stricts (par exemple la dépendance elle-même au pluriel ou singulier) et des schémas beaucoup moins contraints (comme `NOM métallique`). Par exemple, à partir de la dépendance `ordinateur récent` déclarée comme exemple négatif, des

TAB. 7.10 – *Exemple de combinatoire d'attributs linguistiques pour les dépendances à deux unités linguistiques*

NOM	ADJ
Cat Morpho Xcons Coupe Csem Sufx Graphie	Cat Morpho Xcons Coupe Csem Sufx Graphie
Cat Morpho Xcons Coupe Csem Sufx Graphie	Cat Morpho Xcons Coupe Csem Sufx
Cat Morpho Xcons Coupe Csem Sufx Graphie	Cat Morpho Xcons Coupe Csem
Cat Morpho Xcons Coupe Csem Sufx Graphie	Cat Morpho Xcons Coupe
Cat Morpho Xcons Coupe Csem Sufx	Cat Morpho Xcons Coupe Csem Sufx Graphie
Cat Morpho Xcons Coupe Csem	Cat Morpho Xcons Coupe Csem Sufx Graphie
Cat Morpho Xcons Coupe Csem	Cat Morpho Xcons Coupe Csem Sufx
Cat Morpho Xcons Coupe	Cat Morpho Xcons Coupe Csem Sufx Graphie
Cat Morpho Xcons Coupe	Cat Morpho Xcons Coupe Csem Sufx
Cat Morpho Xcons Coupe	Cat Morpho Xcons Coupe Csem
Cat Morpho Xcons	Cat Morpho Xcons Coupe Csem Sufx Graphie
Cat Morpho Xcons	Cat Morpho Xcons Coupe Csem Sufx
Cat Morpho	Cat Morpho Xcons Coupe Csem Sufx Graphie
Cat Morpho	Cat Morpho Xcons Coupe Csem Sufx
Cat	Cat Morpho Xcons Coupe Csem Sufx Graphie

TAB. 7.11 – *Exemple de combinatoire d'attributs linguistiques pour les dépendances à trois unités linguistiques*

NOM_1	PREP	N_2
Cat Morpho Xcons Coupe Csem Sufx Graphie	Prep.	Cat Det Morpho Xcons Coupe Csem Sufx Graphie
Cat Morpho Xcons Coupe Csem Sufx Graphie	Prep.	Cat Det Morpho Xcons Coupe Csem Sufx
Cat Morpho Xcons Coupe Csem Sufx Graphie	Prep.	Cat Det Morpho Xcons Coupe Csem
Cat Morpho Xcons Coupe Csem Sufx Graphie	Prep.	Cat Det Morpho Xcons Coupe
Cat Morpho Xcons Coupe Csem Sufx	Prep.	Cat Det Morpho Xcons Coupe Csem Sufx Graphie
Cat Morpho Xcons Coupe Csem	Prep.	Cat Det Morpho Xcons Coupe Csem Sufx Graphie
Cat Morpho Xcons Coupe	Prep.	Cat Det Morpho Xcons Coupe Csem Sufx Graphie
Cat Morpho Xcons Coupe	Prep.	Cat Det Morpho Xcons Coupe Csem Sufx

généralisation comme « *Un nom d'appareil modifié par un adjectif de localisation temporelle n'est pas une combinaison pertinente* » et comme « *Un nom d'appareil qui à pour suffixe* **-eur** *modifié par un adjectif donc la graphie est* **récent** *n'est pas une combinaison pertinente* » pourront être produites.

Un arbre de décision organise les dépendances aux contraintes linguistiques relâchées

A partir de chaque exemple de l'échantillon sont générés les schémas correspondant à toutes les combinaisons demandées d'attributs. L'apprentissage consiste en la généralisation des propriétés linguistiques que les exemples positifs et négatifs ne partagent pas. Ainsi si des exemples négatifs et positifs ont des dépendances élémentaires dont les combinaisons d'attributs sont communes, ces dernières seront omises. Cela signifie que de telles combinaisons de propriétés sont neutres, qu'elles n'apportent aucun élément pour la distinction entre dépendances pertinentes et non pertinentes. L'ensemble des schémas est ordonné des moins permissifs (le plus d'attributs combinés) aux plus permissifs (le moins d'attributs combinés) sous la forme d'un arbre de décision. A chaque schéma est associée une action d'élagage dans l'arbre syntaxique d'origine.

Différents modes de construction expérimentés

Pour être en mesure d'évaluer par la suite le gain apporté par certaines informations linguistiques, quatre modes de constitution des profils, ou quatre langages des hypothèses, ont été définis. Ils engendrent des généralisations différentes.

mode 0 Ce mode est celui que nous avons déjà décrit : toutes les informations linguistiques disponibles sont utilisées dans le mode 0.

mode 1 Ce mode est identique au mode 0, à ceci près que les informations de genre et de nombre ne sont pas prises en compte. En effet, c'est essentiellement pour traiter des cas d'exception comme *travaux publics* (le pluriel est obligatoire) que nous conservons le genre et le nombre.

mode 2 Ce mode est identique au mode 1, mais les informations de suffixe et de catégorie sémantique sont omises.

mode 3 Ce mode est semblable au mode 1, mais l'information de déterminant dans les dépendances de type NOM_1 PREP (det) NOM_2 est omise.

Pour donner une idée de la taille des profils, nous donnons le nombre des schémas générés en fonction du mode de constitution utilisé.

Le profil I

Ce profil I, défini par un consultant[3] en information d'EDF pour des besoins de veille technologique, est constitué de 17 578 dépendances élémentaires déclarées comme exemples positifs et de 11 629 dépendances déclarées comme exemples négatifs.

Mode	Nbr. schémas positifs	Nbr. schémas négatifs	Nbr. schémas communs (effacés).
0	72 313	49 441	4589 (6% , 9%)
1	62 354	39 547	4114 (6.5%, 10.4%)
2	27 203	18 336	1799 (6.6%, 9.8%)
3	61 419	38 457	4179 (6.8%, 10.8%)

Dans la première colonne figure le nombre de schémas positifs retenus et dans la seconde le nombre de schémas négatifs retenus (ce sont ces derniers qui causent une simplification du syntagme nominal à filtrer). La troisième colonne indique le nombre de schémas communs aux exemples positifs et négatifs avant d'effectuer les comptages des deux premières colonnes. Figure également entre parenthèses la proportion des schémas supprimés dans les exemples respectivement positifs et négatifs. Le mode 2 (pas de suffixe ni de catégorie sémantique) fait décroître de manière importante la taille du profil généré.

Le profil II

Le profil II, de taille beaucoup moins importante que le profil I, est constitué d'un unique échantillon d'exemples positifs (3900). Voici le nombre des schémas générés en fonction du mode de constitution utilisé :

Mode	Nbr. schémas positifs
0	7354
1	6261
2	2105
3	6005

Ce profil vise à filtrer des syntagmes nominaux pertinents pour le terminographe. Nous avons puisé dans le thesaurus EDF des exemples positifs pour le définir. Nous avons choisi les termes des deux domaines les plus représentés dans notre corpus d'après les résultats de l'application d'indexation automatique du département SID (APPAREILLAGE MÉCANIQUE et SCIENCES PHYSIQUES). Après plusieurs essais de filtrage infructueux, il s'est avéré que le profil ainsi constitué ne filtrait rien, ou plutot éliminait tous les syntagmes nominaux qu'on lui présentait. Une telle situation manifeste certainement un décalage important entre les termes du thesaurus et les termes

3. Je remercie Richard Quatrain pour son aide et son travail de validation des échantillons.

rencontrés en corpus[4]. Nous avons abandonné l'idée de filtrer des syntagmes propres à un domaine d'activité à partir des donnés du thesaurus. Pour garder des exemples tirés du thesaurus EDF nous avons alors procédé à l'intersection des dépendances élémentaires de thesaurus (12 605 dépendances *a priori* toutes pertinentes pour identifier des syntagmes pouvant intégrer le thesaurus) avec celles de notre corpus ARD (78 503). Nous avons obtenu un ensemble d'environ 3900 dépendances.

Avec un tel profil constitué exclusivement d'exemples positifs, le deuxième protocole de filtrage (défini au chapitre précédent, voir 6.3.2) doit être utilisé.

7.3 Evaluation

Pour évaluer le système d'apprentissage, nous avons constitué différents échantillons et jeux de tests à partir des exemples du profil I. Nous avons défini 4 profils d'évaluation contenant respectivement 20, 60, 90 et 100% des descriptions de l'échantillon d'apprentissage du profil I. Ces descriptions extraites ont été prélevées en proportions égales pour les exemples positifs (dépendances dites pertinentes) et pour les exemples négatifs (dépendances dites non pertinentes). Pour définir ces profils «partiels», les exemples ont été extraits aléatoirement. Les descriptions non retenues par la sélection aléatoire ont constitué les jeux de tests. Pour le profil à 100% (c'est-à-dire le profil I complet), c'est l'intégralité des exemples positifs et négatifs qui a servi à calculer la précision, c'est-à-dire le nombre de bonnes réponses données par le système sur le nombre de bonnes réponses attendues. La table 7.12 donne le nombre d'exemples sélectionnés pour les profils et les jeux de tests. Pour déterminer la précision, nous avons distingué pour les descriptions pertinentes et non pertinentes le nombre de bonnes réponses, le nombre de non reconnaissances et le nombre d'erreurs. Le nombre de bonnes réponses est le nombre de fois où le système à correctement déterminé la pertinence ou la non pertinence de la dépendance syntaxique. Le nombre de non reconnaissances est le nombre de fois où le système n'a trouvé aucune description dans le profil pour déterminer si la dépendance est pertinente ou non. Enfin, le nombre d'erreurs est le nombre de fois où le système s'est trompé : il a répondu «pertinent» au lieu de «non pertinent» ou inversement.

Dans la table 7.13 figurent les taux de précision, les taux de non reconnaissances et les taux d'erreurs en fonction des profils utilisés et des modes de constitution des profils appliqués.

Précaution Il convient tout d'abord de préciser que les exemples de dépendances alimentant le profil I ont été en partie sélectionnés manuellement avec l'interface conçue dans ce but (pour les dépendances non pertinentes surtout). Lors de cette sélection, seules les dépendances les plus fréquentes ont été jugées. Une autre consé-

4. L'hypothèse la plus probable est que les forts taux de couverture des deux domaines mentionnés sont dus à l'indexation de termes simples, voire de termes simples polysémiques, ce qui expliquerait que le profil de filtrage construit à partir des termes complexes des domaines ne permette de capter aucune forme linguistique issue de ces domaines

TAB. 7.12 – *Effectifs des échantillons et des jeux de tests pour l'évaluation de l'apprentissage à 20, 60, 90 et 100%, en fonction du mode de constitution du profil*

	Profils			
Proportion	20%	60%	90%	100%
Dép. non pertinentes	2325	6977	10 465	11629
Dép. pertinentes	3515	10 545	15819	17578
	Jeux de test			
Proportion	80%	40%	10%	100%
Dép. non pertinentes	9304	4652	1164	11629
Dép. pertinentes	14 063	7033	1759	17578

quence de la sélection manuelle est que l'on ne trouve pas beaucoup de doublons (c'est-à-dire des dépendances dont les descriptions sont identiques à d'autres dépendances) et ce pour des raisons d'économie de temps passé. Des doublons peuvent cependant apparaître en retirant certains attributs (comme les graphies). Ainsi, lorsque la sélection aléatoire retient certains exemples et pas d'autres, il peut en résulter la création de points aveugles dans la description des propriétés qui définissent la pertinence. Parce qu'il y a peu de redondance dans l'échantillon du profil I, il se peut que cette méthode d'évaluation ne soit pas adpatée à un profil défini de la sorte. Quoi qu'il en soit, les tests effectués montrent un certain nombre de résultats que nous allons commenter.

Lecture des résultats en fonction du nombre d'exemples utilisés L'écart entre la précision obtenue pour le profil à 20% et le profil à 60% tourne en moyenne autour de 10%. En triplant la quantité d'exemples, la précision n'augmente que de 10%. En contrepartie, le taux des dépendances non reconnues, pour lesquelles il n'existe pas de description, baisse également de 10% environ. Ces remarques sont valables pour la reconnaissance de dépendances pertinentes. Pour les dépendances non pertinentes, le gain tourne autour de 11%. Le taux d'erreur est par contre en moyenne de 1% supérieur aveec le profil à 60%. Les performances du profil à 90% augmentent peu pour le nombre de descriptions fournies en supplément. Le gain est de l'ordre de 3% pour le dépendances pertinentes et de 1% pour les dépendances non pertinentes (exemples négatifs). Pour les profils de 20 à 60%, le système «apprend vite». Pour les profils entre 60 et 90%, le système retient moins vite. En testant le profil avec les exemples qui ont permi de le définir, le taux de précision monte un peu au dessus de 95% pour les dépendances pertinentes et au alentour de 94% pour les dépendances non pertinentes.

Comment se fait-il que l'on obtienne pas 100% de précision avec le profil à 100%? Pour expliquer cela, il faut considérer le taux des non reconnaissances et le taux d'erreurs. A 100% le taux d'erreur est quasiment nul pour les dépendances pertinentes, il est nul pour les dépendances non pertinentes. C'est la non reconnaissance de cer-

TAB. 7.13 – *Résultats des performances des profils en fonction du nombre d'exemples qui a servi à les définir et en fonction du mode de constitution du profil*

Dépendances pertinentes (Profil à 20%)

Mode	Succès	Non reconnu	Erreur
mode 0	5261 (37%)	8260 (58.7%)	542 (3.8%)
mode 1	5629 (40%)	7809 (55.5%)	625 (4.4%)
mode 2	3203 (22%)	10 550 (75%)	310 (2.2%)
mode 3	5789 (41%)	7640 (54%)	634 (4.5%)

Dépendances non pertinentes (Profil à 20%)

Mode	Succès	Non reconnu	Erreur
mode 0	3621 (38.9%)	5059 (54.3%)	624 (6.7%)
mode 1	4471 (48.5%)	4161 (44.7%)	672 (7.2%)
mode 2	2831 (30.4%)	6111 (65.5%)	362 (3.8%)
mode 3	4594 (49.37%)	4000 (43%)	710 (7.6%)

Dépendances pertinentes (Profil à 60%)

Mode	Succès	Non reconnu	Erreur
mode 0	3264 (46.4%)	3381 (48%)	388 (5.5%)
mode 1	6542 (50%)	3319 (47.2%)	372 (5.2%)
mode 2	2181 (30%)	4610 (65.5%)	242 (3.4%)
mode 3	3657 (52%)	2995 (42.5%)	381 (5.4%)

Dépendances non pertinentes (Profil à 60%)

Mode	Succès	Non reconnu	Erreur
mode 0	2387 (51.3%)	1898 (40.7%)	367 (7.8%)
mode 1	2657 (57.11%)	1611 (34.6%)	384 (7.8%)
mode 2	1992 (42.8%)	2434 (52.3%)	266 (5.8%)
mode 3	2694 (57.9%)	1578 (33.9%)	380 (8.2%)

Dépendances pertinentes (Profil à 90%)

Mode	Succès	Non reconnu	Erreur
mode 0	871 (49.5%)	802 (45.5%)	86 (4.8%)
mode 1	934 (53%)	744 (43.4%)	81 (4.6%)
mode 2	576 (32.7%)	1126 (64%)	57 (3.2%)
mode 3	967 (54.9%)	708 (40.2%)	84 (4.7%)

Dépendances non pertinentes (Profil à 90%)

Mode	Succès	Non reconnu	Erreur
mode 0	589 (50.6%)	472 (40.5%)	103 (8.8%)
mode 1	674 (58.7%)	392 (33.6%)	98 (8.4%)
mode 2	497 (42.6%)	597 (51.2%)	70 (6%)
mode 3	683 (58.6%)	376 (32.3%)	105 (9%)

Dépendances pertinentes (Profil à 100%)

Mode	Succès	Non reconnu	Erreur
mode 0	16836 (95.7%)	678 (3.8%)	64 (0.2%)
mode 1	16673 (94.8%)	853 (4.8%)	51 (0.2%)
mode 2	16829 (95.7%)	729 (4.1%)	20 (0.1%)
mode 3	16634 (94.6%)	893 (5.0%)	51 (0.2%)

Dépendances non pertinentes (Profil à 100%)

Mode	Succès	Non reconnu	Erreur
mode 0	10950 (94.1%)	639 (5.5%)	0 (0%)
mode 1	10768 (92.6%)	861 (7.4%)	0 (0%)
mode 2	10929 (94%)	700 (6%)	0 (0%)
mode 3	10570 (91%)	1055 (9%)	0 (0%)

taines dépendances qui fait baisser la précision. Nous l'expliquons par le fait que lors de la construction du profil, une intersection est réalisée entre les combinaisons d'attributs des exemples positifs et négatifs. Nous supprimons ensuite les descriptions communes aux deux types d'exemples. Les combinaisons ainsi supprimées correspondent à des informations qui ne permettent pas distinguer ce qui est pertinent de ce qui ne l'est pas. Elles sont supprimées pour éviter au système d'être indécis lors du filtrage. Le taux de précision inférieur à 100% s'explique donc par le fait que certaines descriptions sont supprimées dans le profil; les descriptions d'exemples du jeu de test ne peuvent donc pas être correctement reconnues lorsqu'ils sont soumis au système. Le taux de non reconnaissance permet ainsi d'évaluer l'efficacité de la modélisation linguistique pour l'évaluation des dépendances élémentaires.

L'impact du mode de constitution du profil On remarque que quel que soit le type de profil utilisé, le mode de constitution du profil, qui fait varier la modélisation linguistique des dépendances, a toujours le même type d'influence. En comparant les résultats de précision des modes 0 et 1 nous arrivons à la conclusion suivante : ne pas prendre en compte le genre et le nombre permet d'augmenter la précision. Ainsi pour les dépendances pertinentes et pour les profils à 20, 60 et 90%, la précision passe respectivement de 37 à 40%, de 46.4 à 50%, et de 49.5 à 53% de précision. En revanche pour le profil à 100%, on perd un point (de 95.7 à 94.6%) en supprimant l'information de genre et nombre. Cet écart doit correspondre à la proprotion de dépendances qui diffèrent seulement par leur genre et nombre, puisque l'on ne demande pas au profil à 100% de reconnaître de nouvelles formes, mais d'identifier celles avec lesquelles il a été construit. En comparant les taux de précision des modes 2 et 1, il apparaît que l'étiquetage sémantique et le recours aux suffixes permet une augmentation substantielle de la précision. Pour les profils à 20, 60 et 90%, l'utilisation des catégories sémantiques et des suffixes permet un gain de 20% de précision en moyenne, en faisant baisser le taux des non reconnaissances. Par contre, il fait parallèlement augmenter le taux d'erreurs. Cette augmentation du taux d'erreurs s'explique par une trop forte capacité de généralisation des catégories sémantiques. Nous l'expliquons également pas une procédure de reconnaissance de suffixes qui donne des résultats assez médiocres sur les suffixes courts (entre 1 et 3 caractères). En comparant les taux de précision des modes 3 et 1, il apparaît que la suppression de l'information de déterminant sur le deuxième nom des dépendances à trois éléments, fait légèrement augmenter la précision (de 1 à 2%) selon les profils. Ce résultat s'explique certainement par le fait que le profil I, définit pour filtrer des syntagmes pour la veille technologique ne cherche pas à retenir forcément les syntagmes d'allure dénominative, dont une des caractéristique est souvent l'absence de déterminant entre la préposition et le nom. Une évaluation à partir des exemples du profil II (alimenté exclusivement avec des termes) permettrait de vérifier ce point.

7.4 Conclusion

Nous avons présenté notre méthode d'apprentissage symbolique des configurations linguistiques des dépendances syntaxiques élémentaires. Le résultat de l'apprentissage est un profil de filtrage défini à partir de règles interprétables linguistiquement : des descriptions linguistiques associées à une action d'émondage syntaxique. Les formes possibles de ces descriptions linguistiques définissent le langage des hypothèses. Ce langage, qui réduit considérablement l'espace de généralisation, permet de définir et d'attribuer une plus ou moins grande valeur descriptive à chaque type de description admise.

L'évaluation de la procédure d'apprentissage montre que le système apprend assez bien avec le langage des hypothèses défini. Les présupposés d'apprentissage sont donc corrects. Rien ne dit cependant s'ils sont les meilleurs possibles. La suppression de certains attributs linguistiques du langage des hypothèses a permis de montrer que l'utilisation d'informations sémantiques apporte un gain de performance important.

Nous allons maintenant présenter les résultats de filtrage à partir des deux profils sur une sous-partie du corpus ARD.

Chapitre 8

Résultats

Nous donnons dans ce chapitre deux types de résultats. Dans un premier temps (en 8.1), nous appliquons le système de filtrage sur le sous-corpus ARD-94 dans lequel 13 839 syntagmes nominaux (10 936 différents) à filtrer ont été identifiés. Nous donnons alors des résultats quantitatifs en termes de nombre de dépendances syntaxiques filtrées et nombre de syntagmes nominaux pertinents retenus. Ensuite (en 8.2), nous montrons les résultats produits à partir d'un petit extait du corpus : groupes nominaux extraits et SNP retenus. Enfin nous donnons les raisons pour lesquelles il n'était pas possible d'évaluer strictement notre système.

8.1 Résultats de filtrage

8.1.1 Application du profil I

Le profil I a été construit selon les 4 modes qui visent à évaluer l'impact des informations linguistiques (voir chapitre 7, 7.2.5) entrant dans la composition des schémas de filtrage des dépendances.

