Apprentissage de connaissances morphologiques pour l'acquisition automatique de ressources lexicales

Delphine Bernhard

Université Joseph Fourier – Grenoble I **École Doctorale EDISCE**

Apprentissage de connaissances morphologiques pour l'acquisition automatique de ressources lexicales

THÈSE

présentée et soutenue publiquement le 30 novembre 2006

pour obtenir le grade de

Docteur de l'Université Joseph Fourier – Grenoble I

(Spécialité : Sciences cognitives)

par

Delphine BERNHARD

Composition du jury

Président :	Pierre FRATH	Professeur, Université de Reims Champagne-Ardenne
Rapporteurs :	Violaine PRINCE	Professeur, Université de Montpellier II
	Pierre ZWEIGENBAUM	Directeur de recherche CNRS, LIMSI
Examinateurs :	Éric GAUSSIER	Professeur, Université Joseph Fourier-Grenoble I
	Sabine PLOUX	Chargée de recherche CNRS, ISC
	Michel SIMONET	Chargé de recherche CNRS, Laboratoire TIMC-IMAG
	Agnès TUTIN	Maître de conférence, Université Stendhal-Grenoble III

Laboratoire des Techniques de l'Imagerie, de la Modélisation et de la Cognition
TIMC-IMAG — UMR 5525

Remerciements

Le parcours qui m'a menée jusqu'à cette thèse est tout sauf linéaire et doit beaucoup aux enseignants qui m'ont guidée dans mes choix, de la classe de Mathématiques Supérieures à Strasbourg au DEA de Sciences Cognitives à Grenoble. Je voudrais tout particulièrement remercier Pierre Frath, alors Maître de Conférence à l'Université Marc Bloch de Strasbourg, pour le cours de Linguistique Informatique qu'il a dispensé alors que j'étais en Licence d'Anglais. Ce cours a déterminé mes choix ultérieurs et je n'ai jamais eu l'occasion de les regretter.

Je suis reconnaissante à Violaine Prince et Pierre Zweigenbaum pour avoir accepté de rapporter cette thèse. Je tiens également à remercier Pierre Frath, Éric Gaussier et Agnès Tutin pour leur participation au jury.

Je remercie Michel Simonet pour m'avoir accueillie dans son équipe et Sabine Ploux pour avoir co-dirigé cette thèse.

J'ai également bénéficié de nombreux soutiens dans mon équipe. Merci à Sylvain pour m'avoir fait découvrir Python et m'avoir dépannée à plusieurs reprises sous Linux. Je voudrais aussi remercier pour leur bonne humeur mes compagnons de route sur le chemin de la thèse : Lamis, Samer, Gayo, Radja et Houda.

Enfin, je n'oublie pas toutes les personnes de mon entourage, famille et amis, qui m'ont apporté un soutien sans faille. Je voudrais tout particulièrement remercier Olivier pour sa patience au cours de ces années, l'intérêt qu'il a manifesté à l'égard de mon travail et ses relectures attentives de mes articles et mémoires.

Maybe in order to understand mankind, we have to look at the word itself : "Mankind". Basically, it's made up of two separate words : "mank" and "ind". What do these words mean ? It's a mystery, and that's why so is mankind.

Jack Handey
Deep Thoughts

Table des matières

Liste des tableaux

Table des figures

Introduction

L'ordre dans le désordre

La maîtrise de l'information est un enjeu majeur, à une époque où la masse d'informations disponibles ne cesse de croître à un rythme soutenu. Ces informations, qu'il s'agisse d'ailleurs de textes, de sons, d'images ou de vidéos, sont facilement accessibles grâce à Internet. Les progrès dans la diffusion et l'archivage sous forme numérique des informations s'accompagnent toutefois de nombreux problèmes : droit d'auteur, sécurité, caractère éphémère des ressources, multilinguisme, etc. À cela s'ajoute l'absence d'organisation des documents qui complique l'identification des informations pertinentes dans l'énorme volume de données disponibles.

Il est donc nécessaire d'organiser les données pour accéder aux informations d'intérêt : en d'autres mots, le défi consiste à mettre de « l'ordre dans le désordre ». Ce besoin a donné naissance au projet du Web Sémantique qui vise à donner du sens, sous forme de métadonnées, aux documents présents sur Internet et à rendre les machines capables d'effectuer des inférences à partir de ces descriptions sémantiques. Dans ce mémoire, nous allons tout particulièrement nous intéresser aux données textuelles non structurées et non annotées, ce qui exclut les documents XML ou les bases de données.

L'approche la plus ancienne pour catégoriser les informations consiste à décrire les concepts d'un domaine dans des ressources construites manuellement par des experts, comme par exemple les terminologies, les thésaurus ou les ontologies. Ces ressources lexico-sémantiques sont généralement structurées de manière à rendre explicites les liens sémantiques qui existent entre les notions décrites. De plus, elles constituent des listes de référence, offrant un vocabulaire contrôlé rendant possible une description uniforme des informations, par l'élimination des cas de polysémie et de synonymie. Le processus de construction de telles ressources est toutefois long et coûteux, et soumis à de nécessaires révisions à intervalles réguliers, afin de prendre en compte les nouveaux termes et concepts du domaine décrit. Ceci explique pourquoi les ressources construites manuellement ne sont généralement mises en œuvre que pour des sous-domaines restreints comme la médecine.

Nous assistons actuellement à l'émergence d'une nouvelle méthode de catégorisation des données, visant à combler ces lacunes : il s'agit d'une catégorisation collaborative, effectuée de manière spontanée par les usagers et les auteurs eux-mêmes, appelée folksonomie [Sterling, 2005, Wikipedia, 2006a]. La folksonomie consiste à décrire des documents, comme les pages Web ou les images, par des mots clés appelés tags. Ces tags ne sont pas issus de vocabulaires contrôlés et permettent l'accès aux données grâce à une simple recherche par mots clés. Bien sûr, dans la mesure où cet étiquetage des données n'est pas effectué par des experts, il peut être imparfait. Il n'en reste pas moins que cette méthodologie est utile. Elle combine les capacités des logiciels capables de classer et de gérer les informations et la bonne volonté des utilisateurs d'Internet qui effectuent gratuitement un travail de catégorisation. De plus, les mots clés d'une folksonomie ne

sont pas figés mais peuvent évoluer et augmenter en fonction des besoins des utilisateurs et de l'évolution d'un domaine.

Cette approche est toutefois limitée : les folksonomies ne sont pas systématiques et ne permettent généralement pas d'accéder à des informations précises. De plus, elles sont surtout adaptées à la description des documents non techniques, sous forme de mots clés appartenant au vocabulaire commun, ce qui exclut les domaines de spécialité.

Il reste une dernière possibilité pour organiser les connaissances et les données : celle de se passer partiellement ou entièrement de l'intervention humaine, qu'elle soit experte ou non. Pour ce faire, il faut disposer de logiciels capables d'extraire les connaissances pertinentes. Ces connaissances se trouvent naturellement décrites dans les textes, qui constituent de fait un matériau de base pour l'extraction des connaissances. Les méthodes de fouille de textes (*text mining*) visent à donner une structure aux données textuelles en y recherchant les concepts spécifiques au domaine cible et les relations sémantiques qu'ils entretiennent. L'approche que nous présentons dans notre thèse relève de ce domaine et vise à acquérir automatiquement des connaissances (et *in fine* des ressources lexicales) à partir de textes.

Les textes ont une structure qui leur est propre, reposant sur divers niveaux linguistiques : graphèmes (lettre ou suites de lettres), morphèmes, mots, syntagmes, phrases et paragraphes. Cette structure est à distinguer de la structure qu'il est possible d'apporter aux textes après le processus de fouille de texte.

Tous ces niveaux ne bénéficient pas du même traitement. Les travaux en fouille de textes privilégient le niveau du mot ou des groupes de mots et s'intéressent aux relations syntagmatiques et paradigmatiques que ces unités entretiennent dans le texte. Notre approche se différencie des méthodes classiques car nous avons choisi de nous intéresser plus particulièrement au niveau morphologique, qui est celui des plus petites unités porteuses de sens. Nous allons expliciter ce choix dans la section suivante.

Objectifs

Nous avons intégré l'analyse morphologique dans le processus global d'acquisition automatique de ressources lexicales à partir de textes, et notamment l'identification des termes d'un domaine et l'extraction de relations sémantiques entre termes. En effet, dans les domaines techniques, de nombreux termes ont une structure morphologique complexe. De cette structure morphologique dépend généralement l'interprétation sémantique du terme.

Il est bien évident que pour vérifier l'utilité de la morphologie pour l'acquisition de termes et de relations sémantiques il nous a tout d'abord fallu disposer d'un système capable d'effectuer une analyse morphologique complète des termes considérés. Pour atteindre cet objectif, nous avons opéré certains choix méthodologiques, que nous allons expliciter dans la suite de cette introduction.

Méthodologie et matériel

Le cadre général de notre approche peut être résumé par les points suivants : travail sur corpus, langue de spécialité, apprentissage non supervisé et indépendance aux langues. Ceci rejoint les méthodologies généralement utilisées en fouille de texte.

Travail sur corpus

On peut définir un corpus comme un grand ensemble de textes électroniques sélectionnés et variés. Dans cette simple définition se trouvent regroupées trois questions complexes [Péry-Woodley, 1995] :

- taille : quelle est la bonne taille pour un corpus et comment la mesurer (nombre de textes, nombre d'occurrences de formes) ? Les méthodes statistiques nécessitent généralement de gros corpus de textes, contenant plusieurs millions d'occurrences de mots.
- texte : textes entiers ou échantillons de textes de taille constante ?
- choix : comment sélectionner les textes entrant dans la composition du corpus pour qu'il soit représentatif de la langue (ou de la variété de langue) à étudier ? Cette dernière question est centrale dans le domaine de la linguistique de corpus. D'aucuns diront en effet que le Web, une collection d'articles de journaux ou de textes littéraires ne constituent pas des corpus représentatifs de la langue générale car n'étant pas suffisamment diversifiés et caractéristiques de la langue.

Les corpus les plus courants sont monolingues et composés de documents écrits. On trouve également des corpus comparables, contenant des sous-corpus représentatifs de diverses langues, et des corpus parallèles contenant les traductions des mêmes documents dans diverses langues. Certains corpus intègrent des documents oraux, que ce soit sous la forme d'enregistrements ou de transcriptions écrites (c'est le cas par exemple de la composante orale du British National Corpus).

Le travail sur corpus permet l'utilisation de méthodes empiriques basées sur l'apprentissage et l'acquisition automatique. En effet, l'avènement de la linguistique de corpus (traduction du terme anglais « Corpus Linguistics »), rendu possible par les progrès techniques qui ont augmenté à la fois les capacités de stockage et de traitement, s'est accompagné d'une renaissance des méthodes statistiques d'analyse des données textuelles [Church et Mercer, 1993].

Nous avons utilisé des corpus de textes acquis automatiquement à partir d'Internet. En effet, Internet constitue la plus grande source de textes électroniques non structurés ou faiblement structurés. La méthode de construction de corpus à partir d'Internet est détaillée dans le Chapitre 3.

Langue de spécialité

Le choix de travailler sur le vocabulaire propre à des domaines spécifiques, essentiellement la médecine, s'est naturellement imposé du fait de la thématique de notre laboratoire centrée sur l'ingénierie médicale. De plus, de nombreuses applications du traitement automatique des langues sont destinées à des domaines spécifiques, dont le vocabulaire est généralement restreint comme le tourisme [Berger *et al.*, 2004] ou la médecine.

[Kittredge, 2003, p. 432] donne une définition de la notion de sous-langage sous forme de deux pré-conditions :

> - une communauté de locuteurs (c'est à dire des 'experts') partage une connaissance spécialisée d'un domaine sémantique restreint ;
> - les experts communiquent au sujet du domaine restreint dans des situations récurrentes ou un ensemble de situations très similaires.
>
> Lorsque les énoncés (y compris les écrits) des experts du domaine présentent des patrons systématiques qui les distinguent du langage pris dans sa globalité, on dit que ces énoncés appartiennent à un sous-langage [1].

[1] Notre traduction.

Le vocabulaire spécifique aux domaines spécialisés complique l'utilisation de certains outils comme les lemmatiseurs ou étiqueteurs morpho-syntaxiques développés pour la langue générale. Un sous-langage ne se distingue toutefois pas uniquement par son vocabulaire mais également par des associations particulières entre mots. Par conséquent, un sous-langage présente généralement des fréquences d'occurrence et de co-occurrence de certains mots bien différentes d'un corpus général. Ces différences peuvent être mises à profit pour identifier les termes spécifiques d'un domaine. La morphologie des termes d'un sous-langage est également marquée par l'utilisation de morphèmes et de procédés de formation de mots caractéristiques. Par exemple, en médecine, l'utilisation des préfixes, suffixes et éléments de formation suivants est particulièrement fréquente : *anti+*, *micro+* , *semi+*, *+ite*, *+gramme*, ... [Heyer *et al.*, 2006].

Cependant, si l'on se limite à un seul domaine particulier, les méthodes élaborées risquent de n'être valables que pour ce domaine. Nous avons donc choisi de travailler sur deux domaines distincts, l'un relevant des sciences de la terre (la volcanologie) et l'autre de la médecine (le cancer du sein). Cette approche comparative de deux domaines spécialisés différents devrait permettre une confrontation des résultats obtenus et la vérification de l'applicabilité de nos méthodes à divers domaines.

Apprentissage et approche statistique

L'essor, ou plutôt la renaissance, des méthodes empiriques et statistiques est justifié par les résultats de recherches en psychologie cognitive qui montrent que l'apprentissage d'une langue par les êtres humains fait largement appel aux propriétés statistiques des langues et ne serait donc pas lié à des règles pré-câblées présentes dans le cerveau humain [Seidenberg, 1997, Perruchet et Peereman, 2005].

Les méthodes statistiques sont en réalité très anciennes et utilisées depuis des siècles, avant même l'apparition des ordinateurs. Par exemple, les premières méthodes de cryptanalyse, pour le déchiffrement des codes secrets, reposent sur l'analyse de la fréquence des lettres [Singh, 1999]. On a retrouvé une description de cette technique datant du IX^{e} siècle dans un manuscrit du savant Arabe Al-Kindi.

Les méthodes d'apprentissage se divisent en apprentissage supervisé et apprentissage non supervisé. Dans le premier cas, le système « apprend » à partir d'un ensemble d'exemples de résultats attendus. Dans le second, le système apprend à partir des données à analyser, sans règles spécifiques ou données externes au corpus. Les méthodes d'apprentissage non supervisé ont donc un coût réduit par rapport à la construction manuelle de ressources. Même lorsque les ressources sont disponibles, elles nécessitent des mises à jour fréquentes, et ce surtout dans les domaines techniques qui sont en constante évolution. Les méthodes d'apprentissage non supervisé peuvent donc permettre de construire automatiquement ou semi-automatiquement des ressources et d'enrichir les ressources déjà existantes.

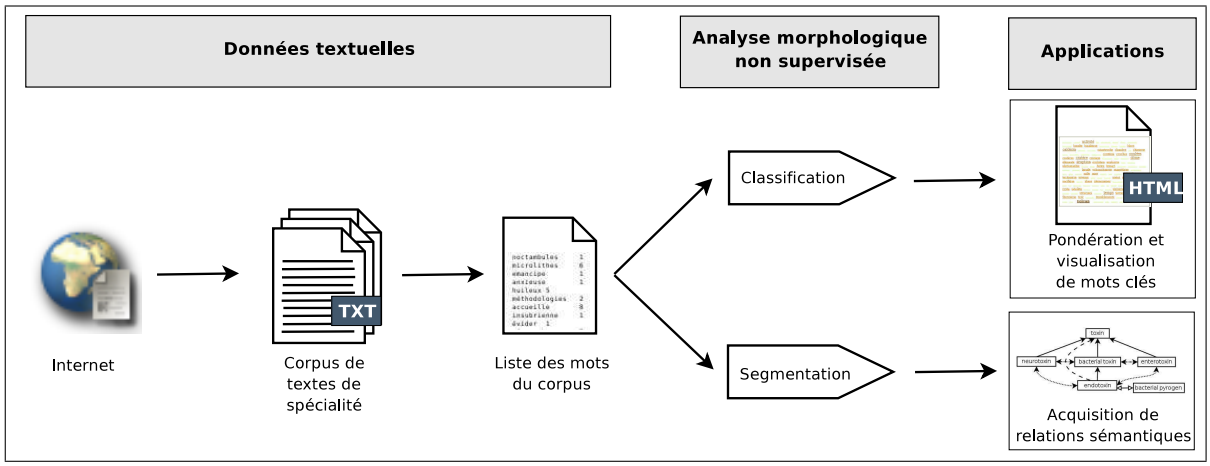
Indépendance aux langues

Une des propriétés souhaitables, qui découlent des principes méthodologiques énoncés précédemment, est l'indépendance aux langues des outils d'analyse développés. Indépendance aux langues ne rime toutefois pas avec universalité : il est illusoire de penser qu'un même outil d'apprentissage puisse être directement utilisable pour toutes les langues existantes. Tout outil se base sur un certain nombre d'hypothèses générales qui restreignent de fait son champ d'utilisation. Ainsi, pour les systèmes d'analyse morphologique que nous allons présenter dans les Chapitres 4 et 5, nous avons fait l'hypothèse que les mots sont formés par *concaténation*

linéaire de morphèmes. Cette contrainte englobe de nombreuses langues européennes comme le français, l'allemand, l'anglais, le finnois ou le turc, mais exclut des langues dont les mécanismes de structuration morphologique sont différents, comme l'hébreu ou l'arabe dont les mots sont formés à partir d'une racine consonantique où vont s'intercaler les voyelles.

Organisation du mémoire

L'organisation du mémoire suit le schéma global de la figure présentée ci-dessous, à l'exclusion des deux premiers chapitres consacrés à l'état de l'art.