Evaluation des dépendances La table 8.1 montre le nombre de dépendances élémentaires du sous-corpus ARD-94 (au total 16 735) retenues pertinentes, non pertinentes et non reconnues par le profil I en fonction du mode de constitution du

TAB. 8.1 – *Nombre de dépendances élémentaires retenus comme pertinentes, effacées et non prises en compte après filtrage avec le profil I*

Mode	Nbr. Dép. retenues	Nbr. Dép. altérées	Nbre Dép. indéterminées
0	3269 (19.5%) [2197]	3499 (20.9%) [3449]	9967 (59.5%) [9879]
1	3764 (22.5%) [3685]	3983 (23.8%) [3932]	8988 (53.7%) [8898]
2	2637 (15.7%) [2562]	2892 (17.2%) [2849]	11 206 (66.9%) [11 104]
3	4041 (24.1%) [3962]	4213 (25.1%) [4165]	8481 (50.67%) [8388]

TAB. 8.2 – *Résultats de filtrage avec le profil I des SNP du sous-corpus ARD94 en fonction du mode de constitution du profil. Les résultats sont indiqués en nombre de SNP retenus, en taille du lexique des SNP retenus et en proportion de SN éliminés*

Mode	SNP polylexicaux	SN monolexicaux
0	11 175 (80.7%) [8728] (79.8%)	4763 [1844] n.a
1	10 975 (79.3%) [8526] (77.9%)	5159 [1943] n.a
2	11 575 (83.6%) [9006] (82.3%)	4245 [1734] n.a
3	10 988 (79.3%) [8430] (77%)	5262 [1970] n.a

profil. Les chiffres indiqués entre parenthèses indiquent les proportions correspondantes. Les chiffres indiqués entre cochets indiquent la taille du lexique des dépendances retenues, altérées et non identifiées. Les dépendances non reconnues n'avaient pas de description qui leur correspondait dans le profil. On remarque que le mode 1 (information de genre et nombre absente) permet de reconnaître un peu plus de dépendances que le mode 0. Lorsque l'on supprime l'information de catégorie sémantique et de suffixe (mode 2, à comparer avec le mode 1), le nombre de dépendances reconnues, qu'elles soient retenues ou effacées, descend fortement : respectivement pour les dépendances pertinentes et non pertinentes de 3685 à 2562 et de 3932 à 2849. En contrepartie, le nombre de dépendances non identifiées augmente, passant de 8898 à 11 104. Par rapport au mode 1, le mode 3 (qui supprime l'information de déterminant sur le deuxième nom des dépendances à trois éléments), augmente un peu la reconnaissance : de 3685 à 3962, et de 3932 à 4165.

Filtrage des groupes nominaux Si nous considérons maintenant la table 8.2, nous pouvons y lire les résultats de filtrage des SNP avec le profil I, selon les modes de constitution utilisés. Pour les SNP polylexicaux figurent dans la même colonne quatre informations de gauche à droite : le nombre total de SNP retenus, entre parenthèses la proportion correspondante par rapport au nombre total de SN, entre crochets la taille du lexique de SNP, et enfin entre parenthèses, la proportion de SNP différents par rapport au nombre de SN dans le corpus ARD94. Les chiffres sont à interpréter avec précaution. En effet, un groupe nominal aura pu être simplifié, ou scindé en deux voire en trois. Cela a un effet multiplicateur sur la proportion évaluée alors qu'il y a pourtant eu filtrage. Les SN monolexicaux (deuxième colonne) correspondent à des groupes nominaux polylexicaux élagués dont il ne reste plus qu'un mot simple. Nous les considérons comme inexploitables mais ils n'ont pas été comptabilisés pour le calcul de la proportion des SN filtrés. En nous appuyant sur les résultats d'évaluation de l'apprentissage faite au chapitre précent, nous pouvons affirmer sans prendre trop de risques que l'utilisation d'informations sémantiques permet augmenter la couvertue du profil de filtrage. Le chiffre de 83% de SNP retenus (mode 2 sans informations sémantiques) s'interprète alors comme une diminution de la couverture de filtrage. Il ressort de ces résultats que la présence ou l'absence des attributs

TAB. 8.3 – *Résultats de filtrage avec le profil II des SNP du sous-corpus ARD94 en fonction du mode de constitution du profil. Les résultats sont indiqués en nombre de SNP retenus et en taille du lexique des SNP retenus*

Mode	SNP polylexicaux	SN monolexicaux
0	1315 (9.50%) [890] (8.1%)	12142 [1464]
1	1490 (10.7%) [999] (9.1%)	12245 [1497]
2	1333 (9.60%) [897] (8.2%)	12166 [1467]
3	1600 (11.5%) [1077] (9.8%)	12394 [1551]

linguistiques utilisés a un impact important, toutefois il est difficile de déterminer si la cause réside dans les descriptions d'exemples positifs ou négatifs. Pour affiner ces mesures il faudrait prendre en compte pour chaque SNP, le nombre de dépendances pertinentes du SNP validées par le profil et le nombre de dépendances qui ont été élaguées dans le SN pour obtenir le SNP correspondant.

Impact des attributs linguistiques utilisés La table 8.2 permet de constater l'effet qu'a la présence ou l'absence de certains attributs linguistiques. En l'absence de genre et de nombre (mode 1) le profil filtre plus (77.9% de SNP en mode 1 contre 79.8% de SNP en mode 0). En l'absence d'informations sémantiques (mode 2), le profil filtre moins (82.3% de SNP en mode 2 contre 77.9% de SNP en mode 1). En l'absence d'information de déterminant, le profil filtre un peu plus (77% de SNP en mode 3 contre 77.9% de SNP en mode 1). Les écarts observés entre modes sont moins importants que ceux observés lors de l'évaluation de l'apprentissage. Mais comme nous le signalions plus haut, un filtrage plus important ne signifie pas forcément diminution du nombre de SNP finaux, étant donné que des SN coupés en plusieurs SNP augmentent la proportion de SNP finaux plutôt qu'ils ne la diminuent.

8.1.2 Application du profil II

Application du protocole de filtrage 2

Ce protocole n'utilise pas d'échantillon négatif mais seulement un échantillon positif. Les dépendances élémentaires qui n'appartiennent pas à l'échantillon positif sont systématiquement supprimées. La table 8.3 donne les résultats pour le profil II. Comme pour les résultats du profil I, figurent dans la première colonne : le nombre de SNP retenus, la proportion correspondante, le nombre de SNP différents retenus, la proportion correspondante. Il apparaît que les profils I et II ont un pouvoir de filtrage nettement différent : environ 20% des groupes nominaux sont éliminés pour le profil I, tandis que 90% sont éliminés avec le profil II. Pour le profil II, le nombre de SN monolexicaux générés est beaucoup plus important (deuxième colonne du tableau 8.3), étant donné que le profil est plus sévère. Ils correspondent la plupart du temps à des débris d'arbres élagués. Les mêmes différences de filtrage que pour le profil I

s'observent avec les différents modes.

SNP communs aux profils I et II Il y a 420 syntagmes sont qui sont communs aux résultats des deux profils (i.e 5% des syntagmes du profil I et 47% des syntagmes du profil II). Ce chiffre s'explique par le fait qu'il avait été introduit dans le profil I un certain nombre d'exemples positifs tirés du thesaurus EDF, ce dernier étant la seule source d'exemples pour le profil II.

8.2 Un exemple détaillé de résultats

8.2.1 Texte de départ

Le texte reproduit en table 8.4 est un extrait d'ARD que nous avons soumis à notre système. Il est d'abord analysé par *AlethIP*. Ensuite il est présenté à notre chaîne de traitement qui l'enrichit (désambiguïsation sémantique entre autres) puis extrait les groupes nominaux identifiés par *AlethIP*. Ces derniers sont listés par ordre d'apparition plus bas en 8.2.2. Un numéro leur est affublé. Ce même numéro a été reproduit dans le texte de la table 8.4 en indice à la fin des syntagmes identifiés. Ensuite ces syntagmes sont soumis au filtrage avec les profil I et II, constitué avec le mode 0 (voir section 7.2.5). Les résultats sont présentés en 8.2.3 et 8.2.4.

8.2.2 Groupes nominaux extraits des analyses d'*AlethIP*

Seuls les syntagmes complexes fournis par *AlethIP* (au moins deux unités lexicales) **sont collectés**, afin de mettre en évidence au moins une dépendance. Nous n'intervenons pas sur l'analyse d'*AlethIP*. Nous laissons tel quel les décompositions opérées sur les groupes nominaux «maximaux» isolés par l'analyseur. La liste des syntagmes retenus est fournie en table 8.5.

8.2.3 Groupes nominaux retenus par le profil I

Chaque groupe nominal «maximal» est décomposé en dépendances. Chaque dépendance est évaluée, filtrée. Le groupe nominal initial est reconstitué avec les dépendances restantes. Les groupes nominaux résultants de ce filtrage sont listés et commentés en table 8.6.

Les dépendances des syntagmes n° 9, 10, 12, 15, 16, 17, 25, 26, 27 de la table 8.5 ont toutes été supprimées (9 groupes nominaux complets ont été supprimés). Dans la plupart des cas, les dépendances à filtrer étaient déclarées sous leur forme complète dans le profil. Il n'était donc pas nécessaire de poursuivre le parcours du profil pour rechercher des schémas moins spécifiques. Notre exemple, de taille trop limitée, n'a permis de mettre en évidence que deux cas d'extrapolation pour les dépendances négatives et quatre cas pour les dépendances positives. Ainsi la dépendance *indication utile* a été supprimée avec le schéma : *un nom en suffixe* -tion *modifié par l'adjectif*

TAB. 8.4 – *Reproduction du document à filtrer sous sa forme lemmatisé*

DES MESURE DE TEMPERATURE ET DE ROTATION$_1$ AVOIR ETE EFFECTUE PENDANT LE SOUDAGE. DES MESURE DE CONTRAINTE RESIDUEL PAR DIFFERENT METHODE$_2$ AVOIR ETE AUSSI EFFECTUE. UNE SYNTHESE BIBLIOGRAPHIQUE SUR LES MESURE DE CONTRAINTE RESIDUEL PAR DIFFRACTION NEUTRONIQUE$_3$ AVOIR ETE REDIGE. LES MESURE PAR DIFFRACTION X SUR LES ASSEMBLAGE SOUDE PAR FRICTION$_4$ AVOIR ETE REALISE.
EFFET DE LES CONTRAINTE RESIDUEL SUR LA TENUE MECANIQUE DE LES COMPOSANTS$_5$.
LES ASSEMBLAGE HOMOGENE ET HETEROGENE$_6$ AVOIR ETE FABRIQUE. LE TRAITEMENT THERMIQUE$_7$ ET LES CARACTERISATION DE MATERIAU$_8$ AVOIR ETE AUSSI EFFECTUE. UN PROGRAMME DE ESSAI ET DE CALCUL$_9$ AVOIR ETE DEFINI.
OBJECTIF ET PRINCIPAL ETAPE DE LA ANNEE 1995$_{10}$. MESURE DE CONTRAINTE RESIDUELLES$_{11}$. ON UTILISER EVENTUELLEMENT DIFFERENT METHODE$_{12}$. CES ETUDE SE FAIRE EN RELATION AVEC LE CEA DANS LE CADRE DE LES FICHE BIPARTITE 2435$_{13}$. ON EXPLOITER LES RESULTAT FOURNI PAR LES MESURE SUR DES PLAQUE REVETU EFFECTUE EN 1994$_{14}$.
LA COMPARAISON DE LES CONTRAINTE OBTENU PAR DIFFERENT METHODE SUR DES MAQUETTE IDENTIQUE$_{15}$ FOURNIR DES INDICATION UTILE SUR LA FIABILITE DE CES METHODE$_{16}$. ON DEVOIR POUVOIR DETERMINER NOTAMMENT SI LES MESURE AINSI OBTENU$_{17}$ POUVOIR SERVIR DE REFERENCE A LES CALCUL. ON POURSUIVRE LA MISE AU POINT DE UNE METHODE DE MESURE DE LES CONTRAINTE RESIDUEL DANS LA EPAISSEUR DE LES COMPOSANT PAR DIFFRACTION NEUTRONIQUE$_{18}$. LES MESURE ETRE EFFECTUE A LE LLB SACLAY$_{19}$. DES MESURE ETRE EFFECTUE SUR DES MAQUETTE PLAN BIMETALLIQUE$_{20}$ ET SUR LES ASSEMBLAGE SOUDE PAR FRICTION$_{21}$ EN PEAU EXTERNE$_{22}$ ET EN PEAU INTERNE APRES DECOUPE$_{23}$.
ON EFFECTUER 2 ESSAI DE FLEXION AVEC DES JOINT SOUDE PAR FRICTION$_{24}$. ON ANALYSER LES RESULTAT ET ON FAIRE DES CALCUL CORRESPONDANT$_{25}$. OBJECTIF ULTERIEUR$_{26}$: L'OBJECTIF ESSENTIEL$_{27}$ ETRE LA ETUDE DE LE ROLE DE LES CONTRAINTE RESIDUEL SUR DES PHENOMENE$_{28}$ TEL QUE LA RUPTURE, LA FATIGUE ET LA CORROSION SOUS CONTRAINTE$_{28}$. IL FALLOIR POUR CELA DISPOSER DE OUTIL NUMERIQUE$_{30}$ PERMETTANT DE INTRODUIRE DES CHAMP DE CONTRAINTE RESIDUEL COMPLET$_{31}$ A PARTIR DE QUELQUES MESURE PONCTUEL$_{32}$, CES CHAMP POUVANT ETRE SUPERPOSE ALORS A LES CHAMP EXTERNE LORS DE CALCUL A LES ELEMENT FINI$_{33}$. CES OUTIL ETRE DEVELOPPE DANS LE CADRE DE UNE THESE. POUR ATTEINDRE CE OBJECTIF, IL APPARAITRE DE PLUS EN PLUS NECESSAIRE DE AUGMENTER LES COMPETENCE EN MESURE DE CONTRAINTE$_{34}$, NOTAMMENT EN COUPLANT LES APPROCHE EXPERIMENTAL ET NUMERIQUE$_{35}$.

TAB. 8.5 – *Liste des groupes nominaux identifiés par* AlethIP

1	MESURE DE TEMPERATURE ET DE ROTATION
2	MESURE DE CONTRAINTE RESIDUEL PAR DIFFERENT METHODE
3	SYNTHESE BIBLIOGRAPHIQUE SUR LES MESURE DE CONTRAINTE RESIDUEL PAR DIFFRACTION NEUTRONIQUE
4	MESURE PAR DIFFRACTION X SUR LES ASSEMBLAGE SOUDE PAR FRICTION
5	EFFET DE LES CONTRAINTE RESIDUEL SUR LA TENUE MECANIQUE DE LES COMPOSANTS
6	ASSEMBLAGE HOMOGENE ET HETEROGENE
7	TRAITEMENT THERMIQUE
8	CARACTERISATION DE MATERIAU
9	PROGRAMME DE ESSAI ET DE CALCUL
10	OBJECTIF ET PRINCIPAL ETAPE DE LA ANNEE 1995
11	MESURE DE CONTRAINTE RESIDUEL
12	EVENTUELLEMENT DIFFERENT METHODE
13	FICHE BIPARTITE 2435
14	RESULTAT FOURNI PAR LES MESURE SUR DES PLAQUE REVETU EFFECTUE EN 1994
15	COMPARAISON DE LES CONTRAINTE OBTENU PAR DIFFERENT METHODE SUR DES MAQUETTE IDENTIQUE
16	INDICATION UTILE SUR LA FIABILITE DE CES METHODE
17	MESURE AINSI OBTENU
18	METHODE DE MESURE DE LES CONTRAINTE RESIDUEL DANS LA EPAISSEUR DE LES COMPOSANT PAR DIFFRACTION NEUTRONIQUE
19	LLB SACLAY
20	MAQUETTE PLAN BIMETALLQUE
21	ASSEMBLAGE SOUDE PAR FRICTION
22	PEAU EXTERNE
23	PEAU INTERNE APRES DECOUPE
24	2 ESSAI DE FLEXION AVEC DES JOINT SOUDE PAR FRICTION
25	CALCUL CORRESPONDANT
26	OBJECTIF ULTERIEUR
27	OBJECTIF ESSENTIEL
28	ROLE DE LES CONTRAINTE RESIDUEL SUR DES PHENOMENE
29	CORROSION SOUS CONTRAINTE
30	OUTIL NUMERIQUE
31	CHAMP DE CONTRAINTE RESIDUEL COMPLET
32	MESURE PONCTUEL
33	CHAMP EXTERNE LORS DE CALCUL A LES ELEMENT FINI
34	COMPETENCE EN MESURE DE CONTRAINTE
35	APPROCHE EXPERIMENTAL ET NUMERIQUE

TAB. 8.6 – *Liste des SNP retenus avec le profil I*

MESURE DE TEMPERATURE MESURE DE ROTATION Ces deux SNP retenus par le profil I résultent de la distribution de la coordination du syntagme n°1 (voir table précédente 8.5). Aucune dépendance n'a été filtrée.
MESURE DE CONTRAINTE RESIDUEL Ce SNP résulte de la supression des deux dépendances *mesure par méthode* et *différentes méthodes* dans le SN n°2.
MESURE DE CONTRAINTE RESIDUEL PAR DIFFRACTION NEUTRONIQUE Ce SNP provient du syntagme n°3 dans lequel la dépendance *synthèse bibliographique* a été supprimée.
MESURE PAR DIFFRACTION X SUR LES ASSEMBLAGE SOUDE PAR FRICTION Syntagme n°4 à l'identique
EFFET DE LES CONTRAINTE RESIDUEL SUR LA TENUE MECANIQUE DE LES COMPOSANTS Syntagme n°5 à l'identique
ASSEMBLAGE HOMOGENE ASSEMBLAGE HETEROGENE Distribution de la coordination dans le syntagme n°6
TRAITEMENT THERMIQUE Syntagme n°7 à l'identique
CARACTERISATION DE MATERIAU Syntagme n°8 à l'identique
MESURE DE CONTRAINTE RESIDUEL Syntagme n°11 à l'identique
FICHE BIPARTITE 2435 Syntagme n°13 à l'identique
PLAQUE REVETU Provient du syntagme n°14 dans lequel seule la dépendance *plaque revêtue* a été retenue.
CONTRAINTE RESIDUEL EPAISSEUR DE LES COMPOSANT DIFFRACTION NEUTRONIQUE Ces SNP proviennent du syntagme n°18 dans lequel la dépendance *méthode de mesure* a été effacée.
MAQUETTE PLAN BIMETALLIQUE Syntagme n°20 à l'identique
ASSEMBLAGE SOUDE PAR FRICTION Syntagme n°21 à l'identique
PEAU EXTERNE Syntagme n°22 à l'identique
PEAU INTERNE APRES DECOUPE Syntagme n°23 à l'identique
ESSAI DE FLEXION JOINT SOUDE PAR FRICTION Coupure du syntagme n°24 au niveau de la préposition *avec* pour la dépendance *essai avec joint*.
CONTRAINTE RESIDUEL SUR DES PHENOMENE Effacement de la dépendance *rôle de contraintes* dans le syntagme n°28. Ce SNP est étrange. Du fait de l'article indéfini, il aurait peut-être fallu déclaré *contrainte sur (art. indéfini) phénomène* comme non pertinent.
CORROSION SOUS CONTRAINTE Syntagme n°29 à l'identique
OUTIL NUMERIQUE Syntagme n°30 à l'identique
CHAMP DE CONTRAINTE RESIDUEL COMPLET Syntagme n°31 à l'identique
CHAMP EXTERNE CALCUL A LES ELEMENT FINI Coupure automatique au niveau de la préposition complexe *lors de* dans le syntagme n°33.
MESURE DE CONTRAINTE Effacement de la dépendance *compétence en mesure* dans le syntagme n°34.
APPROCHE EXPERIMENTALE APPROCHE NUMERIQUE Distribution du syntagme coordonné n°35

utile, et la dépendance *mesures obtenues* a été filtrée avec le schéma *un nom de processus ou d'activité modifié par l'adjectif obtenu*. De plus *maquette bimétallique* a été retenue par le schéma *un nom d'artefact modifié par un adjectif en -ique*, la dépendance *approche expérimentale* a été retenue par le schéma *approche modifié par un adjectif déclarant une propriété générale*, la dépendance *champ externe* a été retenue par le schéma *un nom de lieu modifié par l'adjectif externe* (en l'occurrence, *champ* est ici plutôt un nom de phénomène), la dépendance *assemblage par friction* a été retenue par le schéma *un nom se teminant par -age modifié par un nom de processus*. Enfin le nombre de dépendances non reconnues est de 32, pour 25 dépendances identifiées comme pertinentes et 12 dépendances identifiées comme non pertinentes.

8.2.4 Groupes nominaux retenus par le profil II

Le profil II filtre à l'évidence beaucoup plus sévèrement que le profil I les syntagmes qu'on lui présente. Des 35 syntagmes fournis en entrée, il n'en reste que 10. Ce profil a retenu deux syntagmes qui avaient été écartés par le profil I. Il s'agit de *programme de calcul* et *programme d'essai*. Ce sont deux descripteurs du thesaurus EDF. Au total, 56 dépendances ont été identifiées comme non pertinentes (c'est-à-dire absentes du profil), et 13 dépendances ont été reconnues comme pertinentes.

MESURE DE TEMPERATURE
MESURE DE CONTRAINTE
CARACTERISATION DE MATERIAU
CORROSION SOUS CONTRAINTE
ESSAI DE FLEXION
PROGRAMME DE CALCUL
PROGRAMME DE ESSAI
METHODE DE MESURE
ELEMENT FINI
CHAMP DE CONTRAINTE

Ces résultats s'expliquent par la petite taille du profil II constitué avec 3900 dépendances syntaxiques tirées du thesaurus EDF. Aussi nous pensons que le deuxième protocole de filtrage, exploité avec le profil II, est peut-être trop puissant. Lorsque les descriptions linguistiques qui alimentent le profil sont exclusiment positives, il écarte en effet toutes les dépendances qui n'appartiennent pas au profil. Cette solution trop tranchante ne se veut pas idéale. Mais elle permet de commencer à travailler sur des documents quand bien même l'opérateur ne disposerait que de descriptions positives ou négatives, tirées d'une ressource existante. En ce sens c'est une solution qui permet d'aider à la construction de descriptions négatives (respectivement positives) avec un profil de départ que l'on pourrait qualifier d'incomplet lorsqu'il est exclusivement composé de description positives (respectivement négatives).