La première partie du mémoire décrit le contexte général de la recherche. Nous dressons tout d'abord un panorama des approches utilisées en acquisition automatique de ressources lexicales. Nous décrivons en particulier les méthodes d'identification des termes d'un domaine et des relations sémantiques qu'ils entretiennent. Puis, nous présentons le domaine de la morphologie à travers le point de vue de diverses disciplines : linguistique, psychologie cognitive et traitement automatique des langues.

La deuxième partie est consacrée aux systèmes d'analyse morphologique non supervisée que nous avons développés. Ces systèmes procèdent à l'apprentissage de la morphologie d'une langue à partir d'une liste de mots extraite d'un corpus. Nous décrivons donc tout d'abord l'étape préliminaire d'acquisition des données d'apprentissage par la construction automatique de corpus à partir d'Internet.

Le premier système que nous proposons procède à la segmentation des mots en unités étiquetées. Le deuxième système relève d'une démarche inverse : les mots sont groupés en familles morphologiques par un processus de classification ascendante hiérarchique.

La dernière partie présente des utilisations possibles des résultats de ces systèmes pour l'acquisition automatique de ressources lexicales. La première application consiste en la pondération de mots clés via leur famille morphologique. Cette pondération est ensuite utilisée pour projeter les données sur une carte au format HTML. La seconde application évalue la possible utilisation des résultats de la segmentation pour l'identification de relations sémantiques entre termes morphologiquement complexes.

Première partie

Contexte de la recherche

Introduction

Cette première partie présente un état de l'art de l'acquisition automatique de ressources lexicales, suivi d'une description du domaine de la morphologie.

Les ressources lexicales sont des listes de mots accompagnées d'informations qui sont dépendantes du type de la ressource : informations morphologiques, syntaxiques, sémantiques et terminologiques. Lorsque ces ressources décrivent des concepts, associés à un niveau lexical explicite, nous parlerons de ressources lexico-sémantiques. Les ressources lexicales peuvent être construites manuellement ou automatiquement. L'automatisation de certaines tâches permet de réduire le coût et la durée de la construction. Nous détaillons les méthodes d'acquisition automatique de connaissances lexicales et sémantiques dans le Chapitre 1. Nous présentons en particulier deux des tâches principales : l'identification des termes ou mots clés d'un domaine et l'acquisition de relations sémantiques entre termes. La description des diverses méthodes laisse entrevoir les possibilités offertes par la morphologie pour l'exécution de ces tâches.

Le domaine de la morphologie est décrit dans le Chapitre 2 à travers les points de vue de diverses disciplines du champ des sciences cognitives : linguistique, psychologie cognitive et traitement automatique des langues. Cette présentation vise à mettre en évidence les propriétés et spécificités du niveau morphologique. Celles-ci sont utilisées dans les systèmes d'analyse morphologique décrits dans la deuxième partie du mémoire.

Chapitre 1

Acquisition automatique de ressources lexicales à partir de textes

1.1 Introduction

La construction manuelle de ressources lexicales est une tâche fastidieuse et dont les coûts humains sont importants. À cela s'ajoute l'évolution constante du vocabulaire des domaines à modéliser, surtout lorsqu'il s'agit de domaines techniques ou spécialisés comme la médecine. De plus, les ressources ne sont pas toujours disponibles pour les langues minoritaires ou moins bien étudiées que l'anglais. Le recours à des méthodes automatiques pour la construction ou l'aide à la construction manuelle de ressources lexicales se justifie donc aisément. L'acquisition automatique de ressources comme les thésaurus et autres vocabulaires structurés se divise en diverses sous-tâches comme l'identification des unités linguistiques représentant les concepts du domaine ou la structuration de ces unités par des relations sémantiques. Les corpus de textes utilisés sont généralement spécifiques à un domaine technique restreint.

Les utilisations possibles des résultats sont multiples :

- Aide à la construction manuelle de ressources comme les thésaurus, les réseaux sémantiques ou les ontologies ;
- Indexation ;
- Recherche d'informations : extension de requête avec des mots sémantiquement proches ;
- Désambiguïsation sémantique.

Dans les sections suivantes nous allons tout d'abord faire un tour d'horizon des différents types de ressources visant à représenter et structurer le vocabulaire et les concepts. Puis, nous allons décrire les tâches principales visant à extraire des textes les connaissances utiles à leur acquisition.

1.2 Ressources lexico-sémantiques

Les ressources qui répertorient, définissent et organisent les connaissances lexicales et conceptuelles sont très diverses et peuvent être classées en fonction de différents critères :

- Unités : mots, mots clés, termes, sens, concepts ;
- Domaine : général ou spécialisé ;
- Niveau de formalisation : formel ou informel ;
- Support : électronique ou non ;
- Structuration : linéaire, en réseau ou hiérarchisée.

Le niveau de structuration et de formalisation est généralement corrélé à la technicité du domaine décrit. En médecine, de nombreuses ressources termino-ontologiques, structurées hiérarchiquement, sont disponibles, comme le MeSH (Medical Subject Headings)[1] ou UMLS (Unified Medical Language System)[2] .

1.2.1 Types de ressources lexico-sémantiques

Nous détaillons ci-dessous les principaux types de ressources lexico-sémantiques, mais il faut savoir qu'il n'y pas de frontière stricte entre ces notions.

Lexiques Les lexiques sont des listes de mots, généralement triés par ordre alphabétique, parfois accompagnés de leur définition ou de leur traduction.

Glossaires Les glossaires listent les définitions des termes spécifiques à un domaine.

Dictionnaires Les dictionnaires listent les mots et leur définition ou leur traduction (dictionnaires bilingues). Les mots sont classés en fonction de leur lemme (forme de base, non fléchie, comme l'infinitif pour les verbes ou le singulier pour les noms en français).

Thésaurus Les thésaurus structurent de manière hiérarchisée des notions de la langue générale ou de domaines spécifiques. [Matsumoto, 2003] distingue deux types de thésaurus : les thésaurus de classement (*classification thesaurus*) et les thésaurus taxinomiques (*taxonomic thesaurus*). Les thésaurus de classement groupent les mots dans des catégories en fonction de leur similarité sémantique. Ces catégories sont organisées hiérarchiquement et correspondent à des notions de plus en plus abstraites en fonction de leur hauteur dans l'arbre des catégories. Les mots sont toujours des feuilles (nœuds terminaux) dans une telle hiérarchie. Le thésaurus anglais *Roget's* [Roget, 1911] est un exemple de thésaurus de classement. Les thésaurus taxinomiques n'utilisent pas de niveau catégoriel abstrait et les mots peuvent apparaître à tout niveau de la hiérarchie. WordNet [Miller, 1995] est un exemple de ce type de thésaurus. Les thésaurus sont notamment utilisés pour l'indexation automatique et la recherche d'information. En complément des relations sémantiques hiérarchiques, ils peuvent également comporter des liens transversaux vers des catégories proches (liens *Voir aussi* ou *See also*).

Terminologies Les terminologies sont des classifications normalisées des termes et notions d'un domaine.

Ontologies Les ontologies organisent des concepts, généralement sous forme hiérarchique, et doivent permettre de faire des inférences [Vossen, 2003]. Les ontologies sont souvent représentées en utilisant les logiques de description qui permettent la vérification de contraintes ou le classement automatique en fonction des descriptions associées à chaque concept. Elles ne contiennent que très peu d'informations lexicales (seul un représentant explicite en langue naturelle pour chaque concept). OWL (Web Ontology Language) est le langage de représentation d'ontologies le plus utilisé à l'heure actuelle. Il n'existe cependant pas de réel consensus sur la notion d'ontologie. De plus, l'encodage de connaissances utilisant un formalisme rigoureux est un processus long et coûteux, dont les bénéfices sont difficiles à mesurer. Ainsi des thésaurus ou des terminologies pourront dans certains cas être considérés comme des ontologies même si leur niveau de formalisation est moindre. Dans le cas d'applications opérant sur des textes comme l'indexation ou la recherche d'information, il est nécessaire d'avoir un niveau lexical explicitement rattaché au niveau conceptuel.

[1] `http://www.nlm.nih.gov/mesh/`
[2] `http://umlsinfo.nlm.nih.gov/`

Le système WordNet se situe en réalité à la frontière du dictionnaire, du thésaurus, du réseau sémantique et de l'ontologie. WordNet est gratuit et régulièrement mis à jour, ce qui en fait une ressource très appréciée en traitement automatique des langues. WordNet contient des synsets (groupes de synonymes) qui représentent chacun un concept.

1.2.2 Unités lexicales et conceptuelles

Les différentes ressources que nous avons listées diffèrent par les unités qu'elles décrivent qui sont soit lexicales, soit conceptuelles, soit encore lexico-conceptuelles, comme les termes.

Mots clés

En recherche d'information, les mots clés sont les mots qui décrivent le mieux le contenu d'un document ou d'un corpus. Les mots clés sont souvent des noms, des verbes ou des adjectifs, par opposition aux mots outils comme les prépositions, les déterminants ou les pronoms.

En linguistique de corpus, les mots clés sont les mots qui apparaissent plus fréquemment dans un document que ne le voudrait le hasard (voir Section 1.3.3, p. 22). De nombreuses mesures reposant sur les différences de fréquence d'occurrence permettent ainsi d'extraire automatiquement les mots-clés d'un document ou d'un corpus.

Termes

Le terme est défini de la manière suivante par [Roche, 2005] :

> Élément d'une terminologie. Combinaison indissociable d'un concept et d'une dénomination (désignation).

Du point de vue classique, celui de E. Wüster et du Cercle de Vienne, le terme est la dénomination d'un concept, chaque concept étant désigné de manière non ambiguë par un seul terme [Jacquemin et Bourigault, 2003]. La dénomination ou désignation[1] d'un terme peut prendre diverses formes linguistiques. La distinction la plus courante est faite entre **termes simples** ou **monolexicaux** et **termes complexes** ou **polylexicaux**. Dans le premier cas, le terme est composé d'un seul mot graphique, dans le second d'une succession de mots.

Le terme est donc un élément construit, dont le statut est différent des autres mots de la langue car il répond à un besoin de normalisation sémantique. De plus, on considère généralement que les termes doivent être monosémiques dans le domaine considéré[2] .

Ainsi, les méthodes d'acquisition automatique de connaissances présentées dans les sections suivantes concernent l'acquisition de **candidats termes** qui doivent encore être validés et normalisés par les terminologues au cours d'un processus qui ira du mot ou de l'expression au concept [Rastier, 1995, Jacquemin et Bourigault, 2003].

[1] La différence entre les deux notions correspond à une différence d'approche : la dénomination correspond à une approche onomasiologique allant du concept vers le mot tandis que la désignation relève d'une approche sémasiologique allant du mot vers le concept.

[2] Cette exigence est mise à mal par des études récentes qui tendent à montrer que les mots les plus spécifiques à un corpus technique, correspondant aux termes monolexicaux, ne sont pas les plus monosémiques [Bertels *et al.*, 2006]

Unités de sens

Il existe divers types d'unités de sens dans les ressources lexicales :

- **Concepts ou catégories conceptuelles** : un concept est la représentation mentale d'un ensemble d'objets différents, mais considérés comme équivalents d'un certain point de vue (nom identique, action commune, etc.). Les concepts ne se trouvent pas directement dans les textes. En effet, comme le constate très justement C. Roche [Roche, 2005], « Il n'y a pas de concepts dans un texte, mais uniquement des traces linguistiques de leurs usages ».
- **Classes** : la notion de classe est utilisée en informatique pour décrire un ensemble d'objets ou d'entités partageant un certain nombre de propriétés. On retrouve également ce terme en linguistique, où les classes d'objets sont des « classes sémantiques construites à partir de critères syntaxiques, chaque classe étant définie à partir des prédicats qui sélectionnent de façon appropriée les unités qui la composent »[Le Pesant et Mathieu-Colas, 1998, p. 6].
- **Synsets** : la notion de synset est spécifique à WordNet [Miller, 1995]. Dans cette base de données, les noms, verbes, adjectifs et adverbes sont organisés en ensemble de synonymes, les synsets, qui représentent des concepts lexicalisés.

1.2.3 Relations sémantiques

Après avoir passé en revue les différents types d'unités décrites dans les ressources lexico-sémantiques, nous allons maintenant décrire les relations sémantiques établies entre ces unités. Ces relations sont distribuées sur deux axes :

- **Axe syntagmatique** (horizontal). Deux mots sont en relation syntagmatique s'ils apparaissent ensemble dans un texte : il s'agit donc d'une relation *in præsentia*. On dit également que les mots sont co-occurrents s'ils apparaissent ensemble dans un contexte restreint. Les relations sémantiques définies sur l'axe syntagmatique sont de type associatif comme par exemple entre *tasse* et *café* (La *tasse* contient du *café*) ou *chat* et *lait* (le *chat* boit du *lait*).
- **Axe paradigmatique** (vertical, hiérarchique). Deux mots sont en relation paradigmatique s'ils apparaissent dans des contextes similaires : il s'agit donc d'une relation *in absentia*. C'est à ce niveau que l'on retrouve un certain nombre de relations structurant le lexique telles que la méronymie et l'hyponymie.

En psychologie, et notamment dans les études sur la formation des concepts et des catégories, on distingue deux types de catégories : les catégories taxonomiques et les catégories thématiques [Lin et Murphy, 2001, Nguyen et Murphy, 2003, Wisniewski et Bassok, 1999] :

- Les catégories **taxonomiques** sont organisées en hiérarchies de catégories de plus en plus abstraites comme *bouledogue* < *chien* < *mammifère* < *animal*. Ces catégories sont basées sur des propriétés communes ou la similarité.
- Les catégories **thématiques** groupent des objets qui sont associés ou qui ont une relation de complémentarité (les entités ne jouent pas le même rôle). On trouve différents types de relations thématiques : spatiale (un *toit* se trouve sur une *maison*), fonctionnelle (un morceau de *craie* sert à écrire sur un *tableau noir*), causale (l'*électricité* fait briller l'*ampoule*) et temporelle (le *repas* est suivi de la *note* au restaurant). Celles-ci sont généralement moins représentées dans les ressources lexico-sémantiques.

Nous allons maintenant détailler les différents types de relations sémantiques pour les mots qui sont en relation paradigmatique. Ces relations sont pour la plupart d'entre elles décrites dans les thésaurus et les ontologies.

Relations d'inclusion et d'identité

- **Synonymie.** Selon [Cruse, 2000] les synonymes sont des mots dont les similarités sémantiques sont plus saillantes que les différences. Il est alors possible de distinguer différents degrés de synonymie : synonymie absolue (identité de sens, ce qui est très rare), synonymie propositionnelle (les termes peuvent se substituer l'un à l'autre dans un contexte linguistique particulier sans altérer les conditions de vérité de la phrase) et synonymie proche (comme par exemple *mist* et *fog*). Les termes synonymes correspondent au même concept. La relation de synonymie est symétrique, mais pas nécessairement transitive [Lafourcade et Prince, 2001].
- **Hyponymie.** La relation d'hyponymie (encore appelée subsomption, spécialisation, relation ISA ou EST-UN) implique un rapport d'inclusion entre les sens des mots. Par exemple, *pomme* et *pêche* sont des **hyponymes** de *fruit* et *fruit* est un **hyperonyme** de *pomme* et de *pêche*. On dit également que *pomme* et *pêche* sont subsumés sous *fruit* et qu'ils sont tous deux **co-hyponymes** de *fruit*. La relation de spécialisation est transitive et non symétrique.
- **Méronymie.** La relation de méronymie (aussi appelée relation PART-OF ou PARTIE-DE) correspond à la relation partie-tout. Ainsi, *globule* est un **méronyme** de *sang* et *sang* est un **holonyme** de *globule*.

Relations d'exclusion et d'opposition

- Deux termes co-hyponymes peuvent avoir des sens **incompatibles**, c'est-à-dire qu'ils ne peuvent être vrais en même temps, comme par exemple *chien*, *chat*, *souris* ou *lion* qui sont tous des hyponymes d'*animal*.
- Deux termes sont **complémentaires** si l'un implique le contraire de l'autre, comme par exemple *mort* et *vivant*.
- Deux termes qui sont des **antonymes** stricts appartiennent à la même catégorie syntaxique et sont opposés sur un axe gradué, comme par exemple *long* vs *court*, *chaud* vs *froid*, *bon* vs *mauvais*. On peut considérer l'**antonymie** comme « un cas particulier de la relation de **co-hyponymie** »[Amsili, 2003, p. 37]. Cette relation est notamment utilisée dans WordNet pour l'organisation des adjectifs [Hayes, 1999].

1.2.4 Conclusion

Nous venons de faire un bref tour d'horizon des différents types de ressources lexico-sémantiques et de leur contenu. Les ressources les plus complexes intègrent des informations d'ordre sémantique permettant de structurer les mots ou les concepts. Nous avons déjà évoqué dans l'introduction de ce chapitre le coût de construction élevé de telles ressources, auquel s'ajoutent les nécessaires mises à jour pour rendre compte de l'évolution du domaine décrit. De nombreuses méthodes d'acquisition automatique de connaissances à partir de textes visent à faciliter ce processus de construction, voire à acquérir automatiquement des connaissances susceptibles de remplacer les ressources lexico-sémantiques pour l'exécution de certaines tâches. Nous allons décrire les diverses méthodes existantes dans la section suivante, en commençant par l'acquisition automatique de termes et de mots clés.

1.3 Acquisition automatique de termes et de mots clés

On peut distinguer deux types de méthodes d'acquisition automatique de termes et de mots clés :

– Les méthodes à base de patrons définissant la structure des termes à extraire.

– Les méthodes à base de calculs statistiques (mesures d'association et de comparaison).

Ces deux types de méthodes ne s'excluent pas mutuellement et peuvent être combinées pour obtenir de meilleurs résultats.

1.3.1 Méthodes à base de patrons

Les termes polylexicaux peuvent être caractérisés par des patrons reposant essentiellement sur l'étiquetage morpho-syntaxique. Ce pré-requis n'est toutefois par indispensable car certains systèmes, comme celui proposé par J. Vergne ou le système ANA, fonctionnent à partir de corpus de textes bruts, non étiquetés. Les patrons peuvent également être définis à partir de la structure morphologique des termes, permettant ainsi l'identification de termes simples (composés d'un seul mot graphique) morphologiquement complexes.

Patrons morpho-syntaxiques

Cette famille de méthode nécessite deux types d'informations préalables : l'étiquetage morpho-syntaxique du corpus ainsi qu'un ensemble de patrons reposant sur cette étiquetage et décrivant la structure des termes que l'on cherche à extraire. Les termes ainsi identifiés sont généralement des groupes nominaux [Kageura et Umino, 1996].

Le système LEXTER[1] développé par D. Bourigault [Jacquemin et Bourigault, 2003] est un analyseur syntaxique robuste dédié à l'extraction de syntagmes (nominaux et adjectivaux) à partir de corpus spécialisés, dans une perspective d'acquisition terminologique. Il procède en deux étapes pour extraire les unités terminologiques. Dans un premier temps, il repère les groupes nominaux maximaux en se basant sur des règles permettant d'identifier les limites de syntagmes nominaux les plus vraisemblables. Puis il décompose ces groupes nominaux afin d'en extraire les termes candidats. De plus, les termes candidats sont organisés sous forme de réseau en fonction des éléments lexicaux partagés dans des positions syntaxiques similaires.

Le système ACABIT (Automatic Corpus-based Aquisition of BInary Terms) [Daille, 1996] a pour objectif de préparer la tâche du terminologue en lui proposant une liste ordonnée de candidats-termes pour un corpus préalablement étiqueté et lemmatisé. Les candidats-termes correspondent à un type particulier de co-occurrences où sont prises en compte les nombreuses variations des termes : variations flexionnelles et syntaxiques faibles, variation de modification interne, variation de coordination et variations attributives. Le candidat-terme présenté à l'expert est une forme générique regroupant les différentes occurrences du candidat-terme rencontré dans le corpus sous sa forme de base ou sous la forme d'une de ses variations. Les candidats-termes sont classés suivant un score d'association. Cette méthode ne fait donc pas uniquement appel à des filtres linguistiques permettant de repérer certains types de syntagmes nominaux mais utilise également des mesures statistiques.

[1] Un nouvel analyseur, SYNTEX, a pris la suite de LEXTER.

La méthode de Jacques Vergne

Contrairement aux systèmes que nous venons de présenter, la méthode proposée par J. Vergne ne nécessite aucun étiquetage morphosyntaxique préalable des textes. Elle permet d'extraire des termes de structure contrôlée par des patrons reposant sur l'alternance dans le texte de mots informatifs et non informatifs. Ces catégories sont définies de la manière suivante [Vergne, 2005], selon la distinction introduite par Lucien Tesnière :

> Les mots informatifs sont les mots pleins ou lexicaux (content words), et les mots non-informatifs sont les mots vides, ou grammaticaux (function words).

A cette définition linguistique, J. Vergne fait correspondre des indices facilement mesurables en corpus : « un mot informatif est plus long et moins fréquent que ses voisins », reprenant ainsi les principes de Zipf (« ce qui est d'usage fréquent est court ») et Saussure (« dans la langue, il n'y a que des différences »). La méthode ne faisant usage d'aucune ressource externe au corpus est à la fois endogène et multilingue [Vergne, 2003]. Elle évite donc le recours à une *stop list* (liste de mots outils), nécessairement propre à une langue, et potentiellement ambiguë car contenant des mots correspondant à des homographes qui peuvent être informatifs ou non informatifs suivant le contexte. Les différentes étapes de la méthode, dans sa version la plus récente [Vergne, 2005], sont détaillées dans la Figure 1.1.

Données	Texte à indexer.
Étape 1	Identification des mots informatifs (mots pleins) et non informatifs (mots vides) à l'aide des différences de longueur et d'effectif entre mots contigus.
Étape 2	Génération de candidats termes de structure contrôlée en utilisant des patrons reposant sur l'étiquetage en mots informatifs (I) et mots non informatifs (n) : $I+$, $I+n+I+$, $I+n+I+n+I+$.
Étape 3	Suppression des termes hapax et des termes inclus dans des termes de même effectif.
Étape 4	Calcul du poids de chaque terme dans le document en fonction de son effectif et de sa longueur.
Résultats	Liste de termes pondérés.

Fig. 1.1: Déroulement de la méthode de J. Vergne pour l'acquisition automatique de termes.

Le système ANA, Apprentissage Naturel Automatique d'un Réseau Sémantique

Le système ANA [Enguehard, 1992, Enguehard, 1993, Enguehard et Pantéra, 1995] se distingue également par l'absence de pré-traitement des corpus. Il est inspiré par l'apprentissage humain de la langue maternelle. Le programme extrait une liste initiale de termes du domaine constituant un « bootstrap », ainsi que des listes de mots fonctionnels et de connecteurs de termes complexes. Dans la phase de découverte, les termes complexes contenant un des termes de la liste de bootstrap sont repérés et leurs composants sont ajoutés à la liste de bootstrap. Ils peuvent alors être réutilisés pour la découverte de nouveaux termes, dans un processus itératif. Les différentes étapes de la méthode sont décrites dans la Figure 1.2. Une méthode similaire est utilisée par [Baroni et Bernardini, 2004].

Données	Corpus de textes bruts.
Familiarisation	Extraction automatique de quatre listes : mots fonctionnels, mots fortement liés, mots de schémas, bootstrap (termes du domaine).
Découverte	Extension de la liste des termes du domaine à partir du bootstrap en utilisant les patrons suivants : - expression : terme constitué de deux termes co-occurrents. - candidat : co-occurrence d'un terme, d'un mot de schéma et d'un mot (nouveau terme). - expansion : co-occurrence d'un terme et d'un mot.
Résultats	Liste de termes.

FIG. 1.2: Déroulement de la méthode ANA pour l'acquisition automatique de termes.

Patrons morphologiques

Les méthodes à base de patrons précédemment décrites définissent la structure lexicale des termes polylexicaux. Or, le vocabulaire de domaines spécifiques, comme la médecine, se caractérise également par des patrons de formation incluant des **segments** de mots spécifiques, comme par exemple le suffixe *-ite* en médecine [Heyer *et al.*, 2006]. Ces spécificités, marquées par l'utilisation d'affixes typiques, peuvent être mises à profit pour l'acquisition de termes. Pour ce faire, il est dans un premier temps nécessaire de repérer les affixes spécifiques au domaine.

[Ananiadou, 1994] propose un système d'analyse morphologique des composés savants conférant le statut de termes aux mots contenant certains suffixes et éléments de formation. Les suffixes sont identifiés manuellement à partir de l'analyse d'un corpus de spécialité. Le système décrit par [Heid, 1998] extrait quant à lui les mots contenant certains préfixes, suffixes et éléments de formation caractéristiques de domaines techniques en allemand comme *mega+*, *mikro+*, *+gramm* ou *+graph*. Certaines listes d'affixes spécifiques à des domaines précis comme la médecine sont disponibles. Il est également possible de les identifier de manière automatique. En effet, ces affixes sont rares dans le vocabulaire général et peuvent donc être identifiés par comparaison de leur fréquence d'occurrence dans un corpus de spécialité par rapport à un corpus de langue générale (voir Section 1.3.3, p. 23). Cette méthode permet également la découverte de radicaux spécifiques. Voici quelques exemples de morphèmes identifiés par comparaison à partir de textes légaux en allemand : *Beratung+*, *Amt+*, *Abnahme+*, *+recht*, *+gericht*, *+betrieb*, *+betrag*.

La méthode d'identification d'éléments spécifiques par comparaison des fréquences est également utilisée par [Witschel, 2005] pour la découverte de trigrammes (suites de trois lettres) spécifiques d'un domaine. Les mots qui contiennent ces trigrammes sont les termes candidats. Cette méthode est surtout efficace pour les domaines techniques dont les termes contiennent des éléments de formation grecs ou latins[1]. On trouve par exemple les trigrammes spécifiques suivant, à partir d'un corpus de textes anglais sur l'asthme : *sth* (*asthma*, *anti-asthmatic*), *uco* (*glaucoma*, *mucosa*), *thm* (*asthma*, *dysrythmia*), *apy* (*therapy*, *immunotherapy*) [Heyer *et al.*, 2006].

[1]Le système est accessible en ligne à l'adresse suivante : `http://wortschatz.uni-leipzig.de/~fwitschel/terminology.html`

Limite des méthodes à base de patrons

Les méthodes à base de patrons sont efficaces pour détecter des unités de structure remarquable dans les documents. Toutes ces unités remarquables ne sont toutefois pas des termes. Les méthodes à base de patrons sont donc souvent complétées par des mesures statistiques prenant en compte la fréquence d'occurrence, comme c'est le cas par exemple dans la méthode de Jacques Vergne. L'utilisation d'une mesure simple comme la fréquence d'occurrence se justifie aisément. En effet, une unité qui apparaît fréquemment dans un document ou un corpus joue très certainement un rôle important dans la terminologie du domaine. D'autres mesures, plus complexes, pourront également être utilisées, en combinaison avec les méthodes à base de patrons ou indépendamment. Nous présentons plusieurs de ces mesures dans les sections suivantes et détaillons tout d'abord les mesures *intrinsèques* d'association, puis les mesures *extrinsèques* de comparaison [Streiter *et al.*, 2003].

1.3.2 Mesures d'association

Ces mesures permettent de quantifier l'information partagée par des couples de mots ou termes et de repérer les groupes de mots qui apparaissent ensemble plus fréquemment que ne le voudrait le hasard. Il existe beaucoup de formules différentes pour ces mesures d'association : S. Evert en détaille plus d'une vingtaine sur son site `www.collocations.de` et nous n'en présenterons donc que quelques-unes. Leur calcul se base généralement sur des tables de contingence semblables à la Table 1.1. Cette table de contingence contient les effectifs observés O pour les couples de mots apparaissant dans un contexte donné (co-occurrence directe, phrase, etc.). Les effectifs sont mesurés pour les couples de mots qu'il est possible de former à partir de deux mots x et y et l'ensemble des autres mots du corpus. L'effectif observé pour le couple de mots xy est noté O_{11}, celui du couple $\neg xy$ est O_{21}, etc. La taille de contexte utilisée pour leur calcul est variable, même si pour l'acquisition de termes on ne prend généralement en compte que la co-occurrence directe (mots adjacents).

Les paragraphes suivants détaillent quelques-unes de ces méthodes. La Table 1.2 donne quelques exemples de résultats ainsi obtenus.

	$Y = y$	$Y \neq y$
$X = x$	O_{11} $f(x, y)$	O_{12} $f(x, \neg y)$ $f_1(x) - f(x, y)$
$X \neq x$	O_{21} $f(\neg x, y)$ $f_2(y) - f(x, y)$	O_{22} $f(\neg x, \neg y)$ $N - f_1(x) - f_2(y) + f(x, y)$

TAB. 1.1: Table de contingence pour deux éléments x et y. N correspond au nombre de tokens, $f_1(x)$ au nombre d'occurrences de x en première position dans le couple et $f_2(y)$ au nombre d'occurrences de y en deuxième position dans le couple.

x	y	$f_1(x)$	$f_2(y)$	f(x,y)	Mesures d'association
du	volcan	32 005	7 866	3 736	
le	volcan	43 549	7 866	1 871	
volcan	.	7 602	80 810	1 334	
des	volcans	34 716	3 213	1 206	
les	volcans	37 760	3 213	777	
					Information mutuelle
volcans	copahué	3 160	1	1	2,83
volcans	médio-islandais	3 160	1	1	2,83
volcans	cotopoxi	3 160	1	1	2,83
volcans	fusionnés	3 160	29	29	2,83
volcans	sumbing	3 160	1	1	2,83
					χ^2
du	volcan	32 005	7 866	3 736	114 368,36
des	volcans	34 716	3 213	1 206	26 232,37
volcans	actifs	3 160	837	178	25 537,48
volcans	fusionnés	3 160	29	29	19 771,00
le	volcan	43 549	7 866	1 871	18 913,85
					Rapport de vraisemblance
du	volcan	32 005	7 866	3 736	21 235,22
le	volcan	43 549	7 866	1 871	6 288,12
des	volcans	34 716	3 213	1 206	5 812,61
les	volcans	37 760	3 213	777	2 832,03
volcan	.	7 602	80 810	1 334	2 194,33
					Indice de Jaccard
du	volcan	32 005	7 866	3 736	0,1034
volcans	actifs	3 160	837	178	0,0466
le	volcan	43 549	7 866	1 871	0,0378
des	volcans	34 716	3 213	1 206	0,0328
un	volcan	21 039	7 866	743	0,0264

TAB. 1.2: Associations les plus fortes avec "volcan" selon plusieurs mesures d'association. Les résultats ont été obtenus à partir d'un corpus français sur la volcanologie (N = 2 157 488) avec le logiciel UCS développé par Stefan Evert et disponible sur `www.collocations.de`. Le calcul de l'information mutuelle utilise le logarithme base 10.

Fréquence de co-occurrence

La mesure la plus simple pour déterminer la force d'association entre deux ou plusieurs mots est de compter le nombre d'occurrences de la suite de mots considérée dans le corpus : les suites de mots qui apparaissent fréquemment dans le corpus pourront être considérées comme des termes.

Cette approche est celle adoptée par la technique des **segments répétés** qui consiste à repérer les séquences de mots répétées dans le corpus [Lebart et Salem, 1994]. L'identification des segments répétés repose sur les caractères délimiteurs comme les signes de ponctuation : les segments ne peuvent chevaucher un signe de ponctuation (délimiteur de séquence). Un segment répété est une suite d'occurrences non séparées par un délimiteur de séquence et de fréquence supérieure ou égale à 2.

Les résultats ainsi obtenus peuvent être améliorés par l'utilisation de filtres [Rousselot, 2004] identifiant les mots qui indiquent des frontières de termes, en complément des signes de ponctuation délimiteurs de séquences (filtre « coupant » : verbes courants, adverbes, pronoms relatifs, conjonctions) et ceux qui ne peuvent se trouver aux bornes d'un terme (articles, prépositions). Les redondances sont ensuite supprimées en utilisant le mécanisme de l'**intersection lexicale** (également appelé **contrainte d'autonomie** par [Drouin, 2003]) : par exemple, si l'on trouve *artère coronaire droite* de fréquence 3 et *coronaire droite* de fréquence 3, *coronaire droite* est considéré comme un sous-segment de *artère coronaire droite* et est donc supprimé.

Le test du χ^2

La formule pour le test du χ^2 (ou *test de Pearson*) est la suivante :

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})}$$

Pour le test du χ^2 on fait l'hypothèse que les variables sont distribuées de manière normale. Cependant, cette hypothèse n'est pas très réaliste dans le cas de la fréquence d'occurrence des mots dans un texte car les événements rares sont très fréquents [Dunning, 1993].

Le coefficient de Jaccard

La formule du coefficient de Jaccard est la suivante :

$$\text{Jaccard} = \frac{f(x,y)}{f_1(x) + f_2(y) - f(x,y)} = \frac{O_{11}}{O_{11} + O_{12} + O_{21}}$$

Cette mesure peut également s'appliquer à la comparaison de vecteurs de co-occurrences pour donner une mesure de la similarité entre mots [Manning et Schütze, 1999, page 299], [Oakes, 1998, page 140].

L'information mutuelle

L'information mutuelle compare la probabilité d'observer deux mots x et y ensemble aux probabilités de les observer indépendamment. Elle se calcule de la manière suivante, selon la formule donnée par [Church et Hanks, 1990] :

$$I(x,y) = log_2 \frac{P(x,y)}{P(x)P(y)}$$

où $P(x,y)$, $P(x)$ et $P(y)$ sont estimés par le maximum de vraisemblance tel que $f(x)$ est l'effectif du mot x dans un corpus de taille N (nombre d'occurrences de mots) :

$$P(x,y) = \frac{f(x,y)}{N} \; ; \; P(x) = \frac{f(x)}{N} \; ; \; P(y) = \frac{f(y)}{N}$$

Le calcul de la fréquence de co-occurrence de x et y, $f(x,y)$ peut s'effectuer dans des fenêtres de différentes tailles. Dans le cas des bigrammes (suite de deux mots) de la table de contingence 1.1 la formule devient donc :

$$I(x,y) = log \frac{Nf(x,y)}{f_1(x)f_2(y)}$$

Selon la formule, si x et y sont liés, $P(x,y)$ sera supérieur au calcul de la fréquence de co-occurrence attendue $P(x)P(y)$, sous l'hypothèse nulle de l'indépendance des occurrences de x et de y, et donc $I(x,y) \gg 0$. Cependant, l'information mutuelle ne fonctionne pas très bien pour les données d'effectif faible dans le corpus considéré [Manning et Schütze, 1999, pages 180-181], [Dunning, 1993], comme le montrent les résultats de la Table 1.2. Une autre mesure, le log du rapport de vraisemblance (*log likelihood ratio* en anglais) est plus robuste eu égard au traitement d'événements dont le nombre d'occurrences est faible [Manning et Schütze, 1999, page 174]. Cependant, la valeur de cette mesure est élevée pour des termes fréquents qui apparaissent rarement ensemble [Pantel et Lin, 2001]. Les deux mesures peuvent donc être combinées pour obtenir de meilleurs résultats [Pantel et Lin, 2001].

Limites des mesures d'association

Outre les inconvénients propres à chacune des mesures décrites, les mesures d'associations ne permettent pas de repérer avec précision les unités terminologiques. Ces mesures sont d'une part limitées par le nombre de mots, généralement deux, auxquelles elles peuvent s'appliquer. De plus, les frontières naturelles entre unités peuvent ne pas être respectées [Streiter *et al.*, 2003] car elles n'utilisent aucune information sur la structure des unités à extraire.