8.2.5 Impression générale

Nous avons largement participé à la constitution de profil I. Nous sommes donc en mesure de faire un certain nombre de remarques sur les résultats produits avec le profil I. Dans l'ensemble nous pouvons affirmer que les résultats sont satisfaisants.

Toutefois une lecture détaillée des résultats nous a permis d'identifier un certain nombre de problèmes :

1° Des dépendances fictives sont produites. La cause de ce phénomène réside dans les fréquentes incohérences des structures syntaxiques produites par *AlethIP*. Il serait tout à fait souhaitable d'expérimenter un autre type d'analyseur syntaxique. Il faudrait surtout que cet analyseur exploite une formalisme syntaxique plus rigoureux, plus régulier que celui d'*AlethIP*. L'algorithme d'extraction des dépendances syntaxiques gagnererait en généricité et la qualité des dépendances aussi.

2° Notre traitement des phénomènes de coordination est insuffisant. Ceci est volontaire : il ne servait à rien de prendre en compte les phénomènes complexes de coordination étant donné que l'analyseur syntaxique n'est pas capable de les traiter correctement.

3° De petits «trous descriptifs» se laissent observer dans le profil de filtrage. On remarque par exemple que certains SNP ne sont pas pertinents par rapport aux critères de définition du profil I. Après vérification, on s'aperçoit qu'il n'y a pas de description dans le profil pour traiter ce type de dépendance. Ce problème met en évidence une certaine difficulté à construire des profils exhaustifs, couvrant l'intégralité des phénomènes lexico-syntaxiques pour la tache demandée au profil.

4° De rares phénomènes d'extrapolations abusives. Ils sont toujours dus à une incomplétude du profil du filtrage : une description positive peut parfois entraîner une généralisation abusive si une description négative correspondante ne vient pas limiter la généralisation. Par exemple si on déclare négatif *document intéressant*, la dépendance *document passionnant* sera aussi évaluée négative (dans les deux cas on a un adjectif d'évaluation subjective). Or il se peut que *document passionnant* ne doive pas être éliminé. Pour cela, il faut l'ajouter aux descriptions positives de l'échantillon. On bloque alors la généralisation au niveau du trait sémantique (et *a fortiori* au niveau du suffixe).

5° Longueur des SNP. Les SNP ne sont pas nécessairement exploitables comme des unités minimales porteuses d'information, ni comme des dénominations motivées. Il serait toutefois possible pour les besoin d'une application de redécomposer les SNP (comme *Lexter* le fait pour ses groupes nominaux maximaux, voir figure 6.1) en SNP plus courts.

8.3 Difficulté d'évaluer les extracteurs de groupes nominaux

Nous n'avons jusqu'à présent présenté que des résultats. L'évaluation rigoureuse de ces résultats n'a pu être effectuée. Une telle évaluation ne peut être faite qu'en

demandant à la personne qui a construit le profil de filtrage de contrôler si chaque SNP retenu lui paraît conforme à ce qu'elle attend. Nous n'avons pas trouvé le temps pour cette vérification. Mais au-delà de ce problème pratique, il nous semble bon de préciser qu'une évaluation de notre système rencontre des difficultés objectives propres au domaine de l'extraction de groupes nominaux.

8.3.1 Difficulté technique

Il y a tout d'abord une difficulté technique ou pratique. Celle-ci a été vérifiée à la DER d'EDF pour l'évaluation d'extracteurs de groupes nominaux et de candidats termes, comme *AlethIP* et *Lexter* [Bou94b]. Par exemple, à la DER on ne trouve plus d'experts-terminologues pour faire ce genre de travail de validation de candidats termes. Ils sont partis à la retraite avec leurs connaissances. Ils ne sont pas remplacés. La raison est peut-être que les thesaurus papier n'ont plus d'avenir. Il faudra pourtant trouver de nouveaux experts pour gérer et vérifier les ressources terminologiques électroniques produites à partir des textes en circulation. Si les terminologues ne sont pas disponibles, les concepteurs en sont réduits à évaluer eux-même leurs outils, ce qui n'est pas souhaitable.

Notre méthode souligne d'autant ce problème qu'elle requiert la compétence du spécialiste avant les traitements pour définir un profil, et non pas seulement après pour évaluer les sorties, comme c'est en général le cas pour les extracteurs de groupes nominaux. Le principe de construction d'un profil pour accomplir une tache donnée oblige à avoir une idée très précise de l'objectif à atteindre avant même d'analyser les documents.

8.3.2 Difficulté méthodologique

Il y a un obstacle méthodologique qui se dresse devant l'évaluation des extracteurs de groupes nominaux. Faut-il comparer les sorties de différents extracteurs ou faut-il évaluer isolément les sorties des différents extracteurs? Dans ce dernier cas, comment s'y prend-t-on ?

Evaluation comparée Un exemple d'évaluation comparée est celle qui a été appliquée à des étiqueteurs morphologiques dans le cadre du projet GRACE [ABM+95], en définissant un corpus étalon. Ainsi, un corpus de référence est balisé avec un certain jeu d'étiquettes. On demande alors aux différents étiqueteurs à évaluer d'analyser ce document, en ayant établi au préalable une équivalence entre le jeu d'étiquettes de référence et les jeux d'étiquettes utilisés par les différents outils. On est alors en mesure d'évaluer les performances des différents systèmes.

L'approche comparative nous semble par ailleurs difficile à porter pour notre système étant donné qu'il faudrait prendre en compte la notion de profil. En effet, les groupes nominaux extraits varient d'un profil à l'autre (grammaire d'extraction induite par un échantillon d'apprentissage). Pour cette raison la comparaison avec

un autre extracteur (dont la grammaire d'extraction est généralement figée) est problématique. Cela demanderait la construction d'un profil d'extraction équivalent à un autre extracteur de groupe nominaux.

Evaluation isolée L'autre alternative est d'évaluer le système isolément. Là encore, il faut pouvoir construire des résultats de référence sur un corpus de référence afin de vérifier si les résultats produits sont conformes aux résultats attendus. Comme le suggère T. Saracevic [**?**], une évaluation peut aussi prendre en compte le degré de satisfaction de l'utilisateur final. Nous pensons que c'est ce qui devrait être envisagé dans un premier temps pour l'évaluation de notre système.

8.4 Evolution du nombre de dépendances élémentaires

L'évolution du lexique au fil des ans donne une indication sur l'évolution des activités mentionnées dans les textes scientifiques et techniques. Sans faire d'étude de contenu, nous avons par exemple remarqué sur notre corpus ARD une forte croissante du lexique nominal de 1984 à 1995 (voir figure 3.4 au chapitre 3). Un autre point de vue intéressant pour l'analyse diachronique est de considérer non pas le lexique des mots simples mais le lexique des dépendances syntaxiques élémentaires. Ces dernières sont en effet des objets plus déterminés que des mots simples et servent de «pièces de construction» pour les groupes nominaux. Certaines dépendances du corpus apparaîssent de très nombreuses fois (plus de 5000 fois, comme `circuit primaire`) et entrent dans la constitution de nombreux groupes nominaux différents (c'est aussi le cas de dépendances comme `but de (l') action` ou `objectif principal` mais qui entrent dans la construction de groupes nominaux peu différents les uns des autres).

Nous présentons à présent une brève étude quantitative diachronique du lexique des dépendances élémentaires dans notre corpus. Bien entendu les résultats sont spécifiques au corpus EDF utilisé – **ils n'intéresseront que les chercheurs du groupe ISI familiarisés avec les textes d'ARD**.

8.4.1 Dépendances conservées, abandonnées, renouvelées de 1993 à 1994

On a calculé la cardinalité de l'ensemble résultant de l'intersection des dépendances de 1993 et de 1994 en faisant varier la fréquence d'apparition des dépendances en corpus. On a choisi ces deux années car les sous-corpus correspondants contiennent à peu près le même nombre de dépendances (16 084 et 16 736). Les tableaux 8.8 et 8.7 montrent les résultats et la déduction du nombre de dépendances de 1993 abandonnées en 1994 et du nombre de dépendances de 1994 qui n'apparaissent pas en 1993. Les effectifs sont donnés indépendamment du nombre d'occurrences de chaque dépendance. On note que le nombre de dépendances communes aux deux sous-corpus sont peu importantes (5381, soit 33%), pour des textes abordant normalement des thèmes identiques et relatant l'évolution des mêmes projets d'une année sur l'autre.

TAB. 8.7 – *Proportion des dépendances élémentaires communes et spécifiques aux années 1993 et 1994*

Freq.	Proportion de Dép. abandonnées en 93	Proportion de Dép. 93 communes avec 94	Proportion de Dép. 94 communes avec 93	Proportion de Dép. renouvellées en 94
>= 1	66,54%	33,46%	32,15%	67,85%
>= 2	66,15%	33,85%	34,46%	65,54%
>= 3	58,53%	41,47%	39,98%	60,02%
>= 4	55,82%	44,18%	43,97%	56,03%
>= 5	51,17%	48,83%	47,56%	52,44%
>= 6	49,39%	50,61%	68,44%	77,08%
>= 7	44,98%	55,02%	52,82%	47,18%
>= 8	41,67%	58,33%	58,33%	41,67%
>= 9	48,42%	51,58%	59,39%	40,61%

Lorsque la fréquence des dépendances augmente, la proportion de dépendances communes augmente. C'est certainement le signe que les deux sous-corpus convergent vers les mêmes domaines d'activités.

L'autre point intéressant est que le nombre de dépendances abandonnées de 1993 à 1994 est supérieur au nombre de dépendances communes (66% contre 32%) pour une fréquence supérieure ou égale à un. Ces proportions tendent à s'égaliser avec l'augmentation des fréquences : il y a en moyenne autant de dépendances abandonnées que conservées. Les proportions des nouvelles dépendances en 1994 sont legèrement supérieures à celles des dépendances abandonnées en 1993, ce qui peut s'expliquer par la croissance du lexique d'année en année.

Ces chiffres laissent supposer que les phénomènes de variation lexico-syntaxique sont très importants d'une année sur l'autre.

8.4.2 Dépendances abandonnées, conservées, renouvelées de 1985 à 1995

Nous avons répété la même étude sur une plus vaste échelle de temps. Le tableau 8.9 montre les effectifs et les proportions des dépendances communes, abandonnées et renouvelées des années 1985 à 1995. Le tableau 8.10 montre les mêmes résultats, mais les intersections ont été réalisées sur les dépendances non enrichies, c'est-à-dire sans aucun trait linguistique attaché aux graphies. Le but était de supprimer les formes qui différaient par le nombre, ou la catégorie sémantique.

On observe que la proportion de dépendances communes ne cesse de croître de 1985 à 1995. Elle passe ainsi de 18.73% à 39.38%. Lorsque les dépendances sont appauvries, les proportions augmentent d'environ 5%. de 1985 à 1995, on passe ainsi

TAB. 8.8 – *Effectifs des lexiques de dépendances élémentaires communs et spécifiques aux années 1993 et 1994*

Fréq.	Dép. en 1993	Dép. de 93 non réutilisées en 1994	Dép. communes à 93 et 94	Apport de nouvelles Dép. en 94	Dép. en 1994
>= 1	16084	10703	5381	11355	16736
>= 2	4677	3094	1583	3011	4594
>= 3	1818	1064	754	1132	1886
>= 4	1073	599	474	604	1078
>= 5	600	307	293	323	616
>= 6	407	201	206	232	301
>= 7	289	130	159	142	301
>= 8	228	95	133	95	228
>= 9	190	92	98	67	165

de 23.64% à 45% de dépendances communes sur les intervalles annuels considérés. Ces résultats s'expliquent par la forte croissante du lexique et donc du lexique commun. Là encore, cela laisse supposer d'importants phénomènes de variation lexico-syntaxique au sein des mêmes domaines d'activités abordés dans le corpus.

TAB. 8.9 – *Effecttfs et proportions des dépendances abandonnées, conservées, et renouvelées d'une année sur l'autre entre 1985 et 1995*

Dép. ...	1985-88	1988-90	1990-91
Abandonnées	3737 (81,27%)	4546 (73,97%)	6330 (65,86%)
Communes	861 (18,73%)	1600 (26,03%)	3282 (34,14%)
Nouvelles	5285 (86%)	8012 (83,35%)	8926 (73,12%)
Dép. ...	1991-93	1993-94	1994-95
Abandonnées	9532 (78,08%)	10703 (66,54%)	10146 (60,62%)
Communes	2676 (21,92%)	5381 (33,46%)	6590 (39,38%)
Nouvelles	13408 (83,36%)	11355 (67,85%)	18626 (73,87%)

TAB. 8.10 – *Effectifs et proportions des dépendances abandonnées, conservées, et renouvelées d'une année sur l'autre entre 1985 et 1995. Les dépendances sont décrites sans les traits attachés aux graphies*

Dép. ...	1985-88	1988-90	1990-91
Abandonnées	3511 (76,36)	4231 (68,84)	5823 (60,58)
Communes	1087 (23,64)	1915 (31,16)	3789 (39,42)
Nouvelles	5059 (82,31)	7697 (80,08)	8419 (68,96)
Dép. ...	1991-93	1993-94	1994-95
Abandonnées	8797 (72,06)	9992 (62,12)	9203 (54,99)
Communes	3411 (27,94)	6092 (37,88)	7533 (45,01)
Nouvelles	12673 (78,79)	10644(63,60)	17683 (70,13)

Conclusion

Bilan

L'objectif de la thèse était de proposer au groupe SID-ISI un système de filtrage de groupes nominaux s'appuyant sur une représentation linguistique. Les caractéristiques du système de filtrage que nous avons élaboré sont les suivantes :

1° il doit être calibré pour un champ d'investigation déterminé (domaine d'activité, type de document) avant d'être exploité,

2° il demande à l'opérateur (documentaliste, terminographe, ...) une participation active pour la mise au point des filtres, via la définition d'échantillons d'apprentissage.

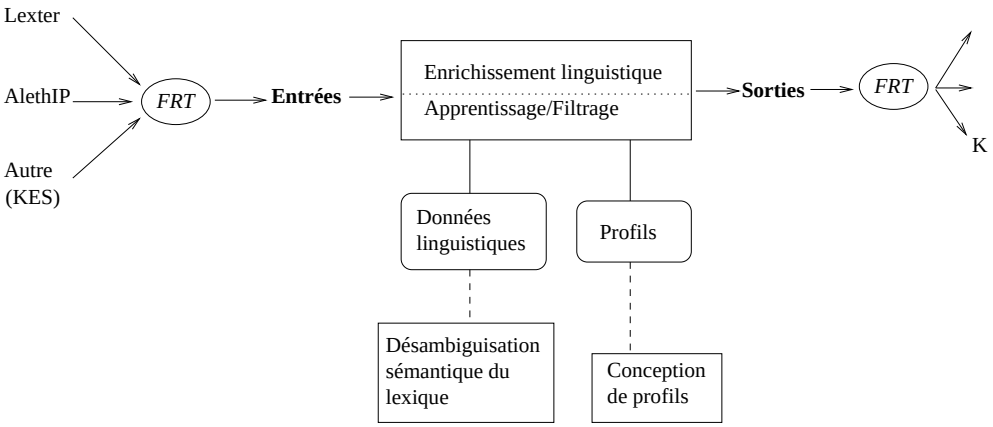
3° l'évaluation de la pertinence des syntagmes nominaux repose sur une description syntaxico-sémantique de leurs dépendances syntaxiques élémentaires.

Intégration du prototype au sein des outils de SID

Notre prototype fournit des résultats exploitables pour le département SID de la DER. Les applications qui pourraient en tirer parti sont : la mise à jour du thesaurus (le profil II a été conçu dans ce but), l'indexation libre de groupes nominaux contrôlée par des profils, la veille technologique (le profil I a été conçu pour cela). Il est également possible d'exporter les résultats vers l'environnement KES pour une manipulation et une exploitation plus aisée des résultats. KES pourrait aussi être utilisé pour aider à la mise au point des échantillons d'apprentissage. La nouvelle version de KES [1] permettra d'intégrer de nouveaux modules. Il est ainsi envisageable d'intégrer notre prototype sous la forme d'un module connecté à KES. La figure 8.1 montre la place que pourrait prendre le module de filtrage. L'application *FRT* de transduction d'arbres [HHPB+97] a pour fonction de normaliser les entrées-sorties afin de tirer parti d'arbres d'analyses autres que ceux d'*AlethIP*, et éventuellement d'exporter dans des formats divers les résultats du filtrage vers d'autres types d'applications (KES par exemple). Notre contribution logicielle est adaptée à un contexte d'utilisation industrielle. Elle montre que les choix linguistiques et méthodologiques que nous avons effectués sont implémentables et donnent des résultats. Du strict point

1. Cette version est en cours de réécriture en langage Tk/Tcl v.8

FIG. 8.1 – *Positionnement de notre chaine de traitement*



de vue des performances et de l'optimisation, des améliorations seraient nécessaires. Nous donnons en annexe D quelques indications sur ce point.

Hypothèses de travail : un changement de cap décisif

Il y a un décalage très important entre les hypothèses que nous avions formulées au commencement de la thèse et les nouveaux points de vue que nous avons adoptés au cours de son déroulement. Après environ deux ans de travail, nous avons déplacé le problème qui nous était soumis et redéfini de nouveaux objectifs.

Au tout début de ce travail, nous nous situions dans une problématique de mise à jour de thesaurus et d'indexation automatique. Il s'agissait d'améliorer les résultats de l'indexation automatique en assurant la mise à jour du thesaurus EDF pour augmenter sa couverture documentaire. Il s'agissait aussi de donner aux termes du thesaurus une représentation linguistique. Car nous faisions l'hypothèse, dans le prolongement de notre travail de DEA, qu'il était possible d'extraire des dénominations complexes motivées sur la base de régularités syntaxiques et sémantiques (des patrons). L'état de l'art dans le domaine de l'extraction de candidats termes montrait que l'exploitation de la seule syntaxe était insuffisante pour isoler des dénominations complexes motivées. Pour cette raison nous souhaitions introduire des informations sémantiques pour décrire les syntagmes.

Or après deux d'expériences menées sur l'enrichissement linguistique de groupes nominaux [NHM96b, NHM96a, Nau96], nous avons dû nous rendre à l'évidence : il n'est pas possible de figer dans une grammaire une représentation linguistique de ce que l'on peut appeler des candidats termes, ou candidats pour la mise à jour d'un thesaurus. Du même coup, cela remettait en cause l'efficacité du processus d'indexation automatique : il n'était plus envisageable de mettre à jour continuellement le thesaurus. L'indexation automatique était mise en difficulté par l'inévitable incomplétude

du vocabulaire contrôlé face à la masse textuelle à prendre en considération.

L'évolution naturelle de cette problématique a été l'adoption d'un point de vue plus relativiste. Puisqu'il apparaissait que cela n'était pas possible, nous avons renoncé à faire assumer à la machine la reconnaissance des candidats termes. Il fallait donc que ce soit l'utilisateur qui juge lui-même ce qu'il considère comme des dénominations, ou plus généralement comme des groupes nominaux pertinents par rapport à ses centres d'intérêt [Nau97]. Nous avons alors exploité la description linguistique des groupes nominaux pour distinguer entre des syntagmes jugés pertinents et des syntagmes jugés non pertinents et pour construire une modélisation de cette distinction.

L'apprentissage s'est alors présenté comme la meilleure solution pour construire des grammaires d'extractions non figées, adaptées à l'évolution des documents et à la tâche demandée à l'application d'extraction. Ainsi nous avons pris une direction quasiment opposée à celle du début, traduite par l'hypothèse suivante : on ne peut avoir de réprésentation linguistique figée d'une terminologie, un modèle descriptif linguistique ne peut pas isoler une réalité terminologique.

Limitations

Une des questions de fond posée dans ce travail était : «Cela vaut-il la peine d'avoir recours à des étiquettes sémantiques ?». Au vu des différences de résultats d'apprentissage et de filtrage avec et sans informations sémantiques, la réponse est : oui, cela vaut la peine car l'information sémantique apporte un gain effectif en augmentant la couverture des profils. Mais si l'on considère l'ensemble de la chaîne de traitement, la réponse est plus mitigée.

En effet, nous ne devons pas oublier l'investissement que représente la catégorisation sémantique. Une fois que les règles de désambiguïsation sont écrites, elles sont réutilisables. Mais pour traiter de nouveaux corpus, il faut vérifier leur comportment et mettre à jour le lexique sémantique correspondant. Nous n'avons pas testé nos règles sur des corpus différents du corpus ARD, nous ne sommes donc pas en mesure de donner une idée du comportement de ces règles sur d'autres textes. Bien que le temps passé à définir les règles de désambiguïsation doive normalement décroître au fil de l'accumulation de contextes-solutions, bien que ce recensement cumulatif doive conduire à une certaine généricité des règles, notre système de catégorisation sémantique n'est pas adapté à la mise à jour constante des règles en raison de la nécessité de les ordonner selon leur portée . Deux solutions se présentent alors :

1. La stratégie de désambiguïsation et le principe d'écriture manuelle des règles sont conservées. Dans ce cas pour pallier la difficulté d'ordonner les règles écrites d'après leur portée (de la plus spécifique à la plus générale), il faut adpoter une technologie d'automate ou de transducteur à états finis. La compilation des expressions régulières sous la forme d'un automate résout le problème de l'ordre de l'application des règles.

2. Conserver le système dans son état actuel et l'utiliser comme une aide à l'étiquetage manuel pour des corpus de gros volume. Dans ce cas, une méthode de catégorisation sémantique basée sur un apprentissage doit être adaptée pour exploiter les corpus de références correctement annotés. Par exemple la méthode basée sur le principe de correction d'erreur [Bri94] dont le code est disponible.

Tests supplémentaires, évaluation Notre protocole expérimental n'a pas permis de distinguer les gains obtenus par l'utilisation des suffixes à l'exclusion des catégories sémantiques et inversement. Nous avons toujours combiné les deux informations. Il conviendrait de faire de nouveaux essais pour déterminer les gains propres à ces deux types d'information. Au bénéfice des suffixes, les résultats pourraient nous orienter vers une piste d'analyse morphologique plus approfondie et possiblement moins coûteuse que la catégorisation sémantique.

De plus, nous n'avons pas effectué d'évaluation des syntagmes nominaux pertinents issus du filtrage. Comme nous l'avons déjà signalé, une telle évaluation demande la participation d'un expert voire de l'expert qui a préparé le profil de filtrage. Le recours à des profils de filtrage nous a conduit à faire participer l'expert à la définition de ce qu'il recherche lui-même, l'amenant à expliciter ce qui est pertinent pour lui. Nous avons ainsi abandonné l'idée de grammaire d'extraction définie *a priori* par un linguiste-informaticien, pour celle de profil de filtrage, accessible à un groupe d'utilisateurs plus large.

Choix des présupposés d'apprentissage Le langage des hypothèses qui contrôle la combinatoire des attributs linguistiques à généraliser pose un problème dans sa définition. La restriction des attributs à combiner ne doit pas être arbitraire. Elle doit reposer sur une connaissance linguistique des phénomènes. Mais on n'est jamais tout à fait sûr, lorsque l'on restreint certaines combinaisons d'attributs, que l'on n'est pas en train de créer un point aveugle dans la description des objets à apprendre. Bien que leurs définitions résultent de nombreux essais et expérimentations, les langages des hypothèses que nous avons proposés sont discutables. Le nombre d'attributs linguistiques impliqués est important mais l'ordre dans lequel ils apparaissent l'est tout aussi. Nous aurions pu choisir d'autres paramètres.

Constitution d'échantillons à partir de documents entiers Nous n'avons pas donné de résultats de filtrage à partir d'un profil construit sur l'intersection de deux documents différents. Outre que cette approche permettrait de mettre à l'épreuve le pouvoir séparateur des profils (en faisant par exemple l'intersection de documents de droit avec des documents sur la chimie), elle permettrait aussi d'envisager une méthode de catégorisation de texte exploitant des informations syntaxiques et sémantiques. En ce sens, nous reprenons la suggestion d' I. Moulinier [Mou96] proposant l'utilisation d'informations sémantiques dans son système de catégorisation de textes à apprentissage symbolique.

Des syntagmes polylexicaux seulement Soulignons le encore, notre système, par construction, n'est pas capable de filtrer des noms simples, étant donné qu'il s'appuie sur des dépendances syntaxiques. Pour filtrer des noms simples, il faudrait étendre l'analyse des dépendances élémentaires à toute la phrase. On trouverait alors des dépendances syntaxiques constituées avec des verbes, des adverbes, des conjonctions, etc. Cette extension, combinée à la prise en compte plus vaste du contexte des dépendances dans l'arbre conduirait à dépasser la limitation d'une évaluation isolée des dépendances syntaxiques.

Perspectives

Du profil de filtrage à une grammaire de groupes nominaux Un profil de filtrage cherche à prédire l'acceptabilité de nouvelles dépendances syntaxiques à partir de la description de dépendances observées. Cette démarche, qui consiste à formaliser le possible à partir de l'observé, s'apparente à l'induction d'une grammaire des groupes nominaux observés en corpus de spécialité. Les profils sont toutefois une forme très appauvrie de grammaire : les dépendances sont décrites isolément sans leurs contextes.

Mais nos besoins en matière d'analyse morpho-syntaxique étaient assez limités : nous avons travaillé avec des formes lemmatisées et catégorisées mises en relation de dépendance syntaxique. Si fallait aller plus avant dans cette description des formes acceptées et des formes rejetées pour construire une grammaire de syntagmes pertinents, il faudrait approfondir l'analyse morpho-syntaxique. Nous pensons par exemple à une analyse morphologique poussée (recherche des préfixes, des suffixes, des affixes savants, de l'étymologie) qui pourrait fournir de l'information sémantique. Nous pensons également à des informations syntaxiques plus précises comme les fonctions grammaticales qui pourraient servir à caractériser les dépendances ou leur contexte. Nous pensons enfin à une information de domaine utilisée à des fins descriptives. Les dépendances pourraient être appréhendées comme des marqueurs de domaine plus précis que les simples mots et plus souples que les groupes nominaux complets qui sont sensibles aux phénomènes de variation.

Outre la prise en compte d'informations encore plus étendues, la prise en compte du contexte arborescent des dépendances est nécessaire s'il l'on souhaite s'orienter vers l'induction de grammaires à partir de corpus spécialisés, comme cela est montré dans [GH97]. Un telle extension au contexte, rendue techniquement possible grâce au formalisme des quasi-arbres, permettrait d'évaluer les dépendances, non plus isolément, mais en fonction de leur contexte dans la phrase. Du même coup, cette extension au contexte rendrait possible le traitement des noms simples qui pourraient être mis en dépendance indirecte avec le contexte.

Il est aussi envisageable d'explorer les dépendances autour des prédicats verbaux. Toutefois, il semble que les analyseurs ne sont pas encore prêts, étant donné qu'ils peinent sur la reconnaissance des structures argumentales des verbes.

Notons enfin, que le choix de la méthode d'apprentissage détermine les perfor-

mances du système. Des expérimentations avec d'autres méthodes d'apprentissage sont souhaitables. Le temps nous a manqué pour tester d'autres méthodes. Des processus de classification sont également envisageables, en vue d'identifier sous forme de classes l'émergence de régularités syntaxico-sémantiques propres aux échantillons.

Consolider les ressources sémantiques Nous soulignions à la fin du chapitre 2 que pour assurer l'adaptation rapide des systèmes d'indexation, il fallait minimiser l'ampleur des tâches d'écriture de règles et de codage lexical. Nous avons avancé sur le premier point, en définissant une fois pour toute autant de règles que de types de dépendance syntaxique que nous prenions en considération. Pour avancer sur le second point il aurait fallu également faire appel à l'apprentissage pour la catégorisation sémantique. Il est donc à souhaiter que des ressources en français équivalentes à *Wordnet* et des corpus sémantiquement annotés soient constitués et distribués librement. Les travaux sur l'étiquetage sémantique en bénéficieraient certainement, de même que les systèmes en tirant parti, comme celui que nous avons présenté.

Bibliographie

[AB96] Houssem Assadi and Didier Bourigault. Acquisition et modélisation de connaissances à partir de textes : outils informatiques et éléments méthodologiques. In *Actes du 10ème congrès Reconnaissance des Formes et Intelligence Artificielle (RFIA'96)*. AFCET, janvier 1996.

[ABM+95] Gilles Adda, Philippe Blache, Joseph Mariani, Patrick Paroubek, and Martin Rajman. Action GRACE : mise en place du paradigme d'évaluation. application au domaine de l'analyse morpho-syntaxique. In P. Blache, editor, *Actes de la conférence Traitement Automatique du Langage Naturel, TALN'95*, pages 72–79, Marseille, 1995.

[BC90] Janine Bouscaren and Jean Chuquet. *Grammaire et textes anglais - Guide pour l'analyse linguistique*. OPHRIS, 1990.

[BFGM90] Richard Beckwith, Christiane Fellbaum, Derek Gross, and George A. Miller. Wordnet: A lexical database organized on psycholinguistic principle. CSL report 42, Cognitive Science Laboratory, Princeton university, USA, March 1990.

[BHNZ97] Jacques Bouaud, Benoît Habert, Adeline Nazarenko, and Pierre Zweigenbaum. Regroupements issus de dépendances syntaxiques en corpus : catégorisation et confrontation à deux modélisations conceptuelles. *Revue Ingéniérie de la connaissance*, pages 207–223, mai 1997.

[Bou94a] Didier Bourigault. Extraction et structuration automatiques de terminologie pour l'aide à l'acquisition des connaissances à partir de textes. In *Actes 9ème Congrès Reconnaissance des Formes et Intelligence Artificielle*, pages 397–408, Paris, 1994.

[Bou94b] Didier Bourigault. *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, 1994.

[BPV93a] Roberto Basili, Maria Teresa Pazienza, and Paola Velardi. Acquisition of selectional patterns in sublanguages. *Machine Translation*, (8):175–201, 1993.

[BPV93b] Roberto Basili, Maria Teresa Pazienza, and Paola Velardi. Semi-automatic extraction of linguistic information for syntactic disambiguation. *Applied Artificial Intelligence*, (7):339–364, 1993.

[BPV93c] Roberto Basili, Maria Teresa Pazienza, and Paola Velardi. What can be learned from raw texts? *Machine Translation*, (8):147–173, 1993.

[BR94] Eric Brill and Philip Resnik. A rule-based approach to prepositional phrase attachment disambiguation. In *Proceedings of COLING-94*, 1994.

[Bri92] Eric Brill. A simple rule-based part of speech tagger. In *Actes, 3rd Conference on Applied Natural Language Processing (ANLP'92)*, pages 152–155, Trento, Italy, march 1992.

[Bri94] Eric Brill. Some advances in transformation-based part of speech tagging. In *Proceedings of the AAAI*, 1994.

[BT91] Bernard Bosredon and Irène Tamba. Verre à pied, moule à gaufres : prépositions et noms composés de sous-classes. *Langue Française*, 91:40–55, 1991.

[Cad92] Pierre Cadiot. *À* entre deux noms : vers la composition nominale. *Lexique*, 11:193–240, September 1992.

[CDGK94] Jacques Courtin, Danièle Dujardin, Damien Genthial, and Irène Kowarski. Analyse et génération morphologique avec le système PILAF. *T.A.L.*, 35(2):93–110, 1994.

[CGE96] Leacock C., Towell G., and Voorhees E.M. Toward building contextual representations of word sens using statistical models. In Branimir Boguraev and James Pustejovsky, editors, *Corpus Processing for Lexical Acquisition*, Language, Speech and Communication, pages 205–216. The MIT Press, Cambridge, Massachusetts, 1996.

[CH90] Kenneth W. Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, march 1990.

[Cha87] Alan F. Chalmers. *Qu'est-ce que la science ? Récents développements en philosophie des sciences*. Editions de la découverte, Science et société, 1987.

[CLB94] Viviane Clavier and Geneviève Lallich-Boidin. Modélisation linguistique de la suffixation en vue de l'analyse automatique. *T.A.L.*, 35(2):129–144, 1994.

[Con91] Patrick Constant. *Analyse syntaxique par couche*. Doctorat de l'ENST, École Nationale Supérieure des Télécommunications, Paris, avril 1991.

[Cou76] Yves Courrier. Analyse et langage documentaire. *DOCUMENTALISTE*, 13(5–6):178–189, Septembre–Décembre 1976.

[Dac94] Roland Dachelet. Sur la notion de sous-langage. Thèse de doctorat en sciences du langage, Université Paris VIII, Décembre 1994.

[Dah81] I. Dahlberg. Les objets, les notions, les définitions et les termes. In GIRSTERM Université Laval, editor, *Fondements théoriques de la terminologie*, pages 223–282. G. Rondeau, H. Felber, Laval, Québec, 1981.

[Dai94] Béatrice Daille. *Approche mixte pour l'extraction de terminologie: statistique lexicale et filtres linguistiques.* Thèse d'Informatique. Université de Paris VII, février 1994.

[Dao96] François Daoust. *SATO, système d'analyse de texte par ordinateur, Manuel de référence, Version 4.0.* Centre ATO, UQAM, Montréal, 1996.

[dL95] Claude de Loupy. Le modèle d'étiquetage d'éric brill. *TAL*, 36(1-2):37–46, 1995. Traitements probabilistes et corpus.

[Erl95] Erli. Présentation du dictionnaire Alethdic/FR, Janvier 1995. Document Erli fourni avec le dictionnaire.

[Fab96] Cécile Fabre. *Interprétation automatique des séquences binominales en anglais et en français. Application à la recherche d'informations.* PhD thesis, Université Rennes I, 1996. Thèse de doctorat en informatique.

[FH96] Helka Folch and Benoît Habert. Les quasi-arbres : un formalisme logique pour exprimer des requêtes en indexation structurée. In *Actes du colloque Informatique et Langue Naturelle*, pages 277–292, Nantes, octobre 1996. IRIN.

[FV92] Catherine Fuchs and Bernard Victorri. Modéliser la levée d'ambiguité à l'aide d'un réseau connexionniste. *Technique et Science Informatique*, 11(2):93–108, 1992.

[GH97] Eric Gaussier and Benoît Habert. Langue spécialisé : des séquences observées aux mots possibles. In Danielle Corbin, Bernard Fradin, Benoît Habert, Françoise Kerleroux, and Marc Plénat, editors, *Mots possibles et mots existants*, Lille, avril 1997.

[GMS+94] Miller G.A., Chodorow M., Landes S., Leacock C., and Thomas R.G. Using a semantic concordance for sens identification. In *Proceedings of the ARPA Human Language Technology Workshop*, San Francisco, 1994.

[Gou90] Daniel Gouadec. *Terminologie: constitution de données.* AFNOR, Paris, 1990.

[Gou93] Daniel Gouadec. *Terminologie et Terminotique - Outils, modèles et méthodes*. La Maison du Dictionnaire, Paris, 1993.

[Gou94] Daniel Gouadec. *Terminoguide no 1 données et informations terminologiques*. La Maison du Dictionnaire, Paris, 1994.

[Gre94] Gregory Grefenstette. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publisher, Dordrecht, The Netherlands, 1994.

[Gre96] Gregory Grefenstette. Evaluation techniques for automatic semantic extraction : Comparing syntactic and window based approaches. In Branimir Boguraev and James Pustejovsky, editors, *Corpus Processing for Lexical Acquisition*, Language, Speech and Communication, chapter 11, pages 205–216. The MIT Press, Cambridge, Massachusetts, 1996.

[Gro94] Gaston Gross. Classes d'objets et description des verbes. *Langages*, (115):15–30, septembre 1994.

[Gui70] Louis Guilbert. Fondements lexicologiques du dictionnaire - de la formation des unités lexicales. *Grand Larousse de la Langue Française*, 1970.

[Gui76] Louis Guilbert. La relation entre l'aspect terminologique et l'aspect linguistique du mot. *Inforterm Series 3*, pages 242–250, 1976. München Verlag Dokumentation.

[Har71] Zellig Harris. *Structures mathématiques du langage*. Dunod, Paris, 1971. Traduit de *Mathematical Structure of Language*, 1968, John Wiley, Chichester.

[HBDJ95] Benoît Habert, Philippe Barbaud, Fernande Dupuis, and Christian Jacquemin. Simplifier des arbres d'analyse pour dégager les comportements syntaxico-sémantiques des formes d'un corpus. *Cahiers de Grammaire*, (20):1–32, 1995.

[HBGN+97] Benoît Habert, Suzanne Bertrand-Gastaldy, Adeline Nazarenko, Fernande Dupuis, Elie Naulleau, Monique Lemieux, and Cynthia Delisle. Recyclage d'analyses syntaxiques automatiques pour le repérage de variantes de termes. In *Actes de RIAO*, pages 751–760, Montréal, juin 1997. CID et CASIS.

[Hei96] Serge Heiden. *Manuel Utilisateur de CorTeCs (version 1.1)*. UMR9952 Lexicométrie et Textes Politiques, ENS Fontenay/Saint-Cloud, 92211 Saint-Cloud, 1996. L'outil est téléchargeable à l'URL suivante : www.ens-fcl.fr/labos/lexico/cortecs.html.

[Her95] Marie-Luce Herviou. Applications d'extraction des connaissances à EDF-DER. In *Actes de IA'95*, Montpellier, 1995.

[HF96] Benoît Habert and Cécile Fabre. Simplifying nominal parse trees to find semantic types in corpus. In Nancy Ide, editor, *Research in Humanities, Proceedings ALLC-ACH 1993 and 1994*. Kluwer, 1996.

[HHPB+97] Benoît Habert, Marie-Luce Herviou-Picard, Didier Bourigault, Richard Quatrain, and Marielle Roumens. Un outil et une mt́hode pour comparer deux extracteurs de groupes nominaux. In *1ères Journés Scientifiques et Techniques FRANCIL*, Avignon, 1997. FRANCIL.

[Hin90] Donald Hindle. Noun classification from predicate argument structures. In *Actes, 28th Annual Meeting of the Association for Computational Linguistics (ACL'83)*, pages 268–275, Berkeley, CA, June 1990.

[HM94] Marie-Luce Herviou and Marie-Gaëlle Monteil. Les projets Eureka GENELEX et GRAAL : quel intérêt pour une entreprise telle qu'EDF. *Revue ICO-Québec*, Vol. 6(1-2):pp. 116 à 118, Printemps 1994.

[HN96] Benoît Habert and Adeline Nazarenko. 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, mai 1996. AFIA.

[HNN96] Benoît Habert, Elie Naulleau, and Adeline Nazarenko. Symbolic word clustering for medium-size corpora. In *16th International Conference on Computational Linguistics*, volume 1, pages 490–495, Copenhague, Danemark, 5-6 Août 1996.

[HP96] Marie-Luce Herviou-Piccard. Les outils d'indexation AlethIP issus du projet GRAAL : principes et utilisation. Technical Report HN-46/96/022, EDF, Direction des Études et Recherches, Clamart, 1996.

[Hut95] Alan Hutchinson. *Algorithmic Learning*. Oxford University Press, Oxford, 1995.

[Jac96] Christian Jacquemin. A symbolic and surgical acquisition of terms through variation. In *Proceedings of IJCAI'95*. Springer-Verlag, 1996.

[Jac97] Christian Jacquemin. *Variation terminologique : Reconnaissance et acquisition automatique de termes et de leurs variantes en corpus*. Habilitation à diriger des recherches, Institut de Recherche en Informatique de Nantes, Université de Nantes, 1997.

[Kar91] Lauri Karttunen. Finite-state constraints. In *The Proceedings of the Interational Conference on Current Issues in Computational Linguistics*, 1991. Article disponible sur www.xerox.fr.

[Kay97] Daniel Kayser. La sémantique lexicale est d'abord inférentielle. *Langages*, (113):92–106, mars 1997.

[KCGS97] Lauri Karttunen, Jean-Pierre Chanod, Gregory Grenfenstette, and Anne Schiller. Regular expressions for language engineering. HTML (www.xerox.fr), 1997. A paraître dans Natural Language Engineering.

[Kle84] Georges Kleiber. Dénomination et relations dénominatives. *Langages*, (76):77–94, 1984.

[Ler95] Pierre Lerat. *Les langues spécialisées*. Linguistique nouvelle. PUF, Paris, 1995.

[LTV95] C. Leacock, G. Towell, and E.M. Voorhees. Learning context to disambiguate word senses. In T. Petsche, Hanson, S.J., , and J. Shavlik, editors, *Computational learning theory and natural learning system*. The MIT Press, Cambridge, Massachusetts, 1995.

[Mar70] André Martinet. *Éléments de Linguistique Générale*. Armand Colin, Paris, 1970.

[MBF+90] George A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. Introduction to WordNet : An on-line lexical database. *Journal of Lexicography*, 3:235–244, 1990.

[MCM83] R.S. Michalsky, J.G. Carbonell, and T.M. Mitchell. *Machine Learning : An artificial intelligence approach*. Morgan Kaufmann, Los Altos, California, 1983.

[Mel88a] Igor Mel'cuk. *Dependancy Syntax*. SUNNY NY, New York, 1988.

[Mel88b] Igor Mel'cuk. Paraphrase et lexique dans la thèorie linguistique sens-texte. *Lexique*, (6):13–54, 1988.

[Mel95] Alan Melby. e-TIF : an electronic terminology interchange format. *Computer and Humanities*, 29(1-3):159–166, 1995.

[MF95] Andrei Mikheev and Steven Finch. A workbench for acquisition of ontological knowledge from natural language. In *Proceedings of 7th conference of the European Chapter of the Association for Computational Linguistics (EACL'95)*, pages 194–201, Dublin, Ireland, march 1995.

[MHF83] Mitchell Marcus, Donald Hindle, and Margaret Fleck. D-theory : talking about talking about trees. In *Proceedings of ACL'83*, pages 129–136, 1983.

[Mit82] T. Mitchell. Generalization as search. *Artificial Intelligence*, 18:203–226, 1982.

[MLH95] José Coch et Richard Leblond Marie-Luce Herviou. Vers une méthodologie de mise en place d'applications linguistiques. In *Actes de TALN'95*, Marseille, Juin 1995.

[Mou96] Isabelle Moulinier. *Une approche de la catégorisation de textes par l'apprentissage symbolique*. Thèse d'informatique, Université Paris 6, novembre 1996.

[Nau96] Elie Naulleau. Complémentarité des approches inductive et déductive pour une lecture terminologique de corpus. In *Actes de Rencontre des Etudiants-Chercheurs en Informatique pour le Traitement Automatique de la Langue (RECITAL'96)*, Dourdan, 1996.

[Nau97] Elie Naulleau. Des syntagmes nominaux pertinents pour le terminographe. In *Actes des Quatrièmes Journées Internationnales de Terminologie*, Barcelone, 1997.

[NHM96a] Elie Naulleau, Benoît Habert, and Marie-Gaëlle Monteil. Recycling a thesaurus to characterize and process terms. In Martin Gellerstam, Jerker Järborg, Sven-Göran Malmgren, Kerstin Noré, Lena Rogström, and Catarina Röjder Papmehl (eds), editors, *Seventh Euralex Proceedings*, pages 807–816, Göteborg, 1996. Göteborg University.

[NHM96b] Elie Naulleau, Benoît Habert, and Marie-Gaëlle Monteil. Tagging term components with semantic information. In Christian Galinski and Klaus-Dirk Schmitz, editors, *Terminology and Knowledge Engineering*, pages 110–117, Vienna, 1996. INDEKS-Verlag.

[OHD94] A. Ogonowski, ML. Herviou, and E. Dauphin. Tools for extracting and structuring knowledge from texts. In *ACOL*, page 1049, 1994.

[PC97] François Nemo Pierre Cadiot. Pour une sémiogenèse du nom. *Langages*, (113):24–34, mars 1997.

[Ré97] François Récanati. La polysémie contre le fixisme. *Langages*, (113):107–123, mars 1997.