1.3.3 Mesures de comparaison

La plupart des méthodes exposées précédemment s'appliquent prioritairement aux termes polylexicaux. L'extraction de termes simples peut prendre pour point de départ l'identification des **spécificités**, c'est-à-dire les mots spécifiques au corpus de spécialité considéré et qui le différencient d'un autre corpus, généralement de langue générale. On les appelle également **formes caractéristiques** [Lebart et Salem, 1994] ou **mots clés**. Il s'agit donc de comparer le vocabulaire de corpus différents, voire de parties de corpus. Des tests permettant de vérifier si la différence entre fréquence attendue et fréquence observée est statistiquement significative sont

alors appliqués pour fournir une mesure statistique indiquant le degré de spécificité. On distinguera ainsi les **spécificités positives**, indiquant un sur-emploi, et les **spécificités négatives**, indiquant le sous-emploi d'une forme.

Comme pour les mesures d'association décrites précédemment, les calculs reposent sur des tables de contingence similaires à la table 1.3.

	Corpus 1	Corpus 2	Total
Fréquence du mot	a	b	a+b
Fréquence des autres mots	c-a	d-b	c+d-a-b
Total	c	d	c+d

TAB. 1.3: Table de contingence pour la comparaison des fréquences de mots entre corpus (tiré de [Rayson et Garside, 2000]).

Comparaison partie-tout

Le calcul des spécificités par comparaison *partie-tout* utilise le corpus entier comme corpus de référence pour identifier les mots spécifiques à une section du corpus. [Lafon, 1980] (cité dans [Labbé et Labbé, 2001]) a ainsi proposé d'utiliser la distribution hypergéométrique, plutôt que la distribution normale, pour mesurer les variations de fréquence d'une forme dans différentes parties d'un corpus. Cette méthode de calcul est utilisée par [Drouin, 2004] et [Lebart et Salem, 1994, chapitre 6].

Comparaison entre documents d'un même corpus

Les mesures utilisées en recherche d'information, notamment dans les modèles à base d'espaces vectoriels, cherchent à quantifier l'information apportée par un mot par rapport à un document. Dans la mesure bien connue du *TD.IDF*, *TF* (Term Frequency) représente la fréquence du terme dans le document, tandis que *IDF* (Inverse Document Frequency) donne une mesure de l'importance du terme dans l'ensemble des documents. Le TF.IDF d'un terme pour un document sera important si le terme est fréquent dans le document considéré et présent dans peu d'autres documents. Cette mesure permet donc d'éliminer les termes communs à plusieurs documents [Salton et McGill, 1983] (cité dans [Manning et Schütze, 1999]).

Comparaison corpus de référence - corpus de spécialité

Les termes identifiés par comparaison de corpus différents sont communément appelés mots clés (*key words*) plutôt que spécificités. Pour appliquer ce type de méthode, il est nécessaire de disposer de deux corpus : un corpus de référence (dit « général ») couvrant divers domaines et un corpus spécifique à un domaine particulier dont on souhaite extraire les termes. La mesure statistique la plus communément utilisée est le log du rapport de vraisemblance (*log likelihood ratio*) [Dunning, 1993, Rayson et Garside, 2000], qui, comme nous l'avons vu précédemment, peut également donner une mesure de la force d'association entre mots. Cette dernière est préférée par [Dunning, 1993] au test du χ^2 dont les résultats peuvent être faussés pour des fréquences attendues trop faibles ou au contraire quand les fréquences sont trop importantes.

La méthode consiste à comparer les effectifs observés du mot dans chacun des corpus : $O_1 = a$ (Corpus 1) et $O_2 = b$ (Corpus 2) aux effectifs attendus selon l'hypothèse d'indépendance :

$E_1 = c \cdot \frac{a+b}{c+d}$ (Corpus 1) et $E_2 = d \cdot \frac{a+b}{c+d}$ (Corpus 2) où c est le nombre d'occurrences de mots dans le Corpus 1 et d le nombre d'occurrences de mots dans le Corpus 2. La formule du log du rapport de vraisemblance donnée par [Rayson et Garside, 2000] est la suivante :

$$-2 \ln \lambda = 2 \left(a \ln \left(\frac{a}{E_1} \right) + b \ln \left(\frac{b}{E_2} \right) \right)$$

Nous venons de présenter deux types de méthodes d'acquisition de termes : les méthodes basées sur des patrons décrivant la structure des termes à identifier et celles basées sur des mesures statistiques. Chacune de ces méthodes présente des avantages et des inconvénients. Les méthodes à base de patrons permettent l'identification de termes rares lorsqu'elles n'utilisent pas de seuil de fréquence, tandis que les méthodes à base de statistiques ordonnent les résultats de l'extraction en fonction de la mesure utilisée, facilitant ainsi l'analyse des données extraites. Certains systèmes combinent donc les deux approches, comme par exemple [Daille, 1996]. Les résultats obtenus peuvent être évalués par diverses méthodes que nous allons décrire dans la prochaine section.

1.3.4 Évaluation des résultats de l'extraction terminologique

Les résultats de l'extraction des termes ou de mots clés sont difficiles à évaluer, et ce quelle que soit la méthode utilisée (patrons, mesures statistiques). Les mesures d'évaluation utilisées viennent du domaine de la recherche d'information. Elles utilisent soit une liste de référence des termes du domaine, généralement produite indépendamment du corpus utilisé par l'extraction, soit une annotation manuelle d'un ou plusieurs documents.

Le **rappel** mesure la capacité de la méthode à identifier tous les termes du document de référence. Il se calcule en divisant le nombre total de termes correctement identifiés par le nombre de termes dans le document de référence.

$$\text{rappel} = \frac{\mid T \cap T_{ref} \mid}{\mid T_{ref} \mid}$$

avec :
- $\mid T \mid$ = nombre total de termes identifiés
- $\mid T_{ref} \mid$ = nombre de termes dans le document de référence
- $\mid T \cap T_{ref} \mid$ = nombre de termes correctement identifiés

La **précision** mesure la capacité de la méthode à identifier des termes corrects. Elle se calcule en divisant le nombre total de termes correctement identifiés par le nombre total de termes identifiés.

$$\text{précision} = \frac{\mid T \cap T_{ref} \mid}{\mid T \mid}$$

On cherche ainsi à obtenir la plus grande précision et le plus grand rappel possible. La **F-mesure** combine la précision et le rappel et correspond à leur moyenne harmonique.

$$\text{F}-\text{mesure} = \frac{2 \cdot \text{précision} \cdot \text{rappel}}{\text{précision} + \text{rappel}}$$

Enfin, la mesure de rappel en fonction du rang des candidats, ou mesure de « **ranked recall** » [Streiter *et al.*, 2003], donne une mesure de la qualité du classement des termes en

fonction d'un poids. Les termes ou mots-clés extraits sont souvent pondérés par une mesure, comme la fréquence d'occurrence dans le corpus, ceci afin de les trier et donc de faciliter leur analyse. Les termes dont le poids se situe en-deçà d'un certain seuil peuvent ainsi être éliminés. Il est donc utile d'évaluer les rangs attribués aux mots-clés par de telles mesures. En effet, une méthode d'extraction qui identifie 5 termes corrects, apparaissant aux rangs 3 à 7 est moins bonne qu'une autre méthode qui identifie les même mots-clés mais qui les classe du rang 1 à 5.

Si r_i est le rang du i-ème mot-clé extrait et $n = \mid T \cap T_{ref} \mid$ alors la mesure de « ranked-recall » se calcule comme suit :

$$\text{ranked recall} = \frac{\sum_{i=1}^{n} i}{\sum_{i=1}^{n} r_i}$$

Malgré l'éventail de mesures qu'il est possible d'utiliser, l'évaluation des résultats de l'acquisition de termes candidats est une tâche difficile. On distingue deux types d'évaluation : utilisation d'une ressource existante ou validation manuelle des termes candidats par des experts du domaine.

La méthode d'évaluation la plus courante consiste à comparer la liste de termes candidats obtenue à une liste de référence des termes du domaine construite manuellement et validée [Daille, 1996]. Cependant, les terminologies de référence existantes sont généralement établies indépendamment d'un corpus textuel précis, tandis que les résultats de l'extraction automatique sont fortement dépendants du contenu du corpus utilisé. Ainsi, tous les termes pertinents par rapport au corpus ne sont pas forcément présents dans la terminologie de référence et peuvent donc pour cette raison être injustement considérés comme du bruit.

S'il n'existe pas de terminologie de référence du domaine, reste la méthode de la validation manuelle. Celle-ci n'est toutefois possible que pour un corpus de taille restreinte ou pour un nombre réduit de termes, généralement les meilleurs selon le poids assigné par la méthode. Ainsi, [Enguehard, 1992] présente plusieurs évaluations manuelles du système ANA, pour des listes de termes d'une taille limitée, comprise entre 300 et 700 éléments. L'évaluation proposée par [Streiter *et al.*, 2003] se base quant à elle sur un corpus très petit de 994 mots (occurrences). De plus, l'évaluation manuelle doit être effectuée par des spécialistes du domaine, qui pourront ne pas s'accorder sur la liste des termes corrects.

Il est également possible de combiner les deux méthodes d'évaluation. Dans [Drouin, 2004], les termes extraits (spécificités) sont d'abord comparés avec une banque de terminologie. Ceux qui sont présents dans la terminologie sont considérés comme pertinents tandis que les autres sont évalués par des terminologues afin de déterminer leur pertinence.

Dans tous les cas, c'est généralement la pertinence (pertinence par rapport au domaine et pertinence par rapport au corpus) qui est mesurée, plus que la mesure classique de rappel, difficile à estimer car il faudrait alors disposer de l'inventaire de l'ensemble des termes à extraire du corpus. [Pantel et Lin, 2001] tentent de contourner les défauts des méthodes d'évaluation précédentes en présentant une méthodologie d'évaluation automatique des résultats de leur système d'extraction de termes complexes basée sur la perplexité. La perplexité mesure la capacité d'un modèle, par exemple un modèle de langue, à effectuer des prédictions sur des données, par exemple un corpus. La méthode d'évaluation proposée consiste à extraire une liste des termes d'un corpus d'apprentissage puis à mesurer la capacité de cette liste à prédire les données d'un corpus de test par la perplexité. C'est donc avant tout la capacité du système à identifier des

unités complexes qui est évaluée, plus que ses performances pour l'acquisition des termes d'un domaine.

Une fois les termes et mots clés identifiés, reste à les structurer. Nous allons donc maintenant décrire les systèmes d'acquisition de relations sémantiques utilisés pour organiser les termes et mots clés.

1.4 Acquisition automatique de relations sémantiques

Il existe deux approches principales pour l'acquisition de relations sémantiques entre termes. Les méthodes dites « externes » se basent sur la comparaison des contextes d'occurrence, tandis que les méthodes dites « internes » reposent sur la structure morphologique des mots ou la structure lexicale des expressions. Nous allons dans un premier temps décrire diverses approches contextuelles, comme les modèles à base de vecteurs et de graphes (Section 1.4.1), les méthodes de classification (Section 1.4.2) et les méthodes par description de patrons lexico-syntaxiques (Section 1.4.3). Puis, nous allons présenter les méthodes reposant sur la structure interne des mots et expressions (Section 1.4.4).

1.4.1 Vecteurs et graphes de co-occurrences

De nombreux modèles d'acquisition de relations sémantiques reposent sur l'idée que le sens d'un mot est lié à ses contextes d'utilisation. Ainsi, les mots sont considérés comme sémantiquement proches s'ils apparaissent dans des contextes similaires. Ces méthodes rejoignent ainsi les théories qui mettent l'accent sur l'importance de l'usage pour la sémantique lexicale telles que celles du linguiste Firth (« You shall know a word by the company it keeps ») et du philosophe Wittgenstein (« Meaning is use »).

Différents niveaux de relations de co-occurrence peuvent être considérés :

- **Co-occurrences de 1er ordre** ou co-occurrences directes : deux mots sont considérés comme proches s'ils apparaissent dans le même contexte, c'est-à-dire s'ils sont directement co-occurrents (relation syntagmatique).
- **Co-occurrences de 2nd ordre** ou co-occurrences indirectes : deux mots sont similaires s'ils apparaissent dans des contextes similaires (relation paradigmatique). Ainsi deux mots M_1 et M_2 pourront être considérés comme sémantiquement proches s'ils partagent des co-occurrents M_i et ce même s'ils n'apparaissent jamais dans le même contexte [Denhière et Lemaire, 2003, Martinez, 2000, Rapp, 2003].

La co-occurrence directe est mesurable de diverses manières, présentées dans la section 1.3.2 : fréquence de co-occurrence, information mutuelle, test du χ^2, rapport de vraisemblance, indice de Jaccard. De plus, suivant la méthode, la taille du contexte (fenêtre de mots, phrase, paragraphe) est variable. Les mots constituant le contexte sont parfois sélectionnés en fonction de leur catégorie morpho-syntaxique.

La mesure de la co-occurrence indirecte s'effectue généralement à l'aide de vecteurs représentant chaque mot. Chaque composante d'un tel vecteur contient la mesure de co-occurrence directe du mot considéré avec un certain mot du lexique. Il faut noter que certaines méthodes représentent les co-occurrences sous forme de graphe [Lemaire et Denhière, 2004, Widdows et Dorow, 2002] dont les nœuds sont les mots et les arcs représentent la relation de co-occurrence directe ou indirecte entre mots. Les graphes de co-occurrence sont surtout utilisés pour découvrir les sens et usages différents des mots par détection des zones de forte densité dans le graphe

[Véronis, 2003, Ferret, 2004]. Les deux modes de représentation, vecteurs et graphes, ne sont toutefois pas totalement dissimilaires, un graphe pouvant être représenté sous forme de matrice.

Les mesures basées sur les vecteurs consistent à calculer la similarité de deux mots en fonction de la similarité ou de la distance des vecteurs les représentant. Les mesures les plus fréquemment utilisées sont la distance euclidienne, la mesure de Kullback-Leibler et le cosinus. Généralement, la taille des vecteurs est réduite de sorte à ne conserver qu'un certain nombre de composantes. En effet, le nombre de mots d'un corpus est de l'ordre des dizaines de milliers et les vecteurs résultants sont donc très grands et qui plus est « creux »(beaucoup de composantes ont pour valeur 0). Les composantes conservées sont sélectionnées en fonction de divers critères : fréquence (les mots les plus fréquents), variance (composantes de plus grande variance), productivité (composantes non nulles pour le plus grand nombre de vecteurs). De plus, le nombre de dimensions conservées est également variable[1].

Les modèles vectoriels les plus connus sont LSA (Latent Semantic Analysis) et HAL (Hypertext Analog to Language). Le principe de LSA [Landauer *et al.*, 1998] consiste à transformer tout corpus en une matrice dans laquelle chaque ligne correspond à un mot et chaque colonne à un contexte. Chaque cellule de la matrice contient le nombre d'occurrences d'un mot dans le contexte correspondant à la colonne. Le contenu de la matrice est pondéré puis soumis à une décomposition en valeurs singulières pour réduire le nombre de dimensions de la matrice à environ 300. Les mots peuvent alors être comparés en calculant la similarité des lignes de la matrice qui les représentent. HAL [Lund et Burgess, 1996, Li *et al.*, 2000] procède différemment pour construire la matrice. Les lignes contiennent les valeurs de co-occurrence pour les mots précédant le mot correspondant à la ligne dans le corpus, tandis que les colonnes représentent les mots suivants. Chaque mot peut ainsi être représenté par un vecteur dont la taille est le double de celle du lexique (concaténation du vecteur ligne et du vecteur colonne). En pratique, la taille du vecteur est réduite aux composantes présentant la plus grande variance.

Il faut également noter que les modèles à base de vecteurs peuvent être utilisés pour l'acquisition des sens des mots polysémiques. Par exemple, le système ACOM [Ji, 2004] sélectionne des mots liés par les contextes, appelés contexonymes puis forme des cliques à partir de ces mots contextuellement liés, une clique correspondant aux mots qui sont tous les contexonymes les uns des autres. Les cliques sont alors projetées dans un espace sémantique multi-dimensionnel et regroupées par un algorithme de classification hiérarchique. Ce système est basé sur la méthode initialement proposée par [Ploux et Victorri, 1998], permettant de caractériser le sens des mots polysémiques à partir de cliques construites à l'aide de dictionnaires de synonymes. Dans ce cas, les cliques sont des unités de sens, similaires aux synsets de WordNet : dans une clique, tous les mots sont synonymes (ou quasi-synonymes) aux autres mots de la clique.

Ces approches sont néanmoins critiquables sur certains points :

- Il est nécessaire de disposer de très gros corpus de textes pour obtenir de bons résultats car la plupart des mots sont rares (problème connu sous le nom de « data sparseness »).
- Ces méthodes donnent une mesure de proximité entre termes. Mais la nature exacte de la relation (i.e. une des relations décrites dans la Section 1.2.3, p. 14) n'est pas connue. Par exemple, une étude des relations sémantiques effectivement extraites par LSA montre que seule une faible proportion de ces relations correspond à des relations d'inclusion, d'identité ou d'opposition [Wandmacher, 2005].
- Les unités contextuelles généralement utilisées sont les mots. Or les mots ne sont pas des entités sémantiquement atomiques [French et Labiouse, 2002] mais décomposables en unités porteuses de sens de niveau inférieur, les morphèmes.

[1] Pour une discussion détaillée des méthodes de réduction de dimensions, voir [Levy et Bullinaria, 2001]

Les mesures de similarité obtenues peuvent être utilisées pour catégoriser automatiquement les mots, grâce à des algorithmes de classification que nous décrivons dans la section suivante.

1.4.2 Classification

L'objectif de la classification est de repérer les mots similaires pour ensuite les grouper en catégories. Les algorithmes de classification de mots sont variés. Dans certains cas, la classification est hiérarchique : les catégories obtenues forment alors une taxinomie.