[Raj95] Martin Rajman. Approche probabiliste de l'analyse syntaxique. *TAL*, 36(1-2):157–201, 1995. Traitements probabilistes et corpus.

[Ras91] François Rastier. *Sémantique et Recherches cognitives*. Presses Universitaires de France, Paris, 1991.

[RCA94] François Rastier, Marc Cavazza, and Anne Abeillé. *Sémantique pour l'analyse : de la linguistique à l'informatique*. Sciences Cognitives. Masson, Paris, 1994.

[Res93] Philip Resnik. *Selection and Information: A Class-Based Approach to Lexical Relationships*. PhD thesis, University of Pennsylvania, December 1993. (Institute for Research in Cognitive Science report IRCS-93-42).

[Res95] Philip Resnik. Disambiguation noun groupings with respect to wordnet senses. In David Yarowsky and Kenneth Church, editors, *Third Workshop on Very Large Corpora*, pages 54–68, Cambridge, Massachusetts, USA, june 1995.

[RM95] Lance A. Ramshaw and Mitchell P. Marcus. Text chunking using transformation-based learning. In David Yarowsky and Kenneth Church, editors, *Third Workshop on Very Large Corpora*, pages 82–94, Cambridge, Massachusetts, USA, june 1995.

[Sag87] Naomi Sager. Information formatting of medical literature. In Naomi Sager, Carol Friedman, and Margared S. Lyman, editors, *Medical Language Processing: Computer Management of Narrative Data*, chapter 10, pages 197–220. Addison-Wesley, 1987.

[Sal89] Gerard Salton. *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer.* Addison-Wesley, Reading, MA, 1989.

[SFe87] Naomi Sager, Carol Friedman, and Margared S. Lyman (editors). *Medical Language Processing: Computer Management of Narrative Data.* Addison-Wesley, 1987.

[Sil93] Max Silberztein. *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX.* Informatique linguistique. Masson, Paris, 1993.

[Slo93] Monique Slodzian. La VGTT et la Conception scientifique du Monde. *Le langage de l'Homme*, decembre 1993. DeBoeck, Bruxelles.

[Slo94] Monique Slodzian. La doctrine terminologique, nouvelle théorie du signe au carrefour de l'universalisme et du logicisme. *ALFA, Terminologie et linguistique de spécialité*, 7/8, 1994.

[Slo95] Monique Slodzian. Comment revisiter la doctrine terminologique aujourd'hui? *La banque des mots*, (7):11–18, 1995.

[Sma93a] Franck Smadja. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177, march 1993. Special Issue on Using Large Corpora: I.

[Sma93b] Frank Smadja. Xtract : An overview. *Computer and the Humanities*, 26:399–413, 1993.

[St94] Simon Sabbagh and GRAAL team. Graal eureka project : re-usable grammars for automatic language analysis. In *Proceedings of the Language Engineering Convention*, Paris, 1994.

[Sta94] Jean-David Sta. Evaluation de méthodes statistiques de filtrage de termes à partir d'un corpus. Note interne EDF-DER HN-46/94/042, 1994.

[Sta95] Jean-David Sta. Comportement statistique des termes et acquisition terminologique à partir de corpus. *TAL*, 36(1-2):119–132, 1995. Traitements probabilistes et corpus.

[Ste23] Rudolf Steiner. *Une théorie de la connaissance chez Goethe.* Editions Anthroposophiques Romandes, Genève, 1923.

[Vic60] B.C. Vickery. *Faceted classification: a guide to construction and use of special schemes.* Aslib, London, 1960.

[Voo93] Ellen Voorhees. Using WordNet to Disambiguate Word Senses for Text Retrieval. In *Actes, 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'93)*, pages 171–180, Pittsburg, PA, June 1993.

[VS92] K. Vijay-Shanker. Using descriptions of trees in a tree adjoining grammar. *Computational Linguistics*, 18(4):482–516, 1992.

[Wö81] Ernest Wüster. L'étude scientifique générale de la terminologie, zone frontalière entre la linguistique, la logique, l'ontologie, l'informatique et les sciences des choses. In GIRSTERM Université Laval, editor, *Fondements théoriques de la terminologie*, pages 55–108. G. Rondeau, H Felber, Laval, Québec, 1981.

[Yar92] David Yarowsky. Word-Sens Disambiguation Using Statistical Models of Roget's Categories Trained On Large Corpora. In *Actes, 14th International Conference on Computational Linguistics (COLING'92)*, Nantes, 1992.

Annexe A

Exploitation du dictionnaire AlethDic

A.1 Présentation du dictionnaire AlethDic v1.5.5

Le projet AlethDic/Fr de la société Erli vise à la constitution de dictionnaires exploitables par des automates d'analyse du langage naturel, et utilisables dans des applications diverses. Il vise à fournir une base étendue de données lexicales nécessaires à des applications TALN de langue française, sans préjuger des ajouts rendus indispensables par l'univers traité par chaque application. Les contenus du dictionnaires sont issus de sources variées : des dictionnaires développés par la société Erli pour des applications spécifiques, et d'autres sources lexicographiques éditoriales. Le dictionnaire AlethDic est conforme au modèle Genelex. C'est-à-dire qu'il s'articule sur trois couches. Une unité lexicale, qui correspond à un cheminement particulier dans les trois couches, est définie par le triplet : unité morphologique, unité syntaxique, unité sémantique.

La première couche dite morphologique recense les unités morphologiques simples et composées, autonomes (chien, chat) et non autonomes (pre-, bêta-, demi-). Des propriétés morphologiques leurs sont associées (mode de flexion par exemple).

L'unité syntaxique est le point d'entrée de la couche syntaxique. Chaque unité syntaxique est en relation avec une unité morphologique. Si une unité morphologique a plusieurs comportements syntaxiques, alors elle est en relation avec plusieurs unités syntaxiques. Chaque unité syntaxique est caractérisée par une description de l'unité morphologique dans une structure positionnelle. Par exemple pour le verbe *aimer*, les constructions suivantes sont décrites :

`USYN:P [SN0 aimer]`	nous aimons
`USYN:P [SN0 aimer SN1]`	il aime la montagne
`USYN:P [SN0 aimer SV1]`	il aime manger des cerises
`USYN:P [SN0 PRONOM aimer]`	nous vous aimons
`USYN:P [SN0 aimer P [CONJONCTION P1]]`	il aime qu'elle chante

Chaque type de construction syntaxique est identifié par une étiquette. Les constructions syntaxiques des noms, verbes, adjectifs et adverbes sont ainsi décrites dans le dictionnaire. Ce qui représente, toutes catégories grammaticales confondues 422 types de constructions.

Une unité syntaxique (et implicitement l'unité morphologique dont elle est issue) peut être associée à plusieurs unités sémantiques de la couche sémantique, c'est-à-dire qu'une même construction syntaxique pour une unité morphologique donnée peut correspondre à plusieurs sens, selon les différentes acceptions de l'unité morphologique.

La description des unités sémantiques est faite selon deux axes : un premier axe componentiel à base de traits valués permettant de décrire l'unité à partir de propriétés élémentaires. Ces traits sont multiples : des traits sémantiques généraux (ex: CONCRET), des traits de classes sémantiques (ex: MATÉRIAU), des traits de domaine (ex: MUSIQUE), des traits spécifiques distinctifs à l'intérieur d'une classe ou d'un domaine (ex: COMESTIBLE). Le second axe rend compte des relations sémantiques entre unités, permettant de décrire une unité sémantique par les relations qu'elle entretient avec les autres. D'une part des relations paradigmatiques entre unités substituables moyennant des modifications de sens : synonymie, antonymie, opposition, proximité, généricité, spécificité. Par exemple, on dénombre ainsi 6835 relations spécifique–générique (hyponyme–hyperonyme) de type *pomme–fruit*, ainsi que 2650 relations partie–tout (méronyme–holonyme) comme *voie–bateau*. On dénombre également 4634 relations de proximité ou de «collocations lexicales quasi-exclusives» de type *aîné–fils, fille, frère, soeur*. D'autre part, des relations sémantiques de dérivation. Elles relient des unités issues de catégories grammaticales différentes, par exemple adjectif–nom (agricole, agriculture), nom–verbe (amour, aimer), adjectif–verbe (blanc, blanchir), adjectif–adverbe (technique, techniquement), adverbe–nom (facilement, facilité).

Pour clore cette présentation très succinte d'AlethDic mais qui donne une idée de la richesse du dictionnaire, nous présentons en table A.1 les effectifs du lexique par catégorie grammaticale pour la version 1.5.5, d'après le même document Erli qui nous a permis de décrire AlethDic [Erl95]. Si le dictionnaire est riche en informations, ces dernières ne sont cependant pas toutes exploitées par l'analyseur AlethIP. Il est clair qu'AlethDic est sous-exploité, mais étant donné le type de grammaire qui l'exploite, il peut difficilement en être autrement. Par exemple, les constructions syntaxiques de même que les classes sémantiques ne sont pas utilisées pour résoudre les attachements prépositionnels. Notre dépendance vis-à-vis d'AlethDic se situe :

1. au niveau des analyses effectuées par AlethIP puisqu'elles s'appuient sur le dictionnaire, que ce soit lors de la lemmatisation-catégorisation ou lors de la phase d'analyse grammaticale.

2. au niveau de nos propres traitements, puisque l'enrichissement linguistique que nous réalisons s'appuie sur les informations extraites d'AlethDic : des informations syntaxiques (approximation de la prédicativité des noms) et sémantiques (réutilisation sous une forme simplifiée des classes sémantiques d'AlethDic).

TAB. A.1 – *Effectifs du lexique d'AlethDic v.1.5.5 (1995) par catégorie grammaticale*

	Effectifs
Noms propres	751
Noms communs	47281
Adjectifs cardinaux	34
Adjectifs ordinaux	34
Adjectifs qualificatifs	12065
Adverbes	2377
Verbes	8384
Prépositions	317
Conjonctions de coordination	10
Conjonctions de subordination	139
Interjections	165
Déterminants possessifs	2
Déterminants démonstratifs	2
Déterminant partitifs	1
Déterminant définis	1
Déterminants indéfinis	24
Pronoms démonstratifs	9
Pronoms indéfinis	24
Pronoms interrogatifs	6
Pronoms relatifs	8
Pronoms personnels	13
Pronoms impersonnels	1

TAB. A.2 – *Exemple de simplification de classes sémantiques*

ESPACE_N_C LOCATIF_T (*parking*)	devient	LIEU
ESPACE_N_C LOCATIF_T PARTIEL_T(*plancher*)	devient	LIEU
ESPACE_N_C LOCATIF_T ELABORE_T (*aérodrome*)	devient	LIEU
ESPACE_N_C LOCATIF_T ELABORE_T ADMINISTRATIF_T (*voirie*)	devient	LIEU
ESPACE_N_C LOCATIF_T ELABORE_T PROFESSIONNEL_T (*banque*)	devient	LIEU
ESPACE_N_C (*écartement*)	devient	LIEU
ESPACE_N_C VEGETAL_T (*rizière*)	devient	LIEU
ESPACE_N_C NATUREL_T (*estuaire*)	devient	LIEU_GEOGRAPHIQUE
GEO_N_C (*pays*)	devient	LIEU_GEOGRAPHIQUE
ESPACE_N_C LOCATIF_T NATUREL_T (*plage*)	devient	LIEU_GEOGRAPHIQUE

A.2 Les étiquettes sémantiques pour les noms

A.2.1 Réutilisation et simplification de l'existant

Nous avons récupéré les étiquettes sémantiques du lexique d' AlethDic, puis projeté celles-ci sur un nouveau jeu d'étiquettes, plus étroit. Nous avons ainsi réduit le nombre d'étiquettes sémantiques mais conservé dans la mesure du possible le plus d'information possible[1]. Le principal critère qui nous a guidé dans cette tâche est la nécessité d'avoir un jeu d'étiquettes sémantiques indépendant de tout domaine d'activité. Les traits de domaine ont donc été systématiquement supprimés. Nous recherchions également une granularité sémantique moins importante que celle donnée par le jeu d'AlethDic (voir table A.2). En réduisant la granularité, nous simplifions le processus de désambiguïsation. Certaines entrées lexicales, porteuses d'un certain nombre d'étiquettes sémantiques - en raison de leur polysémie – voyaient leur nombre d'étiquettes réduit. Du même coup, le système de désambiguïsation gagnait en robustesse et en facilité de mise au point, étant donné que les contextes lexico-syntaxiques à prendre en considération pour sélectionner une catégorie sémantique étaient moins différenciés. Enfin, nous avons trouvé que le système combinatoire de classes et de traits sémantiques distinctifs d'AlethDic, complexe à maîtriser, manquait de généricité et, faisait un mélange d'informations conceptuelles et de traits linguistiques de type restriction de sélection qui présentait parfois des incohérences.

Le jeu d'étiquettes a été donc été fortement réduit (de 322 à 72). La table A.2 montre des exemples de simplification autour de la notion d'espace ou de lieu. Cette seule simplication n'a pu corriger aucune incohérence. Cela n'était pas la finalité. Le but était seulement de tirer parti d'un existant.

1. Par exemple les distinctions entre formes dermique, géométrique, naturelle, organique, symptomatique, végétale ont été conservées, plutôt que regroupées sous la catégorie unique ENTITÉ-ABSTRAIT-FORME.

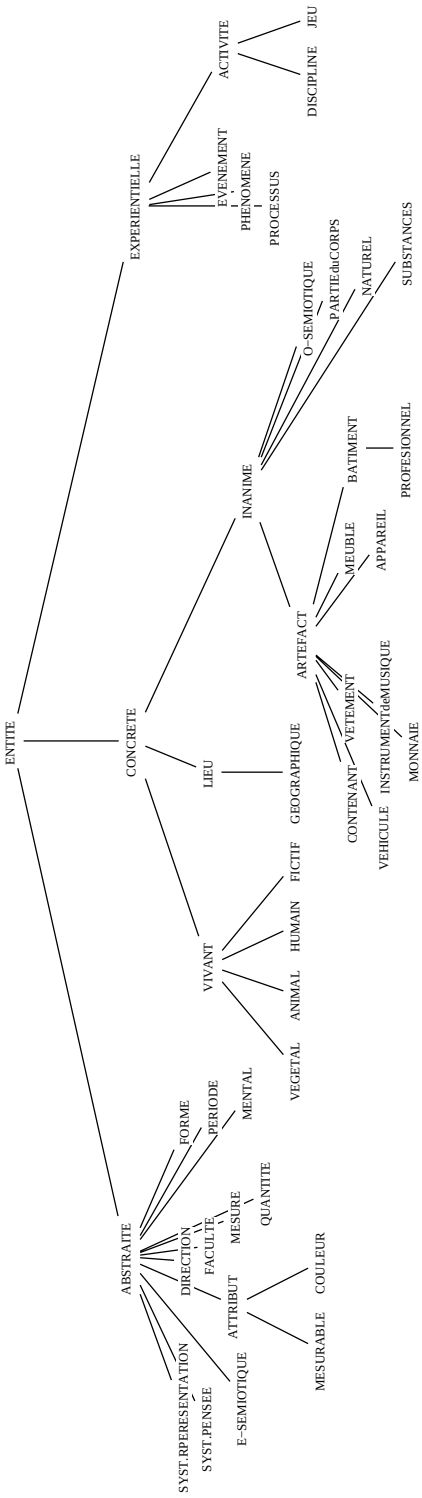
A.2.2 Liste et signification des catégories

Pour des questions d'optimisation et d'occupation disque, toutes les étiquettes sémantiques manipulées par notre application sont codées sous la forme d'entiers. Nous donnons ici la signification de ces étiquettes numériques. Nous reprenons les définitions fournies avec le dictionnaire AlethDic lorsqu'elles restent exactes pour les nouvelles catégories simplifiées, à moins que le nom de l'étiquette soit suffisamment explicite pou en saisir la signification.

TAB. A.3 – *Signification des catégories sémantiques pour les noms*

num.	CATÉGORIE	Commentaire
1	ENTITE-ABSTRAIT	Ensemble des notions universellement reconnues, non classables à un niveau inférieur (l'absolu, le bien , le mal)
2	ENTITE-ABSTRAIT-ATTRIBUT	Type de caractérisation d'un fait ou d'un objet (perméabilité, plasticité, variabilité).
3	ENTITE-ABSTRAIT-ATTRIBUT-COULEUR	Caractérisation de la couleur (le bleu, le jaune, le vert...)
4	ENTITE-ABSTRAIT-ATTRIBUT-MESURABLE	Caractérisation mesurable, quantifiable (largeur, longueur, luminance)
5	ENTITE-ABSTRAIT-DIRECTION	Directions (le centre, le nord, le sud, le dessous)
6	ENTITE-ABSTRAIT-ESEMIOTIQUE	Entité sémiotique: toute production liée à une activité intellectuelle, au langage ou à la pensée (le russe, une langue, une synthèse, une coordonnée)
7	ENTITE-ABSTRAIT-ESEMIOTIQUE-LETTRE	Des symboles ou les lettres d'un l'alphabet.
8	ENTITE-ABSTRAIT-FACULTE	Aptitude, capacité intellectuelle ou physique (le goût,la vue, l'habileté)
9	ENTITE-ABSTRAIT-FORME	Apparence formelle d'un objet (tresse, volute, vrille, rayon , pointillé)
10	ENTITE-ABSTRAIT-FORME-DERMIQUE	Forme visible sur la peau (ride, fossette, cicatrice)
11	ENTITE-ABSTRAIT-FORME-GEOMETRIQUE	Relatif à une forme dans l'espace (rectangle, polygone)
12	ENTITE-ABSTRAIT-FORME-NATURELLE	Forme qui existe à l'état de nature, qui n'a pas subi l'intervention de l'homme (alvéole)
13	ENTITE-ABSTRAIT-FORME-ORGANIQUE	(tubercule, myofibrille)
14	ENTITE-ABSTRAIT-FORME-SYMPTOMATIQUE	Formes en rapport avec des organes (excroissance, carcinome, bubon)
15	ENTITE-ABSTRAIT-FORME-FVEGETALE	Formes en rapport avec des organismes végétaux (gerbe, fane)

FIG. A.1 – *Hiérarchie des principales étiquettes sémantiques pour les noms*



16	ENTITE-ABSTRAIT-MENTAL	Concepts et manifestation issus des facultés humaines de réflexion (velléité, reproche, alibi)
17	ENTITE-ABSTRAIT-MESURE	Toute unité conventionnelle de quantification (cylindrée, décalitre, mètre)
18	ENTITE-ABSTRAIT-PERIODE	Période localisable dans le temps (journée, mardi, septembre, semestre, quinquennat)
19	ENTITE-ABSTRAIT-QUANTITE	Désigne des noms servant à déterminer une collection d'objet ou une portion de matière (kyrielle, gorgée, ensemble, infinité)
20	ENTITE-ABSTRAIT-SONORE	Qui renvoie à la production d'un son (tonalité, vrombissement, mélodie)
21	ENTITE-ABSTRAIT-SYSTPENSEE	Forme de doctrine (rationnalisme, positivisme, spiritisme, théisme)
22	ENTITE-ABSTRAIT-SYSTREPRESENT	(antinomie, référentiel, catégorie, libido)
23	ENTITE-CONCRET	Tout objet tangible
24	ENTITE-CONCRET-INANIME	Tout objet tangible inanimé
25	ENTITE-CONCRET-INANIME-ARTEFACT	Tout objet réalisé par l'homme
26	ENTITE-CONCRET-INANIME-ARTEFACT-APPAREIL	Outil ou assemblage d'éléments constituant un tout doté de fonctionnalités particulières (interrupteur, ponceuse)
27	ENTITE-CONCRET-INANIME-ARTEFACT-BATIMENT	Toute construction couverte (résidence, immeuble)
28	ENTITE-CONCRET-INANIME-ARTEFACT-BATIMENT-PROFESSIONNEL	Toute construction en rapport avec l'exercice d'une pofession ou d'une activité technique (cinémathèque, université)
29	ENTITE-CONCRET-INANIME-ARTEFACT-CONTENANT	Tout objet dont la fonction principale est de contenir un objet ou une substance (fiole, réservoir)
30	ENTITE-CONCRET-INANIME-ARTEFACT-INSTRUMENTde-MUSIQUE	Les instruments de musique (orgue, trompette)
31	ENTITE-CONCRET-INANIME-ARTEFACT-LIEU	
32	ENTITE-CONCRET-INANIME-ARTEFACT-MEUBLE	Tout objet d'ameublement (tabouret, armoire)
33	ENTITE-CONCRET-INANIME-ARTEFACT-MONNAIE	Nom de monnaie (rouble, roupie)
34	ENTITE-CONCRET-INANIME-ARTEFACT-VEHICULE	Engin construit pour servir de moyen de transport de marchandises ou de personnes (bulldozer, camion)
35	ENTITE-CONCRET-INANIME-ARTEFACT-VETEMENT	(costume, gant) vide

36	ENTITE-CONCRET-INANIME-NATUREL	objet qui existe dans la nature (planète, iceberg)
37	ENTITE-CONCRET-INANIME-OSEMIOTIQUE	Les objets sémiotiques regroupent toutes les représentations physiques (écrites, électroniques) de l'activité intellectuelle humaine (thesaurus, document, périodique, base de données)
38	ENTITE-CONCRET-INANIME-PCORPS	Regroupe les noms de partie du corps (doigt, main, coeur)
39	ENTITE-CONCRET-INANIME-SUBSTANCE	Noms de substance (silice, hydrogène)
40	ENTITE-CONCRET-INANIME-SUBSTANCE-ANIMALE	substances d'origine animale (cuir, propolis)
41	ENTITE-CONCRET-INANIME-SUBSTANCE-CHIMIQUE	substances chimiques (fluor, dioxine)
42	ENTITE-CONCRET-INANIME-SUBSTANCE-ELABORE	(savon, rhodoïd)
43	ENTITE-CONCRET-INANIME-SUBSTANCE-NATURELLE	(silice, onyx)
44	ENTITE-CONCRET-INANIME-SUBSTANCE-NOURRITURE	(ricotta, truffade)
45	ENTITE-CONCRET-INANIME-SUBSTANCE-ORGANIQUE	(saccharose, lysine)
46	ENTITE-CONCRET-INANIME-SUBSTANCE-VEGETALE	(fourrage, camphre)
47	ENTITE-CONCRET-LIEU	lieu abstrait ou concret (village, prairie)
48	ENTITE-CONCRET-LIEU-GEOGRAPHIQUE	lieu géographique (Provence, Metz)
49	ENTITE-CONCRET-VIVANT	Tout objet tangible qui doit son existence physique à des tissus vivants (animal, végétal).
50	ENTITE-CONCRET-VIVANT-ANIMAL	Organisme du règne animal (toucan, tortue)
51	ENTITE-CONCRET-VIVANT-ANIMAL-COLLECTIF	Désigne des collections d'animaux (troupeau, horde)
52	ENTITE-CONCRET-VIVANT-ANIMAL-COMESTIBLE	Animaux mangés par les hommes (bigorneau, langoustine)
53	ENTITE-CONCRET-VIVANT-FICTIF	Animaux fictifs ou légendaires (licorne, harpie)
54	ENTITE-CONCRET-VIVANT-HUMAIN	Toute personne humaine (un chauffeur, un égyptien)
55	ENTITE-CONCRET-VIVANT-HUMAIN-COLLECTIF	Groupe de personnes (garnison, tribu, société)
56	ENTITE-CONCRET-VIVANT-HUMAIN-COLLECTIF-UNITEADM	Groupe de personnes qui travaillent de concert (armée, préfecture, ministère)
57	ENTITE-CONCRET-VIVANT-MICROORGANISME	(cellule, pneumocoque)
58	ENTITE-CONCRET-VIVANT-ORGANISME	(foetus, plancton)
59	ENTITE-CONCRET-VIVANT-VEGETAL	(riz, tétragone)
60	ENTITE-CONCRET-VIVANT-VEGETAL-ARBRE	(figuier, épicéa)

61	ENTITE-CONCRET-VIVANT-VEGETAL-CHAMPIGNON	(cèpe, pleurotte)
62	ENTITE-CONCRET-VIVANT-VEGETAL-FLEUR	(rose, pensée)
63	ENTITE-CONCRET-VIVANT-VEGETAL-FRUIT	(citron, datte)
64	ENTITE-EXPERI-ACTIVITE	Noms d'activité physiques (aviron, ski)
65	ENTITE-EXPERI-ACTIVITE-DISCIPLINE	Noms de discipline (phonetique, solfège)
66	ENTITE-EXPERI-ACTIVITE-JEU	(puzzle, quille)
67	ENTITE-EXPERI-EVENEMENT	(accident, péripétie, conférence)
68	ENTITE-EXPERI-PHENOMENE	(feu, irisation)
69	ENTITE-EXPERI-PHENOMENE-ETAT	(pauvreté, stabilité)
70	ENTITE-EXPERI-PHENOMENE-MALADIE	(grippe, oreillons)
71	ENTITE-EXPERI-PROCESSUS	(décroissance, échauffement)
72	ENTITE-EXPERI-PROCESSUS-OPERATION	(brossage, paraffinage)

A.3 Les étiquettes sémantiques pour les adjectifs

Nous avons commencé un codage des adjectifs. Cela s'est rapidement révélé être une entreprise trop ambitieuse étant donné la complexité et la polysémie des valeurs sémantiques des adjectifs. Nous avons donc arrêté ce travail en cours. La classification dans son état actuel ne présente pas ou très peu de cohérence. Un certain nombre d'étiquettes sont quand même utilisées pour enrichir les syntagmes nominaux.

TAB. A.4 – *Signification des catégories sémantiques pour les adjectifs*

num.	CATÉGORIE	Commentaire
101	ALTER-COMPARAISON	Comparaison (analogue, égal, similaire)
102	ALTER-CORRESPONDANCE	Mise en correspondance (correspondant)
103	ALTER-DEICTIQUE	Déictique (autre)
104	ALTER-DUPLICATION	Duplication du référent nominal (nouveau, supplémentaire)
105	ALTER-ENUMERATION	Valeur d'énumération (différent, divers, respectif)
106	ALTER-INTERSECTION	Valeur d'intersection (commun)
107	EVENEMENTIEL	Valeur d'événement (accidentiel,fatal, imprévu)
108	FONCTION	Donne une fonction (préventif, propulsif)
109	LOC-DEFINIE	Défini un lieu non géographique (cranien)
110	LOC-GEO	Détermine un lieu géographique (tropical,alpin)

111	LOC-ORDINAL	Ordinal (dernier, premier)
112	LOC-SPATIAL	Donne une indication spatiale (périphérique, vertical, bas)
114	LOC-TPS-AGE	Donne une indication d'âge (vétuste, vieux, trentenaire)
115	LOC-TPS-ASP	Donne une indication temporelle aspectuelle (immédiat, simultané, repétitif)
116	LOC-TPS-ASP-DUREE	Donne une indication de durée (ponctuel, long, provisoire)
117	LOC-TPS-ASP-FREQUENCE	Donne une indication de fréquence (intermittent, rare)
118	LOC-TPS-ASP-FUTUR	Localisation temporelle future (ultérieur, prochain)
119	LOC-TPS-ASP-PASSE	Localisation temporelle passée (antérieur, ancien, récent)
120	LOC-TPS-ASP-PRESENT	Localisation temporelle présente (actuel, comptemporain)
121	LOC-TPS-ASP-SUCCESSION	Valeur de succession (consécutif)
123	LOC-TPS-PERIODE	Défini une période (trimestriel)
124	LOC-TPS-RYTHME	Donne une indication de rythme (régulier, cyclique)
125	MODALITE1	Valeur d'assertion (vrai, faux)
126	MODALITE2	Valeur d'évaluation d'éventualité (certain, éventuel, prévisible, probable)
127	MODALITE3-APPREC	Valeur de qualification appréciative (délicat, supportable, bénin)
128	MODALITE3-QUALIF	Valeur de qualification qualitative (fondamental, conventionnel)
130	MODALITE4	Valeur déontique, de contrainte (nécessaire, obligatoire)
131	PROP-ABSTRAITE	Définit une propriété abstraite (analytique, universel)
132	PROP-CHIMIQUE	Définit une propriété en rapport avec le domaine de la chimie (électrolytique, soluble)
133	PROP-COULEUR	Définit une couleur (bleu, vermillon)
134	PROP-FORME	Définit une propriété de forme (tubulaire, oval)
135	PROP-GEOM	Définit une propriété géométrique (symétrique, sphérique)

136	PROP-LOGIQUE	Définit une propriété logique (complémentaire, booléen)
137	PROP-MATH	Définit une propriété mathématique (linéaire, binomial)
138	PROP-METROLOGIE	Définit une propriété métrologique (kilométrique)
139	PROP-PHYSIQUE	Définit une propriété relative à la physique (oscillatoire, quantique)
140	PROP-RELATIFA	(torrentiel, salarial, portuaire)
141	PROP-SUBSTANCE	(oxalique, pulvérulent)
142	PROP-X	*Adjectifs de propriétés encore non codés*
144	QUANT-CARD	Adjectifs cardinaux (un, deux, trois, cent)
146	QUANT-LIMITE	Adjectifs définissant une limite (maximal, minimum)
147	QUANT-POIDS	Adjectif qualifiant le poids (léger, lourd)
148	QUANT-TAILLE	Adjectif qualifiant la taille (grand, petit, court)
149	QUANT-X	Autres types de qualification-quantification (moyen, demi, plein) *codage non achevé*

A.4 Les étiquettes sémantiques pour les adverbes

Les adverbes sont catégorisés avec les catégories suivantes. Ce sont les valeurs d'adverbes du dictionnaire AlethDic. Toutefois, étant donné que le filtrage ne prend pour le moment pas en compte les adverbes, ces catégories ne sont pas exploitées.

TAB. A.5 – *Les différentes valeurs adverbiales recensées dans AlethDic*

num.	CATÉGORIE	Exemple d'adverbe
201	AFFIRMATION	volontiers
202	CHRONOLOGIE	ultérieurement
203	FRÉQUENCE	mensuellement
204	HABITUDE	invariablement
205	INTENSITÉ	excessivement
206	NÉGATION	point
207	QUANTITÉ	peu
208	TEMPS	maintenant

A.5 Signification d'autres traits utilisés

A.5.1 Type de déterminant dans les dépendances de type nom_1 préposition nom_2

D=0	Déterminant zéro
D=d	Article défini
D=D	Adjectif démonstratif
D=I	Adjectif indéfini
D=p	Adjectif possessif

A.5.2 Signification du trait Xcons

Xcons=1	le nom accepte des arguments introduits par des prépositions, suffixe en *-tion* ou *-age*, il existe une relation nom–verbe (exemple : *dérivation–dériver*); ainsi les noms : *collaboration, piquage*
Xcons=2	le nom accepte des arguments introduits par des prépositions, suffixe en *-tion* ou *-age*, pas de relation nom–verbe dans le dictionnaire; ainsi les noms : *prestation, adéquation.*
Xcons=3	le nom accepte des arguments introduits par des prépositions, il existe une relation nom–verbe; ainsi les noms : *analyse, collecteur*
Xcons=5	le nom accepte des arguments introduits par des prépositions; ainsi les noms : *forum sur (l'emploi), crue de (du fleuve), but de (l'action).*
Xcons=6	S'applique aux adjectifs, participes passés ou participes présents qui entrent dans des constructions du type *exempt de, accordé par, conduisant à*

Annexe B

Les règles de désambiguïsation

B.1 La syntaxe des règles de désambiguïsation

La syntaxe des règles est la suivante :

```
LEXEME:: Sous-règle-1 Sous-règle-2 ... Sous-règle-n:.
```

L'identifieur de la règle LEXEME, est suivi par la déclaration successive de sous-règles qui décrivent les différents contextes linguistiques de LEXEME à prendre en compte. Une sous-règle déclare des conditions à vérifier, des actions à exécuter si ces conditions sont réalisées et un exemple illustratif de l'action de la règle.

```
Sous-règle-i:=
Si[i] { Condition-1 Condition-2 ...Condition-n }
Alors {
Action-1 Action-2 ... Action-n }
Exemple « ... »
```

Chaque condition, de même que chaque action doit respecter la syntaxe suivante :

```
Condition-i:= MembreGauche Opérateur-conditionnel MembreDroit
Action-i:= MembreGauche Opérateur-action MembreDroit
```

Un membre, gauche ou droit, peut avoir plusieurs statuts:

```
Membre := Fonction | Booléen | $Variable | _Historique |
Identifieur |Graphie | Forme | Prefixe | Suffixe | Liste de
{Identifieur,Graphie,Forme,Prefixe,Suffixe}

Booléen := oui | non

Graphie : Forme.Catégorie

Identifieur := Catégorie |Trait Sémantique | Trait Syntaxique

Opérateur-conditionnel := in | not-in | == | !=
Opérateur-action := = | to
```

Les opérateurs de tests sont l'égalité « == », la différence « != », l'appartenance d'un élément dans une liste « in », ou son absence d'une liste «not-in ». Ils sont applicables sur des entiers et des chaînes de caractères. Les opérateurs pour les actions sont l'affectation « = », et l'envoi « to » (utilisé pour envoyer tout ou partie de la phrase dans un fichier ou vers une autre règle).

Les variables (déclarées dans des blocs d'actions et testées dans des blocs conditionnels) sont globales, c'est-à-dire visibles par toutes les règles pendant le processus de désambiguïsation. Elles sont du type entier ou chaîne de caractères. Il en est de même pour les historiques dans lesquels on peut ajouter des éléments pour constituer des listes (les variables et les historiques sont utiles pour gérer une mémoire de désambiguïsation : une forme a-t-elle déjà été désambiguïsée dans le corpus?).

Un identifieur est une chaîne de caractères représentant une catégorie (le jeu de catégories lexicales fournies par le catégoriseur), un trait sémantique, ou une information syntaxique. Une forme (graphique) est un lexème donné sans catégorie. Une graphie est une forme avec catégorie lexicale. Suffixe et préfixe sont des chaînes de caractères. On peut déclarer une liste d'identifieurs, de graphies, de préfixes, ..., entre crochets, et séparés par des virgules.

Les fonctions retournent des informations linguistiques en accédant au dictionnaire et au corpus. Elles sont décrites dans le tableau B.1. Elles se déclarent ainsi :

```
Fonction := Identifieur_de_fonction ( Argument )

Argument := $Variable | _Historique | Registre | Pattern
Registre := ~1 ~2 ... ~22 , ~~

Identifieur_de_fonction := match | catégorie | graphie | forme | préfixe |
suffixe | vsem | genre | nombre | vsemseek | npred | narg1 | narg2 | ...

Pattern := (Unaire MachOperator Registre)+
MachOperator := * |? | !
Unaire := ^ | ε
```

La première fonction du bloc de conditions est toujours la fonction `match`, dont l'argument suit une syntaxe exprimant une mise en correspondance de la phrase avec un certain schéma linguistique (*linguistic pattern matching*). L'argument de la fonction `match` consiste donc en une suite d'opérateurs de mise en correspondance

déclarant des éléments optionnels ou obligatoires (* pour zéro ou n éléments de la phrase,? pour zéro ou un, '!' un seul obligatoire). Ces opérateurs acceptent l'opérateur unaire de négation '^'. Aussi, pour être en mesure de décrire un schéma linguistique, il faut associer à chacun de ces opérateurs des contraintes linguistiques. Ceci est réalisé en déclarant des tests associés aux opérateurs de mise en correspondance grâce à des registres (∼numéro de registre). Le registre spécial ∼∼ correspond à la forme à désambiguïser (identifieur de la règle).

Les actions suivent la même syntaxe que les tests (à l'opérateur près), mais la ligne est évaluée de droite à gauche, en vue de modifier le contenu du membre gauche (par exemple : catégorie(∼1) = Adjectif donne au contenu du registre ∼1 la catégorie Adjectif); alors que les tests sont évalués de gauche à droite (par exemple catégorie(∼1) in [Adjectif, Article] vérifie que le contenu du registre ∼1 a bien la catégorie article ou adjectif).

Exemple de règle :

```
LIGNE:: si [1]
match ( * !∼∼ !∼1 !∼2 * ) == oui
graphie(∼1) == DE.preposition
categorie(∼2) == Nom
suffixe(∼2) in [-TION,-EMENT,-AGE]
alors {
valeur_semantique(∼∼) = Artefact
}
exemple « ligne d'aspersion »
si [2] {
match ( * !∼1 !∼2 !∼∼ * ) == oui
graphie(∼2) == EN.preposition categorie(∼2) == Nom
valeur_semantique(∼1) in [Processus,Artefact] }
alors {
valeur_semantique(∼∼) = Forme }
exemple «documentation en ligne»:.
```

La règle de l'exemple ci-dessus est écrite pour le nom *ligne*. La première sous-règle vérifie que *ligne* est suivi de la préposition de puis d'un nom qui se termine par le suffixe *-tion, -ement* ou *-age*. Si c'est le cas, le nom *ligne* est alors considéré comme un nom d'artefact. La deuxième sous-règle vérifie que le nom est précédé de la préposition en et d'un nom de processus ou d'artefact. Si c'est le cas, on attribue au nom *ligne* le type de référent « forme », le syntagme prépositionnel « en ligne » constituant pratiquement une sorte d'adverbe. Bien entendu, il y aurait d'autres contextes à énumérer pour le nom ligne.

Fonctions linguistiques implémentées ou restant à implémenter Dans la première colonne du tableau B.1, on trouve le nom de la fonction. Certaines fonc-

tions peuvent être utilisées comme fonctions-tests (T), comme actions-instructions (A) dans la partie action de la règle, ou les deux (T/A). Ceci est mentionné dans la seconde colonne. La troisième colonne décrit ce que fait la fonction ou l'action. La dernière colonne précise si la fonction ou l'action est déjà implémentée ou non.

TAB. B.1 – *Fonctions implémentées*

Indentifieur	Statut	Description	Impl.
match	T	Permet de déclarer un patron linguistique et de tester sa présence dans la séquence linguistique. Les éléments de la phrases retrouvés sont accessibles par des registres.	oui
forme	T	Renvoie la forme graphique du contenu d'un registre.	oui
graphie	T	Renvoie la référence lexicale au dictionnaire du contenu d'un registre.	oui
catégorie	T/A	Renvoie ou définit la partie du discours du lexème contenu dans le registre.	oui
valeur_sémantique	T/A	Renvoie ou définit la valeur sémantique du lexème contenu dans le registre. Valable pour les noms, adjectifs, adverbes et prépositions.	oui
préfixe	T	Renvoie le préfixe de la forme passée en argument.	non
suffixe	T	Renvoie le suffixe de la forme passée en argument.	oui
genre	T/A	Renvoie ou définit le genre de la forme passée en argument.	non
nombre	T/A	Renvoie ou définit le nombre de la forme passée en argument.	non
domaine	T/A	Renvoie ou définit le ou les domaines abordés par le texte. Géré par une variable-historique _Domaine.	non
npred	T/A	Renvoie ou définit le statut prédicatif du nom.	oui
vsem_seek	T	vsem_seek(± n phrases, registre, trait sémantique, $var). Recherche dans les n phrases précédentes (-) ou suivantes(+) la même forme que celle contenue dans le registre, ayant la valeur sémantique « trait sémantique ». Ci cette forme est trouvée, vsem_seek renvoie cette valeur sémantique, sinon renvoie la valeur sémantique contenue dans la variable $var passée en argument.	oui

print	A	print(liste de registres) : écrit dans un fichier le contenu des registres.	oui
tprint	A	tprint(liste de registres) : écrit dans un fichier le sous arbre syntaxique correspondant à la liste des registres fournie.	non
send	A	send(liste de registres) : envoie la séquence linguistique définie par la liste de registres vers une autre règle ou ensemble de règles.	non
SN_tête	T	Renvoie un booléen indiquant si le nom passé en argument est la tête d'un syntagme nominal.	non
nom_parent	T	Renvoie le premier nom parent de l'argument dans la structure syntaxique de la phrase.	non
n_arg	T	Renvoie une propriété sur la structure argumentale d'un nom prédicatif (à partir des données fournies par AlethDic). Par exemple Nom DE-arg1 PAR-arg2 à partir du contexte droit fourni à la fonction.	non

B.2 Ecriture assistée de règles

B.2.1 Opération préliminaire : sélection des formes à désambiguïser

Cette opération consiste à déclarer les formes ambiguës pour lesquelles on souhaite définir des contextes-solutions. Il faut donc sélectionner un corpus, et donner la liste des formes supposées ambiguës ou à inspecter. A la demande, cette liste de formes est générée automatiquement à partir des informations déjà présentes dans le dictionnaire de désambiguïsation ; les formes qui ne sont pas déclarées monosémiques, et qui n'appartiennent pas au dictionnaire de désambiguïsation, sont retenues. A partir de cette liste de formes, le programme construit des concordances, de manière à être capable de lister tous les contextes d'une forme donnée dans le corpus.

B.2.2 Marche à suivre pour l'écriture de règles

Sélection du corpus et la forme pour laquelle on souhaite écrire des règles

L'utilisateur sélectionne un corpus de travail (voir figure B.1). Par exemple ard95. Ensuite, il sélectionne une forme parmi celles disponibles, par exemple BASE, dans le texte ARD95. D'après la figure B.2, la première colonne affiche les formes ambiguës ou non résolues, la seconde indique le nombre d'occurrences de cette forme dans le corpus. La dernière indique le nombre de règles de désambiguïsation écrites pour la forme.