Les cartes auto-organisatrices de Kohonen (ou SOM pour *Self-Organizing Maps*) sont utilisées pour classer divers types de données, et notamment les documents [Kohonen *et al.*, 2000]. Elles permettent également de catégoriser les mots. Les cartes auto-organisatrices sont un type simple de réseau de neurone dont l'objectif est d'organiser les données dans un espace à deux dimensions, la carte, et ceci de manière totalement non supervisée. [Honkela *et al.*, 1995] présentent une utilisation des SOM pour la classification des 150 mots les plus fréquents des contes de Grimm. La carte obtenue après apprentissage reflète à la fois les catégories sémantiques et les catégories syntaxiques des mots étudiés.

Les méthodes de classification utilisent généralement des textes analysés syntaxiquement, afin de repérer diverses relations permettant de former des classes de mots. Ainsi, [Caraballo, 1999] se base sur les groupes nominaux apposés ou joints par une conjonction de coordination pour construire automatiquement une hiérarchie de noms en utilisant une méthode de classification hiérarchique ascendante.

L'utilisation des contextes syntaxiques peut permettre une classification encore plus raffinée, par le regroupement dans des classes distributionnelles des termes qui apparaissent dans le même contexte syntaxique. La similarité de deux mots est alors fonction du nombre de contextes syntaxiques qu'ils partagent. Ce type d'analyse distributionnelle se fonde sur les travaux de Harris [Habert et Zweigenbaum, 2002].

Par exemple, [Hindle, 1990] décrit une méthode de classification de mots anglais en fonction des structures prédicat-argument dans lesquels ils apparaissent. En effet, un nom ne peut généralement être le sujet et/ou l'objet que d'un nombre restreint de verbes. Si l'on prend l'exemple du mot *wine*, il peut apparaître avec les verbes *drink* et *produce* mais pas *prune*. Il est ainsi possible de caractériser chaque nom par un ensemble de verbes. Puis, les noms peuvent être regroupés en fonction des similarités des environnements lexico-syntaxiques dans lesquels ils apparaissent. La première étape du traitement consiste donc à effectuer une analyse syntaxique du corpus, afin de mettre à jour les relations de type sujet-verbe-objet. La pertinence de ces relations est évaluée à l'aide de l'information mutuelle. Les noms sont ensuite regroupés à l'aide d'une mesure de similarité basée sur l'information mutuelle calculée et prenant en compte les verbes partagés.

D'une manière assez semblable, [Lin, 1998] présente une méthode d'identification de mots similaires pour la construction automatique de thésaurus. Tout d'abord, un analyseur syntaxique est appliqué au corpus pour obtenir des couples de mots liés par une relation de dépendance comme sujet-verbe, verbe-objet, nom-adjectif, etc. Puis une mesure de similarité entre mots est calculée en fonction des relations de dépendances partagées par les mots.

Divers autres systèmes se basent sur la même procédure : analyse du corpus pour extraire des relations syntaxiques et agrégation des mots partageant les mêmes relations syntaxiques [Faure et Nédellec, 1999, de Chalendar et Grau, 2000, Bourigault, 2002].

Cependant, même si ces méthodes se basent sur un corpus analysé et sur des relations de co-occurrences bien spécifiques, elles ne permettent pas toujours d'étiqueter les relations sémantiques obtenues. De plus, le bénéfice de l'utilisation d'une analyse syntaxique préalable n'est pas vérifié dans tous les cas. En effet, [Grefenstette, 1996] a montré que les méthodes

utilisant l'analyse syntaxique fournissent de meilleurs résultats pour les mots les plus fréquents mais sont surpassées pour les méthodes utilisant une fenêtre de mots sans autre pré-traitement pour les mots les moins fréquents. De plus, le travail sur de très gros corpus de textes nécessite de prendre en compte également le temps d'éxecution et la taille des représentations fournies par les analyseurs [Curran et Moens, 2001][1] .

1.4.3 Patrons lexico-syntaxiques

Les méthodes décrites précédemment ne permettent pas l'acquisition de relations sémantiques étiquetées. Or certaines relations comme l'hyper/hyponymie se caractérisent par des constructions spécifiques qu'il est possible de repérer dans les textes après les avoir spécifiées sous forme de patrons lexico-syntaxiques. Le Tableau 1.4 liste les patrons proposés par [Hearst, 1992] pour la relation d'hyponymie en anglais.

Patrons lexico-syntaxiques	**Exemple**
NP_0 such as NP_1 , NP_2 ..., (and | or) NP_n	... metabolic disorders **such as** phenylketonuria**,** hypothyroidism **and** cystic fibrosis ... ⇒ hyponyme (phenylketonuria, metabolic disorder) ⇒ hyponyme (hypothyroidism, metabolic disorder) ⇒ hyponyme (cystic fibrosis, metabolic disorder)
such NP as NP,* (or | and) NP	...**such** disorders **as** cystic fibrosis **and** muscular dystrophy... ⇒ hyponyme (cystic fibrosis, disorder) ⇒ hyponyme (muscular dystrophy, disorder)
NP , NP*, or other NP	...penicillin **or other** drugs... ⇒ hyponyme (penicillin, drug)
NP , NP*, and other NP	...antibiotics **and other** medicines... ⇒ hyponyme (antibiotics , medicine)
NP , including NP,* or | and NP	... surgical exploration, **including** biopsy **and** cytology ... ⇒ hyponyme (biopsy, surgical exploration) ⇒ hyponyme (cytology, surgical exploration)
NP , especially NP,* or | and NP	Consult with other medical professionals**, especially** primary practitioners ... ⇒ hyponyme(primary practitioners, medical professionals)

Tab. 1.4: Patrons lexico-syntaxiques définis par [Hearst, 1992].

Afin d'améliorer la pertinence des couples de termes hyponymes extraits par cette méthode, [Cederberg et Widdows, 2003] proposent d'utiliser l'analyse sémantique latente (LSA) pour effectuer un filtrage. Plus la similarité des deux termes est importante suivant cette analyse, plus la relation d'hyponymie qui les relie est plausible.

[Hearst, 1992] décrit également une méthode permettant de découvrir de nouveaux patrons (cette méthode a également adoptée par [Morin, 1998]) :

1. Choisir une relation lexicale pour laquelle on souhaite découvrir les patrons.

[1] Nous remercions Pierre Zweigenbaum pour nous avoir signalé cette référence.

2. Réunir un ensemble de termes liés par cette relation.
3. Rechercher dans le corpus les contextes dans lesquels les couples de termes apparaissent ensemble.
4. Trouver les points communs de ces contextes. [Morin, 1998] définit une mesure de similarité permettant de regrouper ces environnements dans des classes.
5. Lorsqu'un nouveau patron a été identifié, l'utiliser pour rassembler de nouvelles instances de la relation et revenir à l'étape 2.

Une variante de cette méthode, décrite dans [Finkelstein-Landau et Morin, 1999], consiste à la combiner avec une approche non supervisée. Dans ce cas, les relations plausibles statistiquement (selon une mesure d'association des termes telle que l'information mutuelle) sont sélectionnées de manière non supervisée, ce qui permet d'automatiser les phases 1 et 2 de l'extraction.

[Rebeyrolle, 2000] complète les patrons morpho-syntaxiques par l'utilisation de marqueurs typographiques et dispositionnels pour repérer les définitions dans un texte. Ainsi, le terme défini peut être marqué typographiquement par des caractères gras ou italiques, des lettres majuscules ou des guillemets. De plus, les structures définitoires se retrouvent régulièrement en début de paragraphe. Au niveau discursif, le terme à définir est généralement mentionné une première fois avant d'être repris dans un énoncé définitoire.

Les méthodes d'acquisition de relations sémantiques à partir de textes que nous venons de présenter dans les trois sections précédentes ont toutes un point commun, celui d'utiliser le contexte d'occurrence des mots. Dans le cas le plus simple, aucun pré-traitement n'est appliqué au corpus et le contexte est alors constitué de mots. Dans les cas les plus complexes, certains patrons spécifiques, basés sur l'analyse morpho-syntaxique du corpus, sont recherchés dans le corpus. Or, les termes sont souvent des unités polylexicales. Certaines méthodes, que nous allons décrire dans la section suivante, se basent donc sur la structure lexicale des termes pour leur structuration.

1.4.4 Utilisation de la structure interne des termes

Les informations internes aux termes, reposant sur leur structure, peuvent être utilisées, notamment pour repérer les relations d'antonymie et de spécialisation. Les informations internes utilisables se trouvent à deux niveaux :

- Niveau du morphème : la comparaison entre termes simples, monolexicaux mais polymorphémiques, s'effectue sur la base de leur structure morphologique.
- Niveau du mot : la comparaison entre termes polylexicaux s'effectue sur la base des mots qu'ils contiennent.

Nous allons d'abord présenter les méthodes basées sur la structure lexicale des termes polylexicaux, qui sont celles que l'on rencontre le plus fréquemment dans la littérature. Puis, nous présentons les utilisations possibles de la structure morphologique pour l'acquisition de relations sémantiques.

Utilisation de la structure lexicale des termes polylexicaux

Certaines relations sémantiques, et notamment l'hyperonymie et la co-hyponymie sont marquées par des relations structurelles entre les termes.

De nombreux travaux se basent sur l'**inclusion lexicale** pour retrouver des relations d'hypo-/hyperonymie [Bodenreider *et al.*, 2001, Grabar et Zweigenbaum, 2002a]. En effet, l'hyperonymie

se manifeste par une structure lexicale spécifique, notamment pour les noms : l'hyperonyme est un nom, l'hyponyme est un composé, comme par exemple *table gigogne* ou *table de cuisine* [Kleiber et Tamba, 1990]. Le nom hyperonyme est inclus dans le composé qui est son hyponyme. La notion d'inclusion lexicale est définie de la manière suivante par [Grabar et Zweigenbaum, 2002a] : un terme T_1 est lexicalement inclus dans un autre terme T_2 ssi tous les mots informatifs formant T_1 se trouvent également dans T_2. Par exemple, le terme *acide gras* est inclus dans le terme plus long *acide gras libre* : *acide gras* est l'hyperonyme de *acide gras libre*, c'est-à-dire qu'un *acide gras libre* est un type d'*acide gras*.

On peut distinguer trois types de relations d'inclusion lexicale en anglais [Ibekwe-SanJuan, 1998] :

- **Expansion gauche** : $T_2 = M + T_1$. M peut être selon le cas un adjectif [Bodenreider *et al.*, 2001, Ibekwe-SanJuan, 2005], comme par exemple *ventricular aneurysm – aneurysm* ou un nom [Ibekwe-SanJuan, 2005], comme dans *compression fracture – fracture*.
- **Insertion** : dans ce cas, un nouveau mot M est inséré au milieu de T_1 pour former T_2 comme dans *adult brain glioblastoma – adult glioblastoma*.
- **Expansion droite** : $T_2 = T_1 + M$, comme par exemple *cholesterol – cholesterol granuloma*.

Les relations hiérarchiques sont marquées de manière préférentielle par l'expansion gauche et l'insertion, du moins en anglais [Ibekwe-SanJuan, 2005]. L'expansion droite correspond à des relations sémantiques plus faibles, similaires aux liens *Voir aussi* présents dans les thésaurus.

Les résultats obtenus par les méthodes basées sur la structure lexicale montrent qu'elles complètent utilement les méthodes contextuelles et offrent l'avantage d'identifier des liens sémantiques spécifiques comme la spécialisation. Par exemple, [Bodenreider *et al.*, 2001] montrent que moins de la moitié des liens hiérarchiques suggérés par les relations d'expansion gauche adjectivale sont effectivement présents dans le thésaurus qu'ils utilisent (UMLS : Unified Medical Language System). Les liens identifiés pourraient donc permettre de compléter les relations hiérarchiques présentes dans la ressource, après validation.

Une autre relation structurelle, qui présente elle un intérêt pour l'identification de co-hyponymes, est la **substitution** [Ibekwe-SanJuan, 1998]. Elle correspond au remplacement d'un mot informatif de T_1 par un autre mot dans T_2 où T_1 et T_2 contiennent le même nombre de mots. La substitution de modifieurs indique une relation de co-hyponymie entre termes, comme c'est le cas par exemple pour *liver granuloma – cardiac granuloma*.

Les modifications morpho-syntaxiques de l'un des constituants d'un terme complexe peuvent induire des relations sémantiques supplémentaires comme l'antonymie (*organic chemical – inorganic chemical*) ou la succession dans le temps [Daille, 2003].

Utilisation de la structure morphologique des termes simples

Dans la mesure où de nombreux termes techniques ont une structure morphologique complexe, il est envisageable d'appliquer des méthodes similaires aux termes simples. L'utilisation de la structure morphologique des termes simples pour l'acquisition de relations sémantiques est bien sûr dépendante des ressources morphologiques disponibles. La majorité des systèmes combinent donc acquisition de données morphologiques, de manière supervisée ou non supervisée, et utilisation des données ainsi extraites. Nous décrirons plus précisément les méthodes d'acquisition de connaissances morphologiques dans le chapitre suivant.

Les relations sémantiques sont marquées dans la structure morphologique de diverses manières.

L'hyponymie se manifeste sous forme d'inclusion dans les mots à composition savante, comme l'atteste l'exemple suivant : « *Angiosarcome* : type de *sarcome* qui apparaît sur un vaisseau san-

guin ». Ainsi, [Buitelaar et Sacaleanu, 2002] présentent une méthode d'extension des synsets de GermaNet (version germanique de WordNet) reposant sur la décomposition des mots complexes comme *Sauerstofftherapie* (oxygénothérapie) en tête (*Therapie*) et modifieur (*Sauerstoff*). Le terme composé, formé par inclusion d'un terme plus court, est considéré comme un hyponyme de la tête : *Sauerstofftherapie* est donc un hyponyme de *Therapie*.

La relation d'antonymie est marquée morphologiquement par des préfixes ou des suffixes dérivationnels qui s'opposent sémantiquement. Ainsi, [Schwab *et al.*, 2005] décrivent une méthode d'extraction semi-supervisée de couples d'antonymes à partir de leurs préfixes. Les suffixes pourraient également être utilisés mais sont beaucoup plus rares (on peut citer l'opposition *+phile/+phobe*). Le tableau 1.5 décrit différents types d'oppositions et des exemples de leur réalisation sous forme de préfixes découverts au cours de l'extraction. [Grabar et Hamon, 2006] proposent également une méthode d'identification de l'antonymie à partir des préfixes de négation comme *dé+*, *ir+*, *anti+*, *non+* ou *in+* et des préfixes privatifs comme *a+* ou *dys+*.

Les liens sémantiques qui s'expriment par des liens morphologiques ne se limitent pas aux seules relations d'inclusion et d'opposition. On trouve d'autres liens sémantiques [Light, 1996, Namer, 2002, Daille, 2003, Claveau et L'Homme, 2005, Grabar et Hamon, 2006] comme par exemple :

- la répétition : préfixes *re+* ou *ré+* ;
- le changement d'état : suffixes *+iser* ou *+ifier* ;
- la localisation spatiale : préfixes *sur+*, *sous+*, *contre+*, *péri+* ;
- la localisation temporelle : préfixes *pré+*, *post+* ;
- les rôles sémantiques : agent (suffixe *+eur*), résultat (suffixe *+ure*) ;
- la méronymie : suffixes *+age*, *+ade*.

Ces régularités de correspondance entre structure morphologique et liens sémantiques autorisent une approche de l'analyse morphologique de type morphosémantique [Namer, 2003], qui sera décrite plus en détail dans la Section 2.3.1, p. 43.

Type d'opposition	**Préfixes**	**Exemples**
Opposition de degré	hypo+/hyper	hypocalorique/hypercalorique
	micro+/macro+	microcristaux/macrocristaux
	sous+/sur+	sous-alimentation/suralimentation
	infra+/supra+	infra-centimétrique/supra-centimétrique
Opposition de nombre	mono+/poly+	monogénique/polygénique
	uni+/omni+	unidirectionnel/omnidirectionnel
	uni+/bi+	unilingue/bilingue
	uni+/tri+	unicolore/tricolore
Opposition dans l'espace	ex+/in+	exhalation/inhalation
	exo+/endo+	exogène/endogène
	extra+/intra+	extra-cellulaire/intra-cellulaire
Opposition dans le temps	pré+/post+	pré-caldeira/post-caldeira

Tab. 1.5: Récapitulatif des relations d'opposition morphologiquement marquées (d'après [Schwab *et al.*, 2005]).

1.4.5 Évaluation des résultats de l'acquisition de relations sémantiques

Tout comme pour l'extraction terminologique, les résultats des différents systèmes d'acquisition de relations sémantiques sont difficiles à évaluer. Les méthodes d'évaluation sont variées :

- Comparaison à des ressources de référence (*gold standards*) comme WordNet [Hearst, 1992, Lin, 1998, Widdows et Dorow, 2002, Ferret, 2004], le *Roget's Thesaurus* [Lin, 1998], UMLS [Bodenreider *et al.*, 2001] ou le MeSH [Grabar et Zweigenbaum, 2002a].
- Comparaison aux résultats de tests de vocabulaire à choix multiple, comme les tests du TOEFL (Test of English as a Foreign Language) [Landauer *et al.*, 1998].
- Corrélation aux temps de réaction des sujets pour des tâches de décision lexicale en psycholinguistique [Lund et Burgess, 1996].
- Évaluation manuelle [Caraballo, 1999, Buitelaar et Sacaleanu, 2002, Cederberg et Widdows, 2003, Schwab *et al.*, 2005, Wandmacher, 2005].

Les deux méthodes d'évaluation principalement utilisées sont la comparaison à des ressources de références et la validation manuelle. La première suppose la disponibilité de telles ressources pour la langue et le domaine traités, tandis que la seconde n'est possible que pour un petit nombre de données.