FIG. B.1 – *Sélection du corpus de travail*

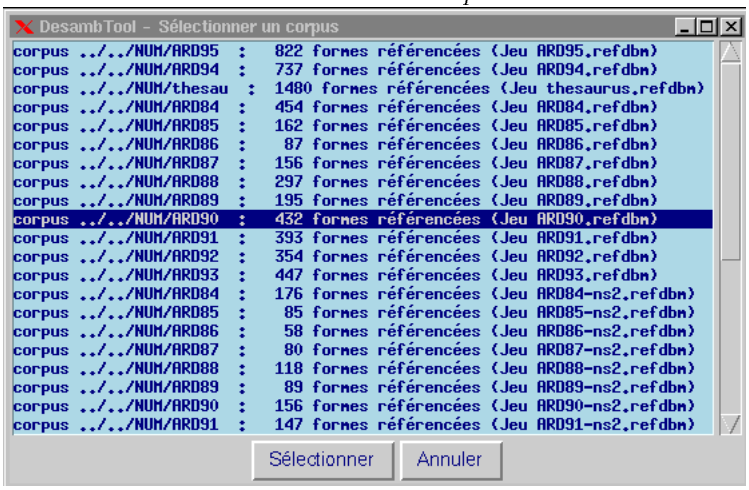


FIG. B.2 – *Sélection de la forme a désambiguïser*

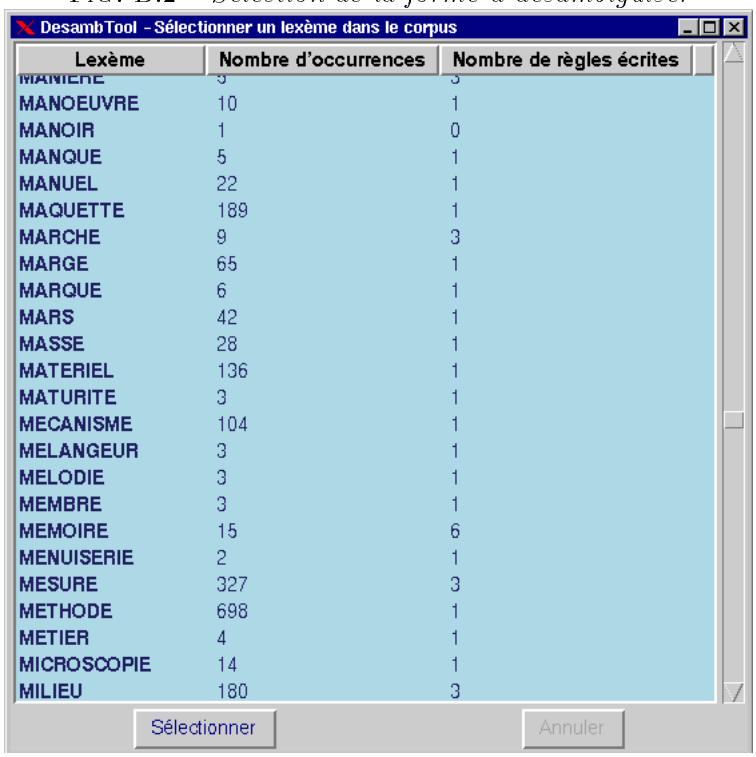
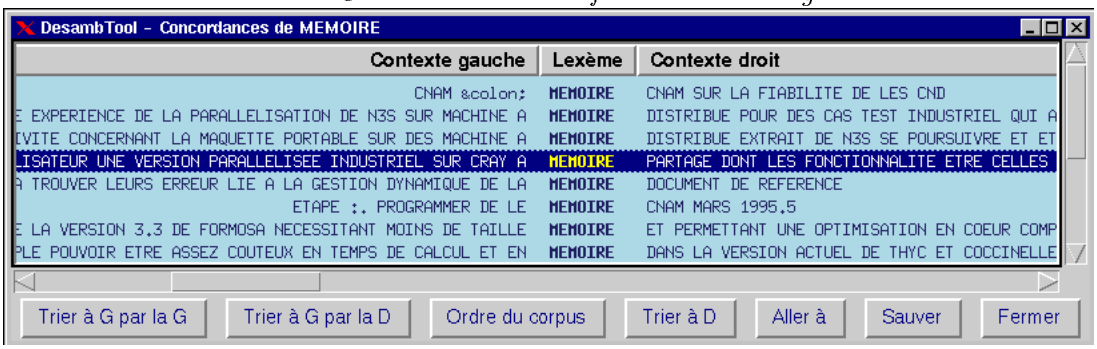


FIG. B.3 – *Concordances de la forme à désambiguïser*



Fenêtre de concordances de la forme à résoudre

La fenêtre de concordances (voir figure B.3) affiche les contextes gauche et droit des occurrences de la forme ambiguë, recensées dans le corpus. On peut effectuer un tri sur le contexte droit, de manière à faire apparaître des régularités à droite, ou sur le contexte gauche, vers la gauche ou vers la droite, pour faire apparaître des régularités à gauche. Les contextes ainsi triés peuvent être sauvegardés en spécifiant le taille des contextes gauche et droit. Un double clic sur une occurrence ou un clic sur le bouton « Aller à » met à jour la vue de travail pour la phrase sélectionnée dans la fenêtre.

Visualisation du contexte de la forme dans la vue de travail

La vue de travail (figure B.4) est la fenêtre de visualisation des phrases du corpus enrichi. Certaines informations syntaxiques et sémantiques y sont représentées.

Il faut noter qu'à ce stade, les textes traités ne sont pas de simples textes ASCII, mais des textes linguistiquement enrichis (voir Annexe B, section 6 Eléments techniques). Les phrases d'un texte sont représentées concurremment sous la forme d'arbres syntaxiques (calculés par l'application AlethGram) et de séquences plates (suite de mots). A chaque mot de la phrase sont associées les informations linguistiques définies plus haut : des étiquettes sémantiques ont notamment été attribuées aux formes monosémiques, et aux formes polysémiques pour lesquelles il existait déjà des règles de désambiguïsation.

L'interface permet de visualiser chaque occurrence de la forme ambiguë dans son contexte phrastique voire extra-phrastique . Une barre de menus donne accès à un certain nombre de fonctions.

L'ascenseur horizontal permet de faire défiler la phrase dans la fenêtre. Chaque mot de la phrase occupe une colonne. Sous la graphie, qui occupe la première ligne de chaque colonne, sont indiquées : la catégorie lexicale (en ligne 2), la valeur sémantique (en ligne 3) et une information de prédicativité (en ligne 4), pertinente pour les noms (consommation de X) et pour les adjectifs (sujet porteur de Y). Ces informations sont codées sous la forme d'entiers (elles ont aussi un équivalent sous forme de chaîne de

FIG. B.4 – *Visualisation d'une séquence*

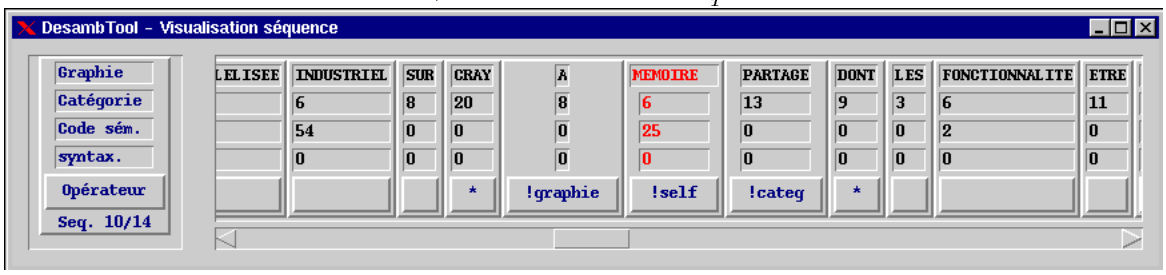
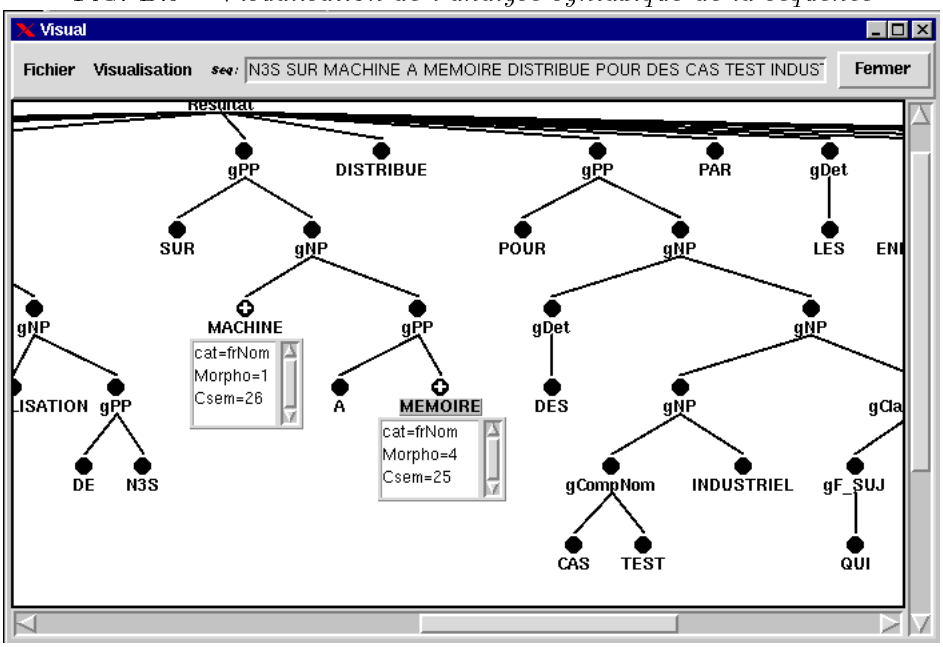


FIG. B.5 – *Visualisation de l'analyse syntaxique de la séquence*



caractères, donné dans la table de conversion, accessible depuis le menu d'aide). Le contenu des lignes 2 à 4 peut être édité directement, modifié et enregistré (commande Séquence-enregistrer du menu), ce qui a pour effet de mettre instantanément à jour le corpus. Cette vue de travail est conçue pour écrire des règles de catégorisation sémantique. En effet, la cinquième ligne de la colonne de la vue de travail est un bouton utilisé pour définir le contexte de la forme à désambiguïser. Cette forme apparaît dans la séquence en gras et en rouge (ici BASE dans la figure B.4).

Visualisation de l'analyse syntaxique de la séquence

Cette fenêtre de la figure B.5 permet l'affichage de l'arbre d'analyse de la phrase sélectionnée et visualisée en même temps sous forme séquentielle dans la vue de travail. La structure syntaxique est représentée sous la forme d'un arbre (figure 4); sur les noeuds terminaux, on peut faire apparaître dans une petite fenêtre, les traits

FIG. B.6 – *Sélection des opérateurs*

associés aux lemmes.

Définition assistée d'un contexte de désambiguïsation

Pour définir un contexte, on utilise les boutons-opérateurs, dessinés sous chaque lemme en 5ème ligne (voir figure B.4). Il faut avoir une idée précise de ce qui va permettre de désambiguïser, quels critères, quels indices dans le contexte de la forme vont être utilisés. Ensuite, on fait apparaître ces critères ou une partie de ces critères, sous les éléments à prendre en compte dans le contexte, en appuyant sur les boutons correspondants. La pression sur un des boutons fait apparaître une boîte de dialogue (figure B.6) décrivant les opérateurs de mise en correspondance, et les principales fonctions linguistiques que l'on peut leur associer.

Génération puis édition de la règle

Une fois le contexte de désambiguïsation défini à l'aide des boutons, la commande Règle-Générer permet de générer une règle de désambiguïsation qui prend en compte les informations définies avec les boutons-opérateurs. La règle est affichée dans une fenêtre d'édition (figure B.7) pour être modifiée et complétée si nécessaire. En effet, toutes les fonctions linguistiques ne sont pas encore disponibles depuis la boîte de dialogue associée aux boutons.

Compilation de la règle

La règle ayant été modifiée ou complétée, elle est enregistrée puis compilée. La compilation vérifie la syntaxe de la règle et transforme celle-ci dans un format dit exécutable, où toutes les références à des objets du dictionnaire (graphies, propriété, traits sémantiques,...) sont résolues et remplacées par des entiers. Lorsque la règle est compilée, elle peut être testée sur le contexte avec lequel elle a été définie, pour

FIG. B.7 – *Edition d'une règle générée*

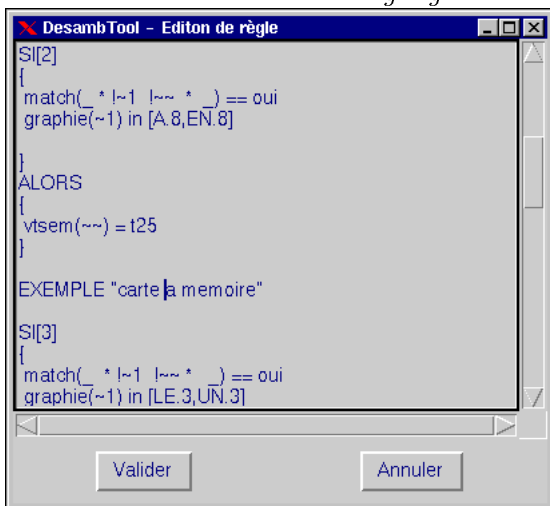
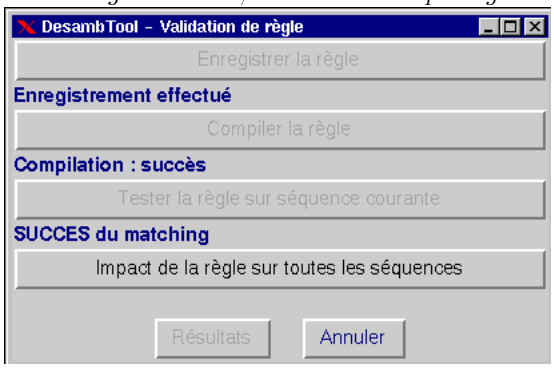


FIG. B.8 – *Enregistrement, validation et pistage de la règle*



vérifier son comportement (voir figure B.8). Après quoi, le comportement de la règle doit être observé sur tout le corpus. Cette dernière opération est appelé pistage.

B.2.3 Pistage des règles

Pour pister une règle il faut l'appliquer à toutes les occurrences de la forme ambiguë du corpus. Après quoi, on sait pour quelles occurrences la règle a été exécutée. Le résultat du pistage est ensuite affiché comme dans l'exemple de la figure B.9. On est alors en mesure d'évaluer la couverture de la règle (nombre d'occurrences touchées). Les problèmes de recouvrement peuvent aussi être détectés. La première colonne dans la fenêtre de résultats indique le numéro de phrase, la seconde indique si la règle a été appliquée ou non ; la troisième montre les règles qui ont déjà été appliquées avec succès à la phrase. La dernière colonne est un bouton permettant de visualiser la phrase du corpus pour vérifier si la règle devait effectivement être ou ne pas être appliquée. Par exemple la figure B.9 montre le résultat de pistage

FIG. B.9 – *Résultats du pistage d'une règle*

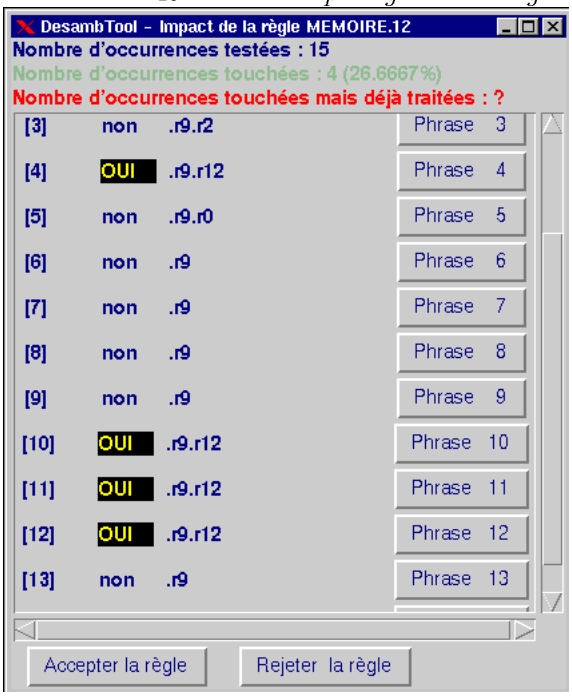
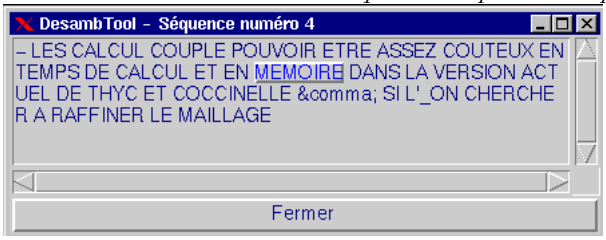


FIG. B.10 – *Visualisation d'une séquence à partir du pistage*

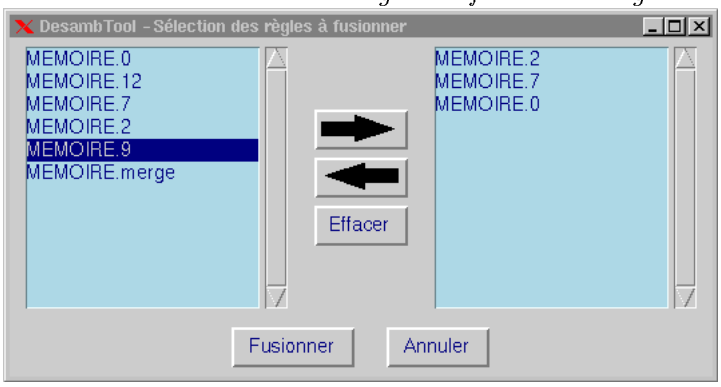


d'une règle nommée[1] MÉMOIRE.9 (titre de la fenêtre). L'état de la fenêtre montre que pour les phrase numérotées 4, 10, 11 et 12, la règle a été appliquée, et qu'il existe une autre règle nommée MÉMOIRE.9 qui s'applique à ces même phrases. Cette règle r9 s'étant appliquée là où r12 a échoué (phrase numérotée 3, 5, 6, 7, ,8 9, et 13), on en déduit pour la fusion ultérieure de ces règles que MÉMOIRE.12 (r12) est plus spécifique que MÉMOIRE.9 (r9).

En clichant sur le bouton correspondant à une l'occurrence dans la phrase 58, on fait apparaître son contexte phrastique comme cela est montré dans la figure B.10

1. On peut donner à une règle le nom que l'on souhaite. Par défaut le nom résulte de la concaténation de la forme lexicale pour laquelle on écrit la règle, d'un point, et du numéro de phrase à partir de laquelle on a construit le contexte de désambiguïsation

FIG. B.11 – *Boîte de dialogue de fusion des règles*



B.2.4 Fusion de règles

Lorsque plusieurs règles ont été mises au point, chacune prenant en compte un contexte-solution différent, elles doivent être regroupées en une seule règle fédératrice. Ces règles sont alors appelées sous-règles. La boîte de dialogue (voir figure B.11) permet de fusionner les sous-règles dans un ordre choisi, l'ordre de spécificité décroissante des contextes linguistiques, pour former une unique règle qui pourra être exportée vers le dictionnaire de désambiguïsation.

B.2.5 Synopsis des menus et des fonctions

Le menu Forme

Nom de la commande	Raccourci	Commentaire
Choisir une forme	Ctrl-F	Sélectionne ou change la forme à désambiguïser ou à définir dans le corpus choisi.
Quitter	Ctrl-Q	Quitte l'application

Le menu Séquence

Suivante hline Précédente	Ctrl-n	Phrase suivante du corpus Phrase précédente
Aller à	Ctrl-A	Aller à une phrase, étant donné son numéro.
Vue arbre syntaxique		Affiche la représentation syntaxique de la phrase active dans la vue séquentielle.
Références		Affiche les références (positions absolues dans le fichier de représentation numérique du corpus) de toutes les phrases ou apparaît une ou plusieurs occurrences de la forme ambiguë ou non définie. Surtout utilisé pour la mise au point et la vérification des données.
hline Concordances	Ctrl-C	Affiche la fenêtre de concordances de la forme à traiter.
Enregistrer	Ctrl-S	Enregistre dans le corpus les modifications apportées à la phrase dans la vue de travail (voir figure 7).

Le menu Règles

Générer	Ctr-G	Génère une règle de désambiguïsation lexicale à partir des informations disponibles dans la vue de travail (voir figure 7, 9 et 10)
Editer	Ctrl-E	Affiche une boîte de dialogue permettant de sélectionner puis d'éditer une règle déjà existante. Permet également l'effacement, et la duplication sous un autre nom de règles existantes.
Appliquer		Permet d'appliquer une règle existante au corpus.
Fusionner	Ctrl-u	Affiche la boîte de dialogue de fusion de règles. Regrouper des sous-règles en une règle fédératrice
Répercussion:Active		Drapeau activant l'enregistrement dans le corpus des actions d'une règle lorsque celle-ci est exécutée.
Répercussion:Inactive		Désactivation de l'enregistrement des actions d'une règle dans le corpus lorsque celle-ci est exécutée.
Importer	Ctrl-I	Permet l'importation de règles stockées dans une autre base de règles que celle utilisée pour le corpus sélectionné.
Exporter	Ctrl-x	
Décharger		Permet de décharger une base de règle au format ASCII. Utile pour transférer des règles d'une machine à une autre lorsque des «magic numbers» différents sont générés pour les fichiers binaires (exemple : SUN <=> Linux/Intel).

Le menu Aide

Ce menu permet l'affichage des codes des catégories lexicales, syntaxiques et sémantiques. Le prototype actuel affiche des codes numériques à la place de catégories plus intelligibles. En attendant d'améliorer l'ergonomie, les commandes du menu d'aide affichent la correspondance entre les catégories

Annexe C

Constitution d'échantillons d'apprentissage

C.1 Méthodes de constitution

Les échantillons d'apprentissage pour la construction des profils peuvent être construits de diverses manières. Dans tous les cas, ces échantillons sont en final composés de dépendances syntaxiques.

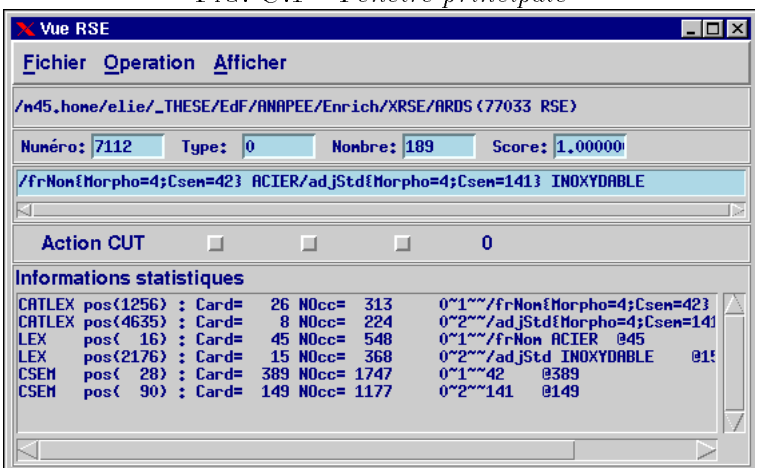
C.1.1 Constitution à partir de documents

Les profils peuvent être construits à partir de documents. Des documents (ou paragraphes) choisis peuvent alimenter les échantillons négatif et positif fournis à la procédure d'apprentissage. Les documents ainsi sélectionnés pour alimenter l'échantillon négatif (et respectivement positif) sont analysés et subissent la même procédure d'enrichissement que les textes dont on souhaite extraire les SNP. Les dépendances lexico-syntaxiques sont calculées à partir de ces analyses et viennent alimenter l'échantillon négatif (respectivement positif). Cette manière de construire les échantillons permet de confronter des documents normalisés sous la forme de dépendances lexico-syntaxiques nominales. Les profils construits à partir de ces échantillons ne retiennent que les dépendances syntaxiques spécifiques à chacune des classes de documents (négative *vs.* positive) construite. Les dépendances syntaxiques communes aux deux échantillons constituent selon l'option choisie soit des éléments pertinents, soit des éléments non pertinents.