1.5 Conclusion

Nous avons décrit dans ce chapitre divers types de ressources lexicales et leur contenu. L'acquisition automatique de connaissances lexicales se subdivise en deux tâches principales : l'identification des mots-clés et termes, qui représentent les concepts d'un domaine, et l'acquisition de relations sémantiques entre termes. Nous avons montré que les approches sont variées, notamment du point de vue des ressources nécessaires : dans certains cas, les connaissances sont extraites à partir de corpus de textes bruts (approches essentiellement statistiques), dans d'autres cas, les corpus sont pré-traités afin de les étiqueter, voire d'effectuer une analyse syntaxique (approches linguistiques). Les deux approches sont souvent combinées pour améliorer les résultats.

Il faut toutefois noter que la majorité des méthodes décrites mettent l'accent d'une part sur l'acquisition de termes polylexicaux et d'autre part sur un mode de structuration des termes reposant sur les connaissances externes (contexte). Or, comme l'ont très justement fait remarquer [French et Labiouse, 2002], les mots ne sont pas des entités atomiques indécomposables, mais contiennent des sous-unités porteuses de sens, les morphèmes.

Certaines méthodes utilisent donc la morphologie pour (a) découvrir les mots clés d'un domaine et (b) identifier des relations sémantiques entre termes morphologiquement complexes. Elles restent toutefois minoritaires dans l'ensemble des travaux consacrés à l'acquisition automatique de ressources lexicales. Nous avons donc choisi d'étudier plus particulièrement le rôle de la morphologie dans un processus global d'acquisition de connaissances à partir de textes, incluant l'analyse morphologique non supervisée.

Nous allons tout d'abord décrire plus précisément le domaine de la morphologie à travers le point de vue de diverses disciplines : psychologie, linguistique et informatique (Traitement Automatique des Langues).

Chapitre 2

La morphologie

Nous présentons dans ce chapitre le domaine de la morphologie, à partir des points de vue de la linguistique, de la psychologie cognitive et du traitement automatique des langues.

2.1 La morphologie en linguistique

La morphologie est le domaine de la linguistique qui traite de la structure interne des mots. Selon l'approche classique, celle de la linguistique structurale, les mots sont formés de **morphèmes** qui sont les unités linguistiques minimales (c'est-à-dire non décomposables) porteuses de sens. Les morphèmes sont des unités abstraites qui peuvent avoir plusieurs formes graphiques et phoniques appelées **morphes**. Ces divers morphes, lorsqu'ils correspondent au même morphème, sont ce que l'on appelle des **allomorphes**. Par exemple, le pluriel en anglais s'exprime généralement à l'aide du morphe *-s* ; mais il arrive qu'il soit réalisé d'une autre manière, comme dans le mot *women* où il s'exprime par un changement de voyelle. Ces deux marques du pluriel sont des allomorphes.

On distingue les morphèmes **libres**, qui peuvent former un mot sans être associés à un autre morphème et les morphèmes **liés** qui sont toujours associés à d'autres morphèmes. C'est le cas des **affixes** notamment qui se combinent avec une **base** ou un **radical**. D'après [Trost, 2003], chaque langue comprend environ 10 000 morphes. Les mots qui ne contiennent qu'un seul morphe sont dits **monomorphémiques**. Un mot **morphologiquement complexe** est un mot qui est composé d'au moins deux morphèmes. Ceci correspond à la majorité du vocabulaire français (80% des mots du *Robert méthodique* selon [Rey-Debove, 1984]).

2.1.1 Procédés morphologiques

Trois procédés morphologiques principaux, utilisés dans de nombreuses langues européennes, peuvent être distingués :

- **Flexion** : phénomènes de déclinaison et conjugaison (changement de nombre, genre, temps, personne, mode et cas). La flexion n'induit pas de changement de catégorie grammaticale. Les différents mots liés par la flexion (ou formes fléchies) sont, par lemmatisation, représentés par une forme unique, le **lemme**, qui correspond à l'infinitif des verbes, au masculin singulier des adjectifs, etc.
- **Dérivation** : formation de nouveaux mots grâce à l'adjonction d'affixes au radical. En français, on peut distinguer trois opérations dérivationnelles : dérivation par préfixation [préfixe + radical] (*précancer* = [*pré* + *cancer*]), dérivation par suffixation [radical +

suffixe] (*cancéreux* = [*cancer* + *eux*]) et formation parasynthétique [préfixe + radical + suffixe] (*intraveineuse* = [*intra* + *vein* + *euse*]). La dérivation peut induire un changement de catégorie grammaticale.

- **Composition** : combinaison de deux ou plusieurs bases pour former un nouveau mot. On peut par exemple ajouter un morphème libre à un autre morphème libre (ex : *cellule-cible*). Les différentes parties peuvent éventuellement être jointes par un élément de liaison, comme la lettre 's' dans le mot allemand *Verteidigungsministerium*. La **composition savante** correspond à la combinaison d'un ou plusieurs **formants** d'origine grecque ou latine susceptibles d'occuper une position et une fonction différentes dans une structure lexicale (ex : pathologie et ostéopathie). Les formants constituent des morphèmes lexicaux liés qui se combinent généralement les uns avec les autres et se terminent souvent par une voyelle comme *+o* [Cottez, 1984].

À ces trois procédés s'ajoute la **conversion** qui consiste à former un mot de catégorie grammaticale différente sans changement de forme.

2.1.2 Types de morphèmes

La description de ces différents procédés de formation repose sur différents types de morphèmes :

- **Radical**. Le radical porte le sens principal du mot. Les mots *base*, *radical* et *racine* désignent des notions très similaires. La **base** (anglais : *base*) correspond à ce qu'il reste du mot une fois les affixes flexionnels supprimés. Elle ne constitue donc pas nécessairement une entité atomique et peut encore être décomposée en affixes dérivationnels et *radicaux*. La **racine** (anglais : *root*) est une entité abstraite, portant le sens commun à tous les mots formés à partir de cette racine. Elle peut être réalisée sous forme de divers *radicaux*. D'un point de vue diachronique, elle sert également à désigner l'origine étymologique. Le **radical** (anglais : *stem, radical*) correspond à ce qu'il reste du mot, une fois tous les affixes supprimés (voir [Gendner, 2002] pour une compilation des différentes définitions de ces termes).
- **Préfixe**. Affixe qui se place avant le radical.
- **Suffixe**. Affixe qui se place après le radical.
- **Circonfixe**. Combinaison d'un préfixe et d'un suffixe qui correspond à un trait morphologique. Par exemple, en allemand, les circonfixes *ge–t* et *ge–n* marquent le participe passé des verbes (Ex : *meinen* – *gemeint*, *lesen* – *gelesen*).
- **Infixe** : affixe qui peut se placer au milieu du radical, suivant des règles phonologiques de placement.

2.1.3 Types de langues

Les langues varient fortement du point de vue de leur mode de formation morphologique [Trost, 2003, p. 28] :

- **Langues isolantes** comme le mandarin : il n'y a pas de morphèmes liés et le seul mode de formation des mots est la composition.
- **Langues agglutinantes** comme le finnois ou le turc. Les mots sont formés par ajout d'affixes au radical, de sorte que chaque affixe ne corresponde qu'à un seul trait morphologique.
- **Langues flexionnelles** comme les langues indo-européennes.

– **Langues polysynthétiques** comme la langue inuit. Ces langues combinent radicaux et affixes pour « synthétiser » des mots très longs, pouvant correspondre à des phrases complètes dans des langues moins synthétiques.

La morphologie de l'anglais est relativement simple et la formation des mots se fait majoritairement par **concaténation** de morphèmes, suivant des règles de combinaison constituant la **morphotactique** de la langue. Le procédé de concaténation est encore plus prégnant dans des langues comme l'allemand ou le finnois. Il existe également des phénomènes qui ne relèvent pas de la concaténation. Ainsi, dans les langues sémitiques, une racine formée par deux à quatre consonnes porte le sens du mot, tandis qu'un patron formé par des voyelles s'intercalant avec les consonnes marque les informations de voix et d'aspect. On notera également les phénomènes d'**ablaut** ou **alternance vocalique** (exemple en anglais : *s*i*ng – s*a*ng – s*u*ng*) et d'**umlaut** (exemple en allemand : *Pl*a*n – Pl*ä*ne*). La **supplétion** correspond quant à elle à une modification complète du radical, comme pour le verbe anglais *go* qui devient *went* au prétérit. Elle se manifeste généralement lorsque deux radicaux d'origine différente co-existent avec le même sens [Grabar et Hamon, 2006] : par exemple *estomac*, d'origine latine, peut être remplacé par *gastr+*, d'origine grecque.

2.1.4 Morphologie et sémantique

Du point de vue classique, les morphèmes participent au sens du mot qui les contient de manière systématique et régulière, ce qui permet notamment aux locuteurs de comprendre les néologismes à partir de leurs constituants. Des expériences de psycholinguistique ont montré qu'il est même possible de donner une définition partielle d'un non-mot comme *unwugable* dérivé de *wug* [Plaut et Gonnerman, 2000] car l'association du préfixe *un+* et du suffixe *+able* se retrouve dans d'autres mots comme *unbreakable*. Lorsque l'on peut déduire le sens des mots de leur constituants, ces mots sont dits **transparents**. C'est d'ailleurs une des techniques naturellement utilisées par tout lecteur lorsqu'il rencontre un mot inconnu dans un texte : soit il s'aide du contexte, soit de la forme même du mot inconnu. Mais les relations entre morphologie et sens sont plus complexes que cela en réalité. En effet, il existe nombre de mots formés de manière irrégulière, comme les verbes dont les formes varient avec leur temps de conjugaison (*aller – je vais – j'irai*). Lorsque le sens d'un mot ne peut être déduit de manière régulière, on dit que mot est **opaque**.

Les morphèmes participent au sens du mot de diverses manières. Les suffixes flexionnels et dérivationnels sont essentiellement porteurs de traits relatifs à la sémantique lexicale, tandis que les radicaux portent l'information sémantique principale. L'analyse sémantique de la composition met en évidence trois types de composés :

– **Endocentriques** : le noyau sémantique se trouve dans le composé et aucun élément extérieur n'est nécessaire à sa compréhension (ex : *timbre-poste*, *oiseau-mouche*, etc.). On appelle **tête** l'élément qui apporte la contribution sémantique principale et **modifieurs** les autres éléments qui qualifient la tête. La tête se trouve à gauche dans les composés populaires comme *timbre-poste* et à droite dans les composés savants comme *hormonothérapie*.
– **Exocentriques** : le sens du composé ne peut être déduit uniquement de ses composants, il est nécessaire de supposer un élément extérieur aidant à la compréhension (ex : un *rouge-gorge* n'est pas un type de *gorge*). Les composés de ce type n'ont pas de tête.
– **Additifs** : le sens du composé est l'addition du sens des éléments (ex : *aigre-doux*, *franco-allemand*, *enseignant-chercheur*, etc.). Les composés de ce type ont deux têtes.

2.1.5 Morphologie et vocabulaire technique

Dans les domaines scientifiques ou techniques comme la médecine, de nombreux termes appartiennent au vocabulaire savant et sont formés par composition de formants classiques grecs ou latins [Cottez, 1984]. Les formants peuvent se trouver à diverses positions dans le mot, soit en tête de mot (*extra+*, *hydro+*, *pharmaco+*), soit en fin de mot (*+graphe*, *+logie*). De plus, les formants gréco-latins se présentent sous une forme caractéristique et qui plus est relativement constante à travers de nombreuses langues européennes [Namer, 2005].

Les possibilités offertes par la composition savante pour la formation de mots techniques sont très étendues et permettent la formation de nombreux néologismes. Selon [Namer et Zweigenbaum, 2004], plus de 60 % des néologismes dans le domaine bio-médical sont formés par composition savante. Ce nombre important de néologismes fait qu'il est difficile de lister l'ensemble des mots possibles et rend nécessaire l'utilisation de systèmes capables d'analyser la structure morphologique des termes [Lovis *et al.*, 1995]. Ces systèmes doivent nécessairement pouvoir analyser les procédés de dérivation, par préfixation et suffixation, et de composition.

Après avoir présenté le domaine de la morphologie du point de vue de la linguistique, nous allons maintenant décrire la manière dont cette discipline est traitée en psychologie cognitive.

2.2 La morphologie en psychologie cognitive

Les travaux de psychologie cognitive et de psycholinguistique sont centrés sur les traitements cognitifs liés à la morphologie [Seidenberg et Gonnerman, 2000]. Les questions principales, auxquelles cherchent à répondre les expériences ainsi que les théories, concernent d'une part le rôle joué par la morphologie dans l'accès au lexique mental et d'autre part l'effet de différents paramètres tels que la longueur, la fréquence ou le genre des mots sur le traitement morphologique.

Les modèles psycholinguistiques cherchent notamment à décrire les traitements allant du mot (sous sa forme orthographique ou phonétique/phonologique) au lexique mental (concept). Le rôle joué par la morphologie ainsi que l'étape à laquelle celle-ci intervient sont très divers suivant les théories. Les modèles se situent sur un continuum allant des théories holistes, selon lesquelles les mots sont représentés de manière globale, sans décomposition, aux théories décompositionnelles postulant que tous les mots sont stockés de manière décomposée dans le lexique. Entre ces deux extrêmes se trouvent les modèles hybrides, selon lesquels décomposition et non-décomposition cohabitent, voire se trouvent en compétition, suivant les mots traités. Selon ces dernières théories, les mots irréguliers et non-transparents seraient stockés de manière globale alors que la décomposition en radical et affixes serait réservée aux mots réguliers.

2.2.1 Méthodes expérimentales

Les principales méthodes utilisées en psychologie cognitive dans le domaine de la morphologie sont l'amorçage et la décision lexicale.

Décision lexicale

Les tâches de décision lexicale consistent pour les sujets à décider si une chaîne de caractères est un mot réel ou non, la durée entre la présentation du mot et la prise de décision étant mesurée. La liste de mots utilisée comprend à la fois des mots et des pseudo-mots, c'est-à-dire des séquences de lettres ou de chiffres qui ne sont pas des mots attestés, figurant dans le lexique, mais qui pourraient réellement exister. Les pseudo-mots peuvent soit être analysés comme contenant

la même racine morphologique qu'un mot réel (ex : le non-mot DEJUVENATE par rapport au mot REJUVENATE qui contient une racine) soit n'incorporer aucune racine morphologique (ex : DEPERTOIRE, qui ne contient aucune racine). D'une manière générale, les sujets mettent plus de temps à rejeter un mot tel que DEJUVENATE qu'un mot tel que DEPERTOIRE [Taft et Forster, 1975], cité par [Seidenberg et Gonnerman, 2000]. Le temps mis par le sujet est corrélé au traitement cognitif. Par conséquent, le temps de traitement est plus long pour les paires morphologiquement liées car les sujets procèdent à une tentative de décomposition du mot présenté.

Amorçage

Les tâches d'amorçage consistent en la présentation visuelle ou auditive d'un mot amorce suivi d'un mot cible. Le mot cible est un mot sur lequel le sujet doit effectuer une tâche, comme par exemple une tâche de décision lexicale. Les deux mots présentés peuvent être morphologiquement liés ou non. En général, il y a facilitation de la tâche lorsque le mot amorce et le mot cible sont reliés morphologiquement (par dérivation ou flexion). Cette facilitation n'est pas observée lorsqu'ils sont reliés uniquement par leur graphie (ex : CARD – CARS, qui constituent un couple de mots dont les graphies sont proches, mais qui ne sont pas liés sémantiquement) [Rastle *et al.*, 2000]. Cet effet a été observé dans diverses langues et il résiste à la transformation typographique (ex : alphabets différents), à la transformation phonologique (ex : HEALTH – HEAL) et à la transformation orthographique et phonologique (ex : DECISION – DECIDE). Il est également indépendant du mode de présentation visuel ou auditif de l'amorce et de la cible.

2.2.2 Paramètres influençant les traitements morphologiques cognitifs

Les expériences sont généralement basées sur des paramètres qui influencent les traitements morphologiques cognitifs tels que la fréquence, la productivité morphologique et les rôles respectifs de l'orthographe, de la morphologie et de la sémantique.

Fréquence

La fréquence d'occurrence joue un rôle dans le temps de traitement des mots morphologiquement complexes. En effet, les mots fréquents sont traités plus rapidement que les mots moins fréquents [Baayen et Schreuder, 2000]. Deux types de fréquence sont distingués en psycholinguistique. La **fréquence de surface** correspond à la fréquence d'une forme, comme par exemple *laitage*, dans la langue. La **fréquence cumulée** est la somme des fréquences de surface de tous les mots de la même famille ($F(lait) + F(laitage) + F(laitier) + F(laiterie) + ...$) [Giraudo et Grainger, 2000]. La fréquence d'usage joue un rôle important dans les tâches de reconnaissance : on reconnaît plus vite un mot fréquemment utilisé. De plus, dans les tâches de décision lexicale, on constate une asymétrie dans le rôle de la fréquence cumulée [Colé *et al.*, 1989] cités par [Meunier et Segui, 1999]. En effet, la fréquence cumulée joue un rôle pour les mots suffixés mais non pour les mots préfixés. Ceci pourrait être dû à la direction de l'analyse, procédant de la gauche vers la droite. Pour les mots suffixés, le radical est rencontré en premier, et donc l'accès à la représentation lexicale du mot peut se faire via ce radical. À l'inverse, le traitement du radical suit le traitement du préfixe pour les mots préfixés et donc les informations associées au radical ne sont pas exploitées pour l'accès au lexique mental.

Productivité morphologique

La productivité morphologique correspond au nombre de fois où un morphème est utilisé pour former de nouveaux mots. Par exemple, le suffixe anglais *+ness* est très utilisé pour dériver des noms à partir d'adjectifs (ex : *lateness*, *blackness*) tandis que le morphème *+ity* est bien moins productif (*brevity*, *specificity*). Les morphèmes liés (c'est-à-dire ceux qui doivent être attachés à un autre morphème tels que *+ceive*, *+mit*, *+cede* en anglais) se situent à la limite de la non productivité. En effet, le sens d'un mot tel que *conceive* ne peut pas se déduire de la combinaison du préfixe et de la base. De plus, la forme des morphèmes liés change généralement lorsqu'ils sont suffixés (ex : *conceive – conception*). Il semblerait donc probable que les mots complexes contenant des morphèmes liés ou d'autres morphèmes non productifs sont stockés et traités comme des ensembles non analysés plutôt que décomposés. Mais [McKinnon *et al.*, 2003] ont montré qu'un mot complexe est décomposé en morphèmes même si ces morphèmes sont non productifs et sémantiquement appauvris.

Rôles respectifs de l'orthographe, de la sémantique et de la morphologie

Des expériences d'amorçage ont montré que la reconnaissance d'un mot cible est facilitée si un mot amorce morphologiquement lié est présenté auparavant. Cependant, cet effet peut être lié aussi bien à la morphologie qu'à l'orthographe ou à la sémantique, ou encore à une combinaison des deux. Il est donc nécessaire de séparer ces sources de variation. [Rastle *et al.*, 2000] prouvent par une expérience d'amorçage variant le lien morphologique, sémantique et orthographique entre l'amorce et la cible qu'en anglais les effets de la morphologie dérivationnelle ne peuvent être réduits à une somme des effets orthographiques et sémantiques. Ceci conforte l'hypothèse d'un niveau de représentation morphologique, différent des niveaux orthographique et sémantique, jouant un rôle important dans le processus de reconnaissance des mots.

Indices statistiques

La problématique de la morphologie en psycholinguistique est aussi fortement liée à celle de la segmentation de la parole en mots [Perruchet et Peereman, 2005]. Une expérience bien connue de [Saffran *et al.*, 1996] a montré que nous exploitons les propriétés statistiques du langage pour découper le flux de parole en mots. En effet, lors d'une première expérience, les sujets adultes ont été capables de découvrir des mots dans un texte d'une langue fictive, constituée de 6 mots tri-syllabiques (*babupu*, *bupada*, *dutaba*, *patubi*, *pidabu*, *tutibu*) construits à partir de 7 phonèmes de l'anglais. Les sujets étaient d'abord exposés à un texte de cette langue fictive, lu par un synthétiseur de parole, sans pause. Dans la mesure où il s'agit d'une langue fictive, il était bien sûr impossible d'utiliser la sémantique. Après avoir écouté le texte, les sujets étaient soumis à une tâche de décision lexicale comprenant des mots de la langue fictive et des non-mots construits à partir des syllabes de la langue fictive. Les résultats de la tâche montrent que les sujets sont capables de distinguer les mots de la langue fictive. Une seconde expérience démontre que les indices prosodiques, tels que l'allongement de la dernière syllabe de chaque mot, sont également utilisés pour le découpage de la parole en mots. Les auteurs attribuent cette capacité à l'utilisation d'indices distributionnels, tels que les probabilités transitionnelles, pour l'apprentissage du langage. La probabilité transitionnelle mesure la probabilité d'observer une syllabe Y compte tenu de la syllabe précédente X :

$$\frac{\text{fréquence de la paire XY}}{\text{fréquence de X}}$$

Une valeur élevée indique que X prédit la présence de Y, tandis qu'une valeur faible indique qu'il est peu vraisemblable de rencontrer Y après X. Dans la mesure où certaines paires de syllabes sont plus fréquentes, car elles apparaissent souvent à l'intérieur de mots, la valeur de la probabilité transitionnelle sera plus grande pour ces paires de syllabes que pour des paires de syllabes qui apparaissent rarement à l'intérieur de mots et sont généralement entrecoupées par une frontière de mot.

Conclusion

Que pouvons nous tirer de cette brève incursion dans le domaine de la psychologie cognitive ? Les expériences décrites montrent que divers paramètres influencent les traitements morphologiques cognitifs et qu'ils sont pour la plupart mesurables en corpus. Ces divers éléments – fréquence, similarité orthographique et sémantique, probabilités transitionnelles – pourront donc servir à l'acquisition automatique de connaissances morphologiques à partir de corpus.

Nous allons maintenant présenter la manière dont la morphologie est abordée en traitement automatique des langues.

2.3 La morphologie en traitement automatique des langues

L'analyse morphologique est une tâche importante dans les systèmes de traitement automatique des langues comme la reconnaissance de la parole [Hacioglu *et al.*, 2003], la communication alternative et augmentée [Baroni *et al.*, 2002b], la traduction automatique [Lee, 2004, Goldwater et McClosky, 2005] ou la recherche d'informations [Daille *et al.*, 2002].

En analyse morphologique automatique, on peut distinguer diverses problématiques :

- **Désuffixation** ou **racinisation** : elle a pour objectif de supprimer la terminaison des mots. Cette méthode est surtout utilisée en recherche d'informations.
- **Lemmatisation** : elle consiste à ramener les variantes (flexionnelles) d'un même mot à une forme canonique, le lemme.
- **Analyse morphosyntaxique** : l'analyse morphosyntaxique consiste à analyser chaque mot pour lui associer divers types d'informations telles que la catégorie grammaticale, des traits morphologiques ainsi que le lemme correspondant.
- **Décomposition** : elle est surtout utilisée pour des langues comme l'allemand ou le néerlandais et consiste à segmenter un mot contenant plusieurs autres mots afin de retrouver ses composants.
- **Segmentation** : la segmentation consiste à découper un mot en segment morphémiques.

Les sections suivantes décrivent les différentes méthodes utilisées en traitement informatique de la morphologie ainsi que leurs objectifs. On trouve deux types de méthodes :

- Les méthodes qui reposent sur des dictionnaires et / ou des règles.
- Les méthodes par apprentissage, supervisé ou non supervisé, qui consistent à découvrir la morphologie d'une langue.

Lors de l'analyse des différents systèmes présentés, nous mettrons l'accent sur les propriétés suivantes, dérivées du cahier des charges que nous avons présenté dans l'introduction générale : applicabilité à diverses langues, indépendance au domaine, traitement de l'ensemble des procédés morphologiques, intervention humaine minimale.

2.3.1 Analyse morphologique à base de lexiques et de règles

Nous allons décrire trois types de systèmes d'analyse à base de lexique et de règles : les algorithmes de désuffixation, la morphologie à deux niveaux et l'analyse morphosémantique.

Désuffixation

Les méthodes de désuffixation, également connues sous le nom de racinisation (*stemming* en anglais) sont surtout utilisées en recherche d'informations. L'objectif de la désuffixation est de regrouper sous une même forme de base les variantes flexionnelles et dérivationnelles qui partagent un sens commun [Porter, 2001]. L'ensemble des variantes ainsi regroupées constitue une classe d'équivalence pour l'indexation et la recherche d'informations. L'objectif n'est pas d'effectuer une lemmatisation au sens linguistique du terme et la forme de base peut donc être un mot inexistant de la langue. Les deux algorithmes de désuffixation les plus connus sont ceux de J. Lovins et de M. Porter ou « Porter Stemmer »[Porter, 1980]. Ce dernier supprime les suffixes des mots de l'anglais au cours d'étapes successives qui peuvent également conduire à une modification de la racine. L'algorithme de Porter a été originellement élaboré pour l'anglais. Son principe est applicable à diverses langues, mais les données et l'ordre d'application des règles doivent alors être adaptées. À l'heure actuelle, il existe des versions de l'algorithme de Porter pour le français, l'espagnol, le portugais, l'italien, le suédois, le russe, le finnois etc.[1]. Le Tableau 2.1 donne quelques exemples des racines obtenues après désuffixation. L'algorithme de Porter donne généralement des résultats assez précis, mais ne couvre pas tous les cas de dérivation comme le montre la racinisation de *visionner* à *vision* tandis que la racine de *visionnage* est *visionnag*.

Mot	**Racine**
vision	vision
visions	vision
visionner	vision
visionnage	visionnag
visible	visibl
visibilité	visibil

Tab. 2.1: Exemples de racines obtenues après désuffixation avec la version française du raciniseur de Porter.

Afin d'améliorer les résultats de ce type de systèmes, [Krovetz, 1993] propose de vérifier les résultats aux différentes étapes du traitement dans un dictionnaire pour aboutir à des racines correspondant à des mots attestés de la langue (et non des racines qui n'existent pas de manière indépendante comme *visibl*).

Les effets de la désuffixation sur les performances des systèmes de recherche d'information sont toutefois assez mitigés [Tzoukermann *et al.*, 2003]. La désuffixation augmente le rappel, c'est-à-dire le nombre de documents pertinents retournés par rapport à l'ensemble des documents pertinents, mais diminue la précision, c'est-à-dire la proportion de documents pertinents parmi tous les documents retournés. [Krovetz, 1993] démontre néanmoins que la racinisation, surtout lorsque la racine est un mot attesté de la langue, améliore la performance, et ce d'autant plus que les documents sont courts, et donc contiennent peu de variantes.

[1]Les implantations du système pour diverses langues sont disponibles sur le site `http://snowball.tartarus.org/`. Elles sont également utilisables sous Python via l'outil TextIndexNG.

Transducteurs et morphologie à deux niveaux

Contrairement à la racinisation, la morphologie à deux niveaux [Koskenniemi, 1984] est une méthode non-directionnelle, c'est-à-dire qu'elle peut être utilisée à la fois pour l'analyse et la génération. Il existe deux niveaux de représentation dans ce modèle : la représentation lexicale et la représentation de surface. Le niveau de surface correspond à la forme écrite du mot tandis que le niveau lexical correspond à l'analyse souhaitée, incluant aussi bien des lettres que des symboles spécifiques encodant d'autres informations. Des règles, représentées sous forme de transducteurs, mettent ces deux niveaux en correspondance, imposant éventuellement des contraintes sur cette mise en correspondance. Par exemple, la règle suivante[1] spécifie qu'une frontière morphémique ('+') située entre une consonne sifflante à gauche et un *s* à droite correspond à un *e* dans le niveau de surface :

$$+ : e \Longleftarrow \{s\ x\ z\ [\{s\ c\}\ h]\} : \_s$$

Cette règle couvre certains des cas d'insertion d'un *e* entre un radical et un affixe flexionnel commençant par *s* en anglais (pluriel, $3^{ème}$ personne, superlatif), comme dans *bliss + s* (niveau lexical) – *blisses* (niveau de surface).

La morphologie à deux niveaux est bien adaptée aux langues fortement suffixées comme le finnois ou le turc car elle est essentiellement tournée vers le traitement de la flexion [ten Hacken et Lüdeling, 2002].

Analyse morphosémantique

Les systèmes d'analyse morphosémantique ont un double objectif : (a) effectuer l'analyse morphologique des mots et (b) définir le sens des mots en fonction de leur structure morphologique. Ces systèmes sont donc particulièrement adaptés aux applications comme la recherche d'information.

L'intérêt pour l'analyse morphosémantique est assez ancien en traitement automatique des langues comme en témoigne le système décrit par [Pratt et Pacak, 1969], appliqué à l'analyse du vocabulaire des diagnostics en anglais. Le système repose sur le lexique SNOP (Systematized Nomenclature of Pathology), qui est organisé de manière hiérarchique. L'analyseur morphologique utilise des règles de transformation des suffixes ainsi que des règles de décomposition des composés savants incluant une analyse sémantique de leurs composants.

Des systèmes plus récents visent à offrir le même type de services. [Hahn *et al.*, 2003] présentent une approche morphosémantique pour le traitement du vocabulaire médical de l'allemand. Leur système incorpore un dictionnaire de sous-mots (*subwords*) comprenant des éléments de formation d'origine grecque ou latine, des noms propres et des affixes. Les liens de synonymie entre sous-mots sont représentés de manière explicite, de sorte que les sous-mots synonymes ont un identifieur commun dans le dictionnaire. Par conséquent, l'indexation peut se faire soit sur la base des sous-mots identifiés par segmentation des mots, soit sur la base des identifieurs associés, permettant ainsi de regrouper les synonymes. [Lovis *et al.*, 1995] décrivent également un système destiné à l'analyse morphosémantique du vocabulaire médical en français. Le programme est basé sur un dictionnaire contenant un ensemble de mots et d'éléments de formation, associés à leur sens, ainsi que des éléments de liaison. Les éléments de formation sont également décrits par une étiquette syntaxique décrivant leurs possibilités de combinaison avec d'autres formants.

Le système DériF, développé par F. Namer dans le cadre du projet MorTAL (Morphologie pour le TALn) [Dal *et al.*, 2005], effectue quant à lui l'analyse morphosémantique des lemmes du français, en mettant l'accent sur l'analyse morphologique non affixale [Namer, 2003]. Ce

[1] Exemple tiré de [Trost, 2003, p. 41]

système est basé sur la théorie de la morphologie constructionnelle proposée par D. Corbin. Il a également été adapté au traitement du vocabulaire médical [Namer et Zweigenbaum, 2004]. La Figure 2.1 donne un exemple d'analyse produite par DériF, pour le lemme *encéphalite*[1]. Cette analyse comprend :

- Un arbre de l'analyse du lemme, indiqué par des crochets.
- La famille du lemme sous forme des bases successives reconnues.
- Une glose de la relation sémantique liant le lemme donné en entrée avec sa base. Le lien entre *encéphal* et *cerveau* est rendu possible grâce à une table d'environ un millier de bases non autonomes, sortes d'allomorphes supplétifs de lexèmes français.
- Les relations sémantiques candidates :
  - synonymie (eql) : une *encéphalite* est une *cérébrite*.
  - hyponymie (isa) : une *encéphalite* est un sous-type de *cranite*.
  - approximation (see) : une *encéphalite* a à voir avec (see) une *encéphalalgie/encéphalodynie* (cette relation sémantique est attestée).

L'analyse est effectuée de manière récursive, en appliquant des règles de formation de mots à l'entrée, de manière ordonnée et récursive, jusqu'à obtention d'une base non analysable. Seules les relations sémantiques concernant des lemmes attestés du lexique sont conservées.

encéphalite/NOM ==> [[encéphal N*] [ite N*] NOM]
(encéphalite/NOM, [encéphal,N*] :ite/N*)
" (Partie de – Type particulier de) inflammation en rapport avec le(s) cerveau "
Constituants = /encéphal/ite/
Type = maladie
Relations possibles = (eql :cérébr/ite, eql :cérébr/phlogo, eql :encéphal/phlogo, isa :adréno/ite, isa :adréno/phlogo, isa :crani/ite, isa :crani/phlogo, see :cérébr/alges, see :cérébr/algie, see :cérébr/odyn, see :duro/ite, see :duro/phlogo, see :encéphal/alges, see :encéphal/algie, see :encéphal/odyn)

Fig. 2.1: Exemple de résultat d'analyse morphosémantique par le système DériF.

Les systèmes d'analyse morphosémantique que nous venons de décrire présentent le point commun de décrire la langue médicale, à l'exclusion de tout autre domaine. Seul Dérif est également destiné à l'analyse de la langue générale. De plus, ils nécessitent la construction de dictionnaires de segments de mots et l'élaboration de règles d'analyse. Ce travail de description linguistique, qui dépend au moins partiellement de la langue traitée, est un processus long et coûteux. À cela s'ajoute l'absence d'analyse des mots dont les composants et procédés de formation ne sont pas explicitement décrits dans le système. Enfin, alors que les systèmes d'analyse morphosémantique traitent de la composition savante, les méthodes de désuffixation et d'analyse morphologique à deux niveaux ne traitent que de la flexion et partiellement de la dérivation.

Les méthodes d'apprentissage non supervisé visent à combler certaines de ces limites en permettant, après apprentissage, d'appliquer les régularités extraites du corpus d'apprentissage à de nouvelles données.

[1]Nous remercions Fiammetta Namer de nous avoir fourni et explicité cet exemple.

2.3.2 Apprentissage supervisé

L'apprentissage supervisé nécessite des données d'apprentissage associant les données d'entrée aux résultats désirés. Le corpus d'apprentissage restreint de fait le champ des possibilités du système : le système ne sait faire que ce qu'il a appris à faire. Les capacités du système ne dépendent donc pas uniquement de ses propriétés intrinsèques mais avant tout du contenu, de la qualité et de la taille du corpus d'apprentissage.

L'apprentissage supervisé consiste à découvrir les régularités présentes dans un ensemble d'exemples pour ensuite les appliquer à des données différentes. Les techniques d'apprentissage gardent pour ainsi dire « en mémoire » les instances rencontrées lors de l'apprentissage pour les réutiliser au moment de l'analyse.

Le système MOSES procède par apprentissage pour segmenter les mots de l'allemand. Il repose sur un corpus d'apprentissage segmenté manuellement. Diverses informations sont extraites de ce corpus lors de la phase d'apprentissage : pour chaque bigramme de lettres, le système extrait le nombres d'occurrences dans le corpus d'une frontière morphémique à gauche, au milieu et à droite du bigramme, ainsi que le nombre d'occurrences du bigramme sans aucune frontière morphémique à une de ces positions. Après apprentissage, ces valeurs sont utilisées pour déterminer les frontières morphémiques d'un mot, en calculant la somme des fréquences d'occurrence d'une frontière à toute position dans le mot. Ce même principe a été adapté au français [Janssen, 1992] et à l'espagnol [Klenk, 1992, Flenner, 1994].

Dans [van den Bosch et Daelemans, 1999] les exemples sont tirés de la base CELEX [Baayen *et al.*, 1995] pour le néerlandais. Le système d'apprentissage est basé sur le stockage en mémoire des instances de la base d'apprentissage. Chaque mot de CELEX est transformé en autant d'instances qu'il y a de lettres dans le mot selon une technique de fenêtrage centrée sur une lettre et comportant un nombre fixe de lettres voisines à droite et à gauche. De plus, à chaque instance (ou fenêtre de lettres) est associée une classe correspondant à des traits indiquant la catégorie morphosyntaxique ainsi que des informations sur la morphotactique et d'éventuels changements orthographiques. Après apprentissage, le système classe les nouvelles instances en les comparant aux instances stockées en mémoire et en calculant une distance entre la nouvelle instance et les instances en mémoire.