C.1.2 Constitution à partir de listes de syntagmes

Les profils peuvent également être construits à partir de données terminologiques ou documentaires existantes. Par exemple, un thesaurus pourra être recyclé sous la forme de profils de pertinence. Des listes de syntagmes nominaux respectivement jugés pertinents et non pertinents seront alors fournies au système qui, après analyse, enrichissement linguistique et décomposition en dépendances élémentaires, les

FIG. C.1 – *Fenêtre principale*



transformera en profils de filtrage.

C.1.3 Constitution à partir des dépendances lexico-syntaxiques

Une troisième solution pour constituer des profils est de sélectionner manuellement des dépendances syntaxiques élémentaires. L'interface de gestion des dépendances élémentaires que nous avons développée permet d'assister cette tâche. Valider manuellement les dépendances syntaxiques élémentaires est long mais permet d'obtenir une plus grande précision dans la définition des profils. Nous décrivons maintenant les fonctionnalités de l'interface.

C.2 L'interface utilisateur

L'interface permet de visualiser et de modifier les paramètres associés aux dépendances syntaxiques élémentaires. Elle permet d'avoir une vision globale de l'utilisation de ces dépendances sur tous le corpus. Elle permet également d'exporter vers d'autres applications des dépendances qui répondent à certains critères spécifiés. Par exemple, il est possible d'exporter vers le logiciel *ZELLIG* [HNN96] des dépendances élémentaires préfiltrées avec un profil de filtrage. Les composantes construites par *ZELLIG* pourraient ainsi reflète l'impact du filtrage dans les réseaux d'unités lexicales mises en relation.

C.2.1 Informations associées à une dépendance élémentaire

Chaque dépendance lexico-syntaxique est représentée par son type et un numéro unique. Pour chacune d'entre elles, on connaît sa fréquence dans le corpus et la position des groupes nominaux dans lesquels elle apparaît. Ainsi on peut lire dans la fenêtre principale (figure C.1) les informations suivantes : une base

contenant 77033 dépendances qui est nommée ARDS et placée dans le répertoire /m45.home/elie/_THESE/EdF/ANAPEE/Enrich/XRSE/, est ouverte. Le lecteur est positionné sur la dépendance numéro 7112, de type 0 (NOM ADJ) qui apparaît 189 fois dans le corpus. Sa forme est acier inoxydable analysée comme : /frNom{Mopho=4;Csem=42} ACIER /adjStd{Mopho=4;Csem=141} INOXYDABLE.

Ensuite on peut lire un bandeau nommé «Action cut» suivi de trois boutons à cocher. C'est dans cette zone que l'on lit ou que l'on définit les actions d'élagage associées à la dépendance. Les trois boutons correspondent aux trois positions d'une dépendance de type NOM PREP NOM. Seulement deux boutons sont actifs pour les dépendances à deux positions, comme celle couramment affichée dans la fenêtre. On trouvera au chapitre 6 les explications concernant les coupures des dépendances syntaxiques aux positions un, deux ou trois.

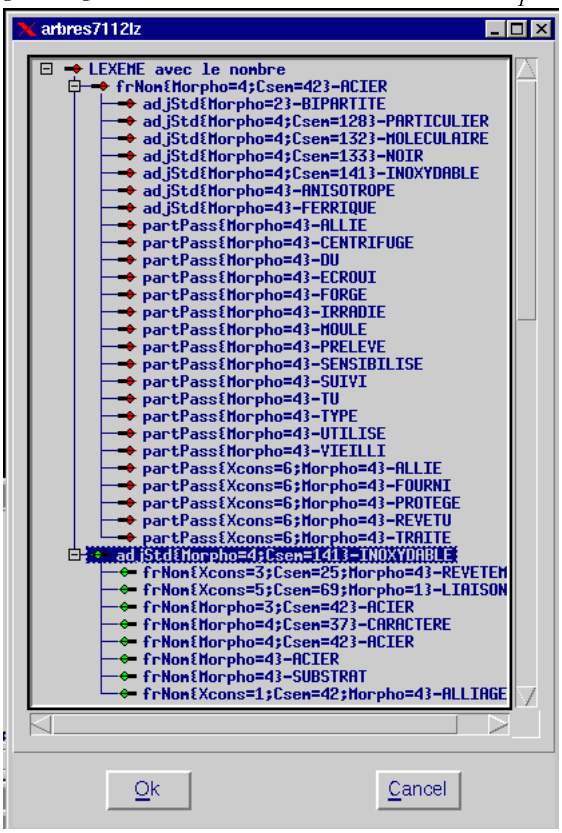
Ensuite, on trouve une zone nommée «informations statistiques». Chaque ligne commence par le mot-clef CATLEX, LEX ou CSEM. Ceux-ci correspondent à des comptages de propriétés pour la généralisation de dépendances sous forme de schémas productifs. Pour les dépendances à deux positions cela consiste à remplacer le contenu de chaque position l'une après l'autre par un autre lexème qui a des proriétés communes avec le lexème originel. Le mot-clef CATLEX signifie que le lexème substitué a en commun la catégorie lexicale (par exemple ADJECTIF), le genre et le nombre. Le mot-clef LEX signifie que le lexème substitué a seulement en commun la catégorie lexicale. Enfin, Le mot-clef CSEM signifie que le lexème substitué a la catégorie sémantique du lexème originel. On trouve également sur chaque ligne une information Card=n et NOcc=n. Ces sont des comptages qui distinguent la cardinalité d'un ensemble de modifieurs ou de modifiés du nombre d'occurrence de ces modifieurs et modifiés dans le corpus.

Prenons l'exemple de la fenêtre principale pour expliciter cela : pour la dépendance acier inoxydable, la première ligne indique qu'il y a dans le corpus 313 schémas de type acier ADJ, soit 313 adjectifs modifiant le nom acier (NOcc=313); mais sur ces 313 adjectifs, il n'y en a que 26 différents (Card=26). On peut donc formuler la chose ainsi : la taille du lexique des modifieurs adjectivaux du nom acier accordé au pluriel est de 26 adjectifs. La seconde ligne indique qu'on dénombre 224 schémas de type NOM inoxydable, c'est-à-dire 224 noms modifiés par l'adjectif inoxydable. Ces 224 occurrences de noms représentent un lexique de 8 noms seulement.

La troisième ligne donne les mêmes informations que précedemment, mais en omettant le genre et le nombre : cela donne 45 adjectifs différents modifiant le nom acier (au singulier ou au pluriel) pour un total de 548 occurrences. En quatrième ligne on peut lire qu'il y a un lexique de 15 noms modifiés par l'adjectif inoxydable (quel que soit son genre et son nombre) pour un total de 368 occurrences dans le corpus.

La cinquième ligne CSEM indique qu'il y a un lexique de 389 adjectifs (pour un total de 1177 occurrences) qui modifient des noms dont la catégorie sémantique est Csem=42 c'est-à-dire des noms de substances élaborées, comme l'acier. La dernière ligne indique qu'il y un lexique de 149 noms (pour un total de 1177 occurrences) qui sont modifiés par des adjectifs dont la catégorie sémantique est Csem=141, c'est-à-dire

FIG. C.2 – *Commutations lexicales sur une dépendance*



des adjectifs qualifiant des substances, comme `inoxydable`.

Pour les dépendances à trois positions du type `vibration de fluide`, les mêmes comptages sont effectués : `NOM1 de fluide`, `vibration de NOM2`, avec et sans information de genre et nombre, puis `Csem(NOM1) de fluide`, `vibration de Csem(NOM2)`. Ensuite les mêmes comptages sont effectués, en relâchant la contrainte sur la forme de la préposition. De cette manière on peut voir si le nom tête accepte des modifieurs prépositionnels introduits par différentes prépositions - donner exemples contrastifs.

Dépendances syntaxiques associées par commutations lexicales Si l'on souhaite connaître le lexique des modifieurs d'un nom ou le lexique des noms modifiés par un adjectif ou un syntagme prépositionnel, la commande «Diversité des têtes-arguments pour `CATLEX` et `LEX`» du menu Affichage fait apparaître une fenêtre comme celle représentée en figure C.2 pour `acier inoxydable`. On voit ainsi dans la figure C.2 à quoi correspond le lexique des 26 adjectifs modifieurs du nom `acier`: {`bipartite + particulier + moléculaire + noir + inoxydable + anisotrope + ...`}. La même fonctionalité pour les dépendances à trois positions permet de visualiser sous forme de grappes les modifieurs du nom introduits par chaque préposition différente.

FIG. C.3 – *Dépendances triées par fréquence décroissante*

C.2.2 Construction assistée de profils

La démarche proposée pour aider à la constitution d'échantillons d'apprentissage pour la construction de profils de sélection est d'associer les actions d'élagage directement aux dépendances lexico-syntaxiques. Pour ce faire, on choisit de commencer par les dépendances lexico-syntaxiques les plus fréquentes du corpus. La figure C.3 présente une fenêtre dans laquelle on trouve un calepin à quatres onglets : «Nombre d'occurrences, CATLEX, LEX, CSEM». Il permet de visualiser respectivement et par ordre de fréquence décroissante : les dépendances lexico-syntaxiques élémentaires, les schémas élémentaires construits sur le mode `CATLEX` (par exemple `NOM`$_1$ `de vapeur`), les schémas élémentaires construits sur le mode `LEX` , les schémas élémentaires construits sur le mode `CSEM` (par exemple `Csem(NOM`$_1$`) de vapeur`).

Lorsque l'on sélectionne une dépendance sur le premier onglet, la fenêtre principale est mise à jour (voir figure C.1) et la barre d'élagage est activée (voir figure C.5); cette dernière est pourvue de boutons qui permettent d'élaguer la dépendance syntaxique en ses différentes positions (voir paragraphe 6.2 du chapitre 6 pour plus d'explications). La fenêtre de visualisation des syntagmes nominaux complets est également mise à jour (voir figure C.4) : la dépendance est ainsi replacée dans son contexte d'extraction. Après un examen des contextes dans lesquels la dépendance prend place, l'opérateur lui associe une action d'élagage. De cette manière, il classe la dépendance dans la catégorie «non pertinente» ou «pertinente à condition d'effacer tel élément lexical ou de couper à tel endroit».

FIG. C.4 – *Visualisation des syntagmes nominaux complets*

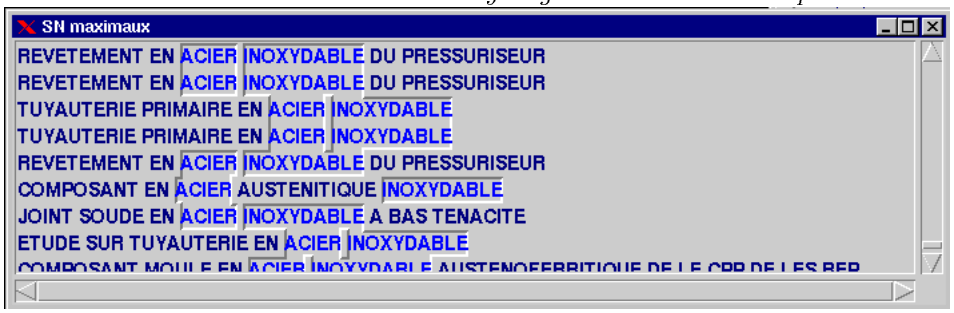


FIG. C.5 – *Boutons d'élagage des dépendances syntaxiques*

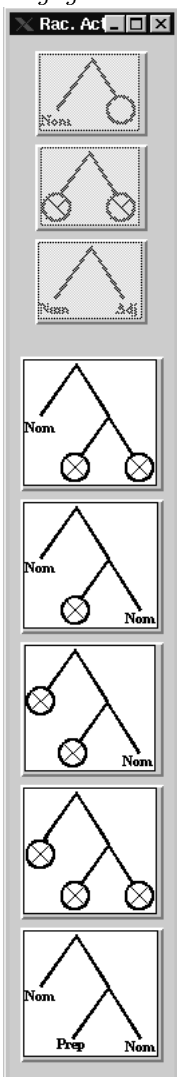
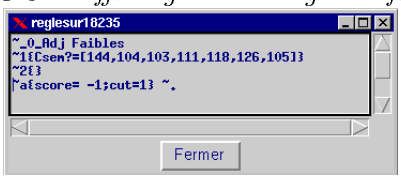


FIG. C.6 – *Affichage d'une règle de filtrage*



Les règles de filtrage et d'exportation

Ces régles décrivent sous forme d'attributs/valeurs des dépendances élémentaires. Elles permettent d'extraire des dépendances de la base courante ou de modifier les scores de celles-ci. La syntaxe des règles/requêtes en est la suivante :

```
~_<type de dépendance> Nom facultatif de la règle ou requête
~1 {Conditions}
~2 {Conditions}
~3 {Conditions}
~a {Actions} ~.
```

La déclaration débute par l'identification du type de dépendance à traiter. Ensuite pour chaque élément de la dépendance, des contraintes à réaliser sont déclarées. Un exemple de règle est donné en figure C.6. Cette règle recherche les dépendaces de type 0 (NOM ADJ) avec un trait sémantique affecté à l'adjectif qui appartient à la liste spécifiée. Dans les blocs conditionnels entre accolades, les tests sont séparés par des points virgules. Chaque test est exprimé par le triplet `trait operateur valeur`. Les traits sont ceux qui proviennent de l'analyse et de l'enrichissement linguistiques. Les valeurs peuvent être du type entier, chaîne de caractères, liste d'entier, liste de chaînes de caractères, arbre ou flottant. Le type arbre correspond à la description d'une hiérarchie sous la forme d'un arbre n-aire. Il est utilisé avec des opérateurs de subsomption pour vérifier si tel trait sémantique est bien subsumé par tel ou tel autre. La table C.1 fait la liste des opérateurs de test et de leur signification en fonction des types de valeurs auquels ils s'appliquent. Les traits sémantiques ont un type mixte : ils sont considérés aussi bien comme des entiers que comme des feuilles d'arbres. Le trait Freq correspond à la fréquence de la dépendance dans le corpus.

Les actions suivent la même syntaxe que les tests aux opérateurs près. Les actions peuvent modifier le score d'une dépendance, modifier l'action d'élagage d'une dépendance et exporter une dépendance vers un fichier. La table C.2 donne la syntaxe des opérateurs.

TAB. C.1 – *Usage et signification des opérateurs de test*

Traits concernés	Opérateurs	Types	Signification
Csem	<	arbre	est subsumé strictement par
Csem	<=	arbre	est subsumé par
Csem	>	arbre	subsume strictement
Csem	>=	arbre	subsume
Csem	? =	liste	appartient
Csem	^=	liste	n'appartient pas
Csem, Xcons, Morpho, Freq	==	chaîne, entier, flottant	égale
Csem, Xcons, Morpho, Freq	!=	chaîne, entier, flottant	est différent de
Freq	<	entier, flottant	est inférieur à
Freq	>	entier, flottant	est supérieur à

TAB. C.2 – *Usage et signification des actions*

Actions	Opérateurs	Type	Commentaire
s	+=	flottant	Ajoute
s	-=	flottant	Retranche
s	=	flottant	Affecte
c	=	entier	Produit l'action d'élégage
e	@	nom de fichier	Exporte la dépendance vers le fichier spécifié

Annexe D

Données techniques

D.1 Traitement des données

D.1.1 Représentation du corpus

La représentation du corpus adoptée permet de :

- considérer chaque phrase du texte concurremment comme une séquence plate (suite de mots) et un arbre syntaxique. Ainsi une modification effectuée dans l'arbre se répercute dans la séquence et vice versa. Cependant si toute modification de la séquence se répercute dans l'arbre, cela se limite strictement aux informations attachées aux feuilles de l'arbre (genre, nombre, catégorie lexicale, propriétés syntaxiques et sémantiques, etc.), cela n'affecte en rien la structure de l'arbre. Il est cependant tout à fait envisageable de soumettre une seconde fois la phrase à un analyseur syntaxique, de manière à prendre en compte les corrections ou les ajouts. Cela n'est cependant pas possible avec l'outil Aleth qui regroupe catégorisation et analyse syntaxique en un processus indissociable.

- s'affranchir du caractère séquentiel du discours (analyse de gauche à droite). Le corpus est numérisé sous forme d'un fichier à accès direct, de format nommé NUM. On peut parcourir le corpus de gauche à droite, de droite à gauche, à partir de n'importe quel point de départ dans le texte, explorer avant et après une phrase donnée.

Le corpus de format NUM dit «numérisé» est une représentation du texte sous la forme d'entiers. Ceci permet d'optimiser les traitements ultérieurs, en minimisant les traitements sur les chaînes de caractères, au profit de traitements sur des entiers. Dans ce fichier, chaque unité lexicale est représentée par un entier qui correspond à une référence dans le dictionnaire général. De même les propriétés associées à cette unité lexicale sont des nombres, correspondant à des propriétés déclarées dans des dictionnaires spécifiques.

Le corpus est subdivisé en documents. Chaque document est segmenté en phrases. Différents documents peuvent être regroupés au sein d'un corpus à l'aide du gestionnaire de corpus.

Compilation du dictionnaire AlethDic Afin d'accéder aux informations du dictionnaire AlethDic, assez volumineux dans son format texte (8Mo), on a dû lui faire subir un traitement pour qu'un accès direct soit possible pour toute unité lexicale pourvu que l'on connaisse sa forme graphique et sa catégorie lexicale. Les informations morphologiques, syntaxiques et sémantiques associées à chaque lexème sont codées sous forme d'entiers, ainsi que les références aux lexèmes. Le tout est enregistré dans une base GNU Dbm (taille finale : environ 7Mo) permettant d'associée une clef à des données en bénéficiant de la technique de de hashage. L'accès à l'information lexicale se fait en donnant une clef qui est la forme graphique et la catégorie lexicale du lexème.

D.1.2 Réutilisabilité logicielle de notre prototype de filtrage

Le prototype a été developpé au fur et à mesure des avancées de la thèse et des résultats des expériences effectuées. Il n'a pas été conçu d'après une vision globale. C'est une des raisons pour lesquelles il pourrait être optimisé, principalement en ce qui concerne les formats de fichiers et certaines structures de données, inutilement encombrées. L'autre raison est que pour garder la possiblité d'exécuter ce prototype sur une machine dotée de peu de mémoire, nous avons minimisé au long du développement la consommation de mémoire vive en préférant le stockage de données intermédaires sur disque. Cette approche pénalise fortement les performances, mais elle autorise le prototype a fonctionner sur une machine PC/Linux dotée 16 Mégaoctets de mémoire vive. Toutefois étant donné que nous utilisons intensivement un système de cache mémoire, les performances s'améliorent avec la quantité de mémoire disponible.

D.1.3 Optimisations peu coûteuses en développement

Abandonner le système de référence au dictionnaire La ressource la plus volumineuse et la plus lourde à gérer est certainement le dictionnaire AlethDic. Celle-ci pourrait être laissée de côté étant donné que l'on en fait une exploitation assez restreinte. Les seules données exploitées pourraient être placées dans des fichiers autonomes, comme c'est le cas pour le lexique sémantique qui se présente sous la forme d'un fichier de règles de désambiguïsation.

Abandonner le type de fichier dit «corpus numérisé» (NUM) Ceci n'est possible qu'en abandonnant le système de référence au dictionnaire. Cela aurait pour effet de supprimer la représentation concurrente du corpus sous formes d'arbres d'analyse et de de séquences «plates». Cela augmenterait les temps de calculs (ana-

lyse de structures arborscentes, lectures récursives d'arbres) mais cela diminuerait notoirement les accès disque.

D.2 Langages utilisés

Tous les traitements sont implémentés en langage C. L'environnement d'écriture des règles de désambiguisation et l'outil de constitution assistée de profils d'apprentissage sont implémenté en C et en Tk/Tcl. Les traitements linguistiques, et l'accès aux dictionnaires et aux corpus sont écrits en langage C. L'interface est réalisée en Tk/Tcl. Le dialogue entre le noyau (supportant les traitements complexes et l'accès aux données) et l'interface se fait en mode client-serveur. L'analyse des arbres syntaxico-sémantique est effectuée au moyen d'une grammaire Lex/Yacc. La librairie GNU Dbm (*Database manager*) est utilisée intensivement pour l'accès aux données stockées sur disque. Cette librairie offre une fonction de cache mémoire paramètrable qui optimise le temps d'accès aux données fréquemment consultées.

Durée des traitements en secondes de CPU Ces estimations ont été réalisées avec la commande Unix `time` sur une sous-partie du corpus d'environ 137 000 mots (ARD-94) avec une machine Sun Ultra Sparc utilisée en `rlogin` depuis une Sparc 20. Une telle configuration accélère les temps de calcul, mais pénalise les accès disque. Ces durées ne rendent compte que du temps de calcul CPU. La durée réelle qui comptabilise les temps entrées/sorties n'est pas pris en compte. Pour donner une estimation, il a fallu par exemple 45 minutes pour effectuer l'étape de désambiguïsation. Les autres durées réelles varient entre 5 et 10 minutes, sauf pour la «Numérisation» qui a durée près de 25 minutes.

Etape	Temps CPU en secondes
1. Normalisation	11
2. Numérisation	46
3. Désambiguïsation	605
4. Répercussion	32
5. Extraction GN	26
6. Extraction Dépendances	46

D.3 Facteurs d'expansion des fichiers

Texte	Document ASCII
GraalDoc	Document balisé au format GraalDoc
AlethIP-Gram	Sorties d'analyses syntaxique du document
IDX	Format normalisé des analyses d'AlethIP
NUM	Document sous forme d'entiers
XGN	Fichier de SN extraits du document
XRSE	Dépendances syntaxiques extraites des SN

Texte ASCII→GraalDoc	1.5x
GraalDoc→AlethIP-Gram	4.2x
AlethIP-Gram→IDX	0.7x
IDX→NUM	1.1x
NUM→XGN	0.5x
XGN→RSE	5x
Echantillon→Profil m0	6x
Echantillon→Profil m1	4.5x
Echantillon→Profil m2	3x
Echantillon→Profil m3	4.5x

Une fois les fichiers de type IDX poduits, les fichiers de type AGRAM peuvent être supprimés. Une fois les fichiers de type XGN poduits, les autres (IDX, NUM) peuvent être supprimés.