2.3.3 Apprentissage non supervisé

Les méthodes d'apprentissage supervisé que nous venons de présenter nécessitent des corpus d'apprentissage, constitués à partir de ressources existantes ou construites manuellement. Ces corpus déterminent ce que le système est capable de faire à l'utilisation. Or ces données ne sont pas toujours disponibles et leur construction est une tâche longue est fastidieuse, car il est nécessaire de disposer de corpus d'apprentissage de taille suffisante pour obtenir de bons résultats. Contrairement aux méthodes supervisées, les systèmes d'apprentissage non supervisé ne nécessitent pas de disposer d'exemples type des résultats souhaités. Les seules données nécessaires sont une liste de mots, éventuellement complétée par un corpus ou des ressources facilement disponibles comme des dictionnaires ou des thésaurus. On pourrait croire – à tort – que l'ensemble des informations utilisables dans une liste de mots pour effectuer une analyse morphologique complète est très limité. Pourtant la diversité des stratégies adoptées semble au contraire montrer que la plus grande difficulté réside dans la capacité à combiner l'ensemble des indices disponibles pour obtenir les meilleurs résultats possibles.

Comparaison des graphies

Une des manières les plus immédiates d'acquérir des informations sur la morphologie d'une langue est de procéder à la comparaison des graphies de mots, soit pour obtenir une mesure de leur similarité orthographique, soit pour repérer leurs différences et points communs.

Distances orthographiques Les distances orthographiques peuvent être utilisées pour retrouver des couples de mots morphologiquement liés [Adamson et Boreham, 1974]. Elles constituent toutefois une technique peu raffinée car elles ne permettent pas d'identifier la structure morphologique des mots : la comparaison de deux mots morphologiquement très éloignés ou très proches peut correspondre à la même mesure de distance orthographique. Ces mesures sont donc utilisées en combinaison avec d'autres indices, comme la mesure de la similarité contextuelle des mots comparés [Baroni *et al.*, 2002a].

Méthode du plus long préfixe ou suffixe commun Il s'agit de comparer les mots deux à deux et de repérer la plus longue chaîne initiale commune [Jacquemin, 1997, Gaussier, 1999, Zweigenbaum et Grabar, 2000]. Par exemple, le plus long préfixe commun à *chants* et *chanteurs* est *chant*. Les chaînes de caractères finales, qui distinguent les mots comparés, sont des suffixes potentiels de la langue : pour l'exemple précédent, les suffixes identifiés sont *+s* et *+eurs*. Une manière efficace d'implémenter cette méthode est d'utiliser les arbres et d'y repérer les points d'embranchement [Schone et Jurafsky, 2000, Schone et Jurafsky, 2001]. Un seuil de longueur minimale pour le plus long préfixe commun est généralement utilisé : il est de 3 pour [Grabar et Zweigenbaum, 1999, Hathout, 2002] et de 5 pour [Gaussier, 1999]. De même, les paires de suffixes extraites sont sélectionnées en fonction de leur fréquence d'occurrence : le seuil de fréquence est fixé à 2 par [Gaussier, 1999, Hathout, 2002], à 10 par [Schone et Jurafsky, 2001] et à 200 par [Schone et Jurafsky, 2000]. La même méthode peut être utilisée pour l'identification de préfixes, par la recherche du plus long suffixe commun.

Inclusion d'autres mots [Keshava et Pitler, 2006] observent que souvent, lorsque l'on enlève un suffixe à un mot, le radical restant est également un mot. Par exemple, si l'on supprime le suffixe *+ing* à la fin du mot *laughing*, le radical restant, c'est à dire *laugh* est un mot de l'anglais. La même observation peut être faite pour les préfixes : la suppression du préfixe *un+* en tête du mot *unmake* correspond à un mot attesté de l'anglais, à savoir *make*. Cette propriété ne semble toutefois pas adaptable à toutes les langues. [Demberg, 2006] constate qu'en allemand, les radicaux des verbes et des noms apparaissent rarement de manière autonome.

Analogies et différences L'utilisation du principe d'analogie permet de repérer les couples de mots qui non seulement partagent une sous-chaîne commune mais qui diffèrent également de la même manière qu'un autre couple de mot. Ce principe a été énoncé par [de Saussure, 1916, Chapitre IV] et appliqué à l'analyse morphologique automatique par [Lepage, 1998, Hathout, 2002, Lepage, 2003, Claveau et L'Homme, 2005, Hathout, 2005]. De manière plus formelle, une analogie entre 4 symboles peut se noter A : B = C : D (A est à B ce que C est à D). C'est le cas par exemple pour *exact : inexacte = fini : infinie* où *exact* est à *inexacte* ce que *fini* est à *infinie* (voir Figure 2.2). Les points communs et différences correspondent alors aux segments attendus : le préfixe *in+* et le suffixe *+e* sont identifiés par cette analogie. L'approche adoptée par [Neuvel, 2002b, Neuvel, 2002a, Neuvel et Fulop, 2002] met l'accent sur les différences plutôt que sur les analogies, même si le principe est très similaire. Ainsi, deux mots sont considérés

comme morphologiquement reliés s'ils diffèrent exactement de la même façon que deux autres mots du lexique. Par exemple, *complete* et *completely* partagent les mêmes différences que *direct* et *directly*.

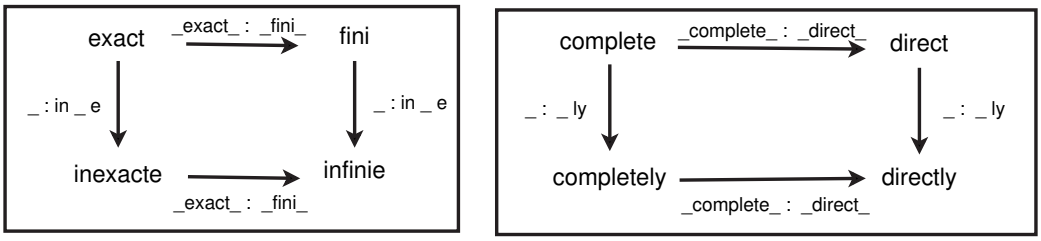


FIG. 2.2: Exemples d'analogies.

Les méthodes basées sur la comparaison de graphies sont très largement utilisées car la similarité orthographique est, avec la similarité sémantique, une des deux facettes de la notion de similarité morphologique (si l'on exclut bien sûr les cas de supplétion). Sa mise en œuvre est assez simple, du moins pour les comparaisons des débuts et fins de mots, et applicable à l'ensemble des langues formées par concaténation. Elle procède par comparaison de doublets ou quadruplets de mots, voire un plus grand nombre de mots lorsqu'ils sont insérés dans des arbres. Dans la section suivante nous allons présenter une méthode de segmentation initialement proposée par Z. Harris, qui est assez proche des méthodes basées sur la comparaison des graphies par insertion des mots dans un arbre et repérage des points d'embranchement.

La méthode de Z. Harris

La méthode proposée par [Harris, 1955] consiste à découper les représentations phonémiques des mots en utilisant le nombre de phonèmes différents qui peuvent suivre une séquence initiale de phonèmes. Cette méthode est bien sûr adaptable aux mots écrits, en comptant le nombre de lettres différentes qui peuvent suivre une chaîne de caractères initiale (*letter successor variety*). Une augmentation de ce nombre de lettres indique une frontière de morphème. Par exemple, après la séquence *direc* on ne trouve que la lettre *t*. Par contre, après *direct*, il est possible de trouver 4 lettres différentes en anglais : *i, l, o* et *e* (*direction, directly, director, directed*). Cette augmentation du nombre de lettres possibles indique une frontière entre la racine *direct* et les suffixes *+ion, +ly, +or* et *+ed*. Il est également possible de procéder en sens inverse et de calculer le nombre de lettres précédentes (*letter predecessor variety*).

Cette méthode a été utilisée et implémentée par [Hafer et Weiss, 1974]. Ils l'ont évaluée pour différentes stratégies de segmentation : utilisation du nombre de lettres suivantes et précédentes avec un seuil, combinaison des deux chiffres et calcul de l'entropie qui permet de prendre en compte à la fois la variété et la distribution des lettres et donc de pondérer l'importance des lettres sur la décision de segmentation.

[Déjean, 1998] propose une extension de l'algorithme de Z. Harris, basée sur les étapes suivantes :

1. Recherche des préfixes ou suffixes les plus fréquents en utilisant la méthode de Z. Harris. Une règle supplémentaire est ajoutée qui détecte les morphèmes les plus longs (comme par exemple le suffixe *+ion* à la place du suffixe *+on*, normalement détecté par la méthode de Harris).
2. Utilisation de ces affixes pour segmenter les mots et recherche de toutes les chaînes de caractères possibles associées à la racine ainsi obtenue. Par exemple, si l'on enlève le suffixe

+s à *lights*, il reste la racine *light*. Cette racine peut être associée à divers suffixes : *+s*, *+ed*, *+ing*, *+ness*, etc. Si la moitié de ces morphèmes correspond à des affixes identifiés à l'étape 1 alors les autres morphèmes sont validés. Par exemple, la racine *light* est associée aux suffixes *+s, +ed, +ing, +ly* et *+er* trouvés lors de la première étape ainsi qu'aux suffixes *+ness, +est* et *+en*. Ces derniers sont donc validés car ils correspondent à moins de la moitié de la liste des affixes. De plus, seuls les morphèmes qui ont une fréquence supérieure à un seuil sont conservés.

3. Segmentation des mots à l'aide de la liste de morphèmes obtenue aux étapes 1 et 2. Les mots sont segmentés en recherchant l'affixe le plus long qui correspond à la fin ou au début du mot. Les mots les plus fréquents du corpus, qui sont généralement des mots grammaticaux, ne sont pas segmentés.
4. Utilisation d'informations contextuelles pour corriger les segmentations. Les contextes correspondent à des bigrammes de morphèmes : par exemple le bigramme *des* racine *+es* correspondant au génitif en allemand. Chaque mot ou morphème apparaissant dans un tel contexte va être segmenté : par exemple, le mot allemand *Hauses* dans *des Hauses* va être segmenté en *des Haus-es* à cause du contexte *des* racine *+es*.

La méthode de Z. Harris est à rapprocher des études selon lesquelles nous segmentons les énoncés oraux en mots en nous aidant de la prévisibilité des segments suivants en fonction des segments précédents. Deux mesures sont proposées pour mesurer la prévisibilité : les probabilités transitionnelles (ou conditionnelles) [Saffran *et al.*, 1996, Keshava et Pitler, 2006] et l'information mutuelle [Brent, 1999, Rytting, 2004].

Les deux méthodes que nous allons décrire dans les prochaines sections sont assez différentes : elles visent à l'apprentissage du système morphologique d'une langue dans sa globalité. Les propriétés locales aux mots ne sont prises en compte que dans la mesure où elles participent à la cohérence et à la complétude du système global.

Segmentation par compression

Les méthodes de segmentation par compression reposent sur une idée dérivée du principe du rasoir d'Occam (ou principe d'économie), connue sous le nom de longueur minimale de description (MDL : Minimum Description Length). Le principe de la longueur minimale de description a été initialement proposé par Jorma Rissanen.

L'intuition sous-jacente au principe d'économie en morphologie est la suivante : il y a des régularités dans la langue (notamment la flexion) qui permettent par exemple aux lexicographes de « gagner de la place » dans les dictionnaires. Ainsi, on ne trouve généralement qu'une entrée par nom ou par verbe dans les dictionnaires, sans le détail de leurs différentes variantes flexionnelles. L'objectif de la segmentation est alors de trouver un dictionnaire de segments morphémiques et un encodage des mots du corpus à l'aide de ces segments qui soient les plus courts possible. Une manière simple d'estimer le « taux de compression » offert par le dictionnaire est de faire la somme du nombre de symboles contenus dans le dictionnaire et du nombre de symboles nécessaires pour l'encodage du vocabulaire. La segmentation par optimisation repose donc sur la fréquence d'apparition des segments et leur longueur : plus les segments sont longs, plus le nombre de segments contenus dans un mot est faible, ce qui favorise donc un encodage relativement court mais augmente la taille du dictionnaire. A l'inverse, plus les segments sont courts, plus le nombre de segments contenus dans un mot est important, ce qui donne un encodage long mais un dictionnaire de segments relativement court. Il s'agit donc de trouver le meilleur compromis.

Par exemple, considérons la suite de mots suivante :

`défaire brancher débrancher refaire débrancher refaire`

Il est possible de construire différents dictionnaires de segments morphémiques permettant de décrire et générer ces mots : nous présentons 4 dictionnaires possibles dans les Tableaux 2.2. Les dictionnaires associent des segments de mots à un nombre qui est utilisé pour encoder la liste de mots précédente.

Dictionnaire : 36 symboles	**Encodage :** 6 symboles
1 défaire 2 brancher 3 débrancher 4 refaire	1 2 3 4 3 4

(a) Analyse 1.

Dictionnaire : 21 symboles	**Encodage :** 11 symboles
1 faire 2 dé 3 brancher 4 re	1*2 3 2*3 4*1 2*3 4*1

(b) Analyse 2.

Dictionnaire : 28 symboles	**Encodage :** 8 symboles
1 défaire 2 dé 3 brancher 4 refaire	1 3 2*3 4 2*3 4

(c) Analyse 3.

Dictionnaire : 22 symboles		**Encodage :** 53 symboles
1 d	7 e	1*2*3*4*5*6*7 8*6*4*9*10*11*7*6
2 é	8 b	1*2*8*6*4*9*10*11*7*6
3 f	9 n	6*7*1*2*3*4*5*6*7
4 a	10 c	1*2*8*6*4*9*10*11*7*6
5 i	11 h	6*7*1*2*3*4*5*6*7
6 r		

(d) Analyse 4.

Tab. 2.2: Exemples d'analyses MDL : (a) **Analyse 1.** Longueur totale : 36 + 6 = 42 symboles ; (b) **Analyse 2.** Longueur totale : 21 + 11 = 32 symboles ; (c) **Analyse 3.** Longueur totale : 28 + 8 = 36 symboles ; (d) **Analyse 4.** Longueur totale : 22 + 53 = 75 symboles.

Les analyses s'ordonnent comme suit, celle qui obtient la longueur la plus courte étant la meilleure : Analyse 2 > Analyse 3 > Analyse 1 > Analyse 4. La meilleure analyse, celle qui donne lieu à une longueur minimale de description est donc l'analyse 2, qui correspond à la meilleure segmentation.

Le calcul de la longueur de description que nous avons utilisé pour les besoins de l'exemple est bien sûr simpliste et les formules effectivement utilisées sont généralement plus complexes.

De plus, il n'est pas possible de tester tous les dictionnaires possibles. Il est donc nécessaire d'employer des heuristiques pour guider la recherche du meilleur modèle. La méthode de segmentation basée sur la longueur minimale de description est utilisée par de nombreux auteurs, parmi lesquels [Brent et Cartwright, 1996, Kazakov, 1997, Kit et Wilks, 1999, Goldsmith, 2001, Creutz et Lagus, 2002, Baroni, 2003, Argamon *et al.*, 2004].

Nous allons plus particulièrement détailler le système Linguistica [Goldsmith, 2001], qui présente une caractéristique intéressante. En effet, Linguistica combine l'utilisation du principe MDL avec la notion de *signature* qui correspond à une liste d'affixes qui apparaissent avec le même radical dans un corpus. La signature du radical *betray* est `NULL.ed.ing` (où `NULL` est la chaîne vide), ce qui signifie que l'on trouve les mots *betray*, *betrayed* et *betraying* dans le corpus. Plusieurs radicaux partagent la même signature : la signature `NULL.ed.ing` s'applique également aux radicaux suivants : *applaud*, *arrest*, *astound*, etc.

Dans un premier temps, Linguistica propose une segmentation des mots en utilisant des heuristiques. Une de ces heuristiques repère les chaînes de caractères finales, d'une longueur comprise entre 1 à 6, qui sont susceptibles d'être des suffixes de la langue, grâce à une mesure appelée information mutuelle pondérée. Les mots sont ensuite décomposés en radical et suffixe. Il est alors possible de former les signatures de suffixes associées à chaque radical. Les signatures qui n'apparaissent qu'avec un seul radical et celles qui ne contiennent qu'un seul suffixe sont éliminées. Enfin, des heuristiques sont appliquées pour améliorer les segmentations obtenues et les résultats de l'application de ces heuristiques sont évalués par le principe de la longueur minimale de description : chaque nouvelle analyse n'est adoptée que si elle permet de réduire la longueur de la description.

Modèles probabilistes

Les méthodes basées sur des modèles probabilistes sont assez proches de celles qui reposent sur le principe d'économie car elles permettent de sélectionner un modèle des données parmi un ensemble de modèles possibles en ordonnant les différentes interprétations en fonction de leur vraisemblance.

Reprenons les exemples d'analyse précédents[1]. Nous allons tout d'abord estimer les probabilités des dictionnaires donnés pour chaque analyse : c'est ce que l'on appelle la probabilité a priori des dictionnaires. La probabilité d'un dictionnaire dépend de sa taille. Plus le dictionnaire est grand, plus le nombre de configurations possibles des segments qu'il contient est important, et plus la probabilité d'avoir un tel dictionnaire est faible. Si l'on considère qu'il y a 40 caractères en français, y compris les caractères accentués, alors la probabilité d'avoir le dictionnaire 1, qui comprend 36 symboles est $\mathrm{P}(\text{Dictionnaire 1}) = \left(\frac{1}{40}\right)^{36} \approx 2,1.10^{-58}$. Nous pouvons de manière similaire calculer la probabilité des autres dictionnaires : $\mathrm{P}(\text{Dictionnaire 2}) = \left(\frac{1}{40}\right)^{21} \approx 2,3.10^{-34}$, $\mathrm{P}(\text{Dictionnaire 3}) = \left(\frac{1}{40}\right)^{28} \approx 1,4.10^{-45}$ et $\mathrm{P}(\text{Dictionnaire 4}) = \left(\frac{1}{40}\right)^{22} \approx 5,7.10^{-36}$. Ces probabilités permettent d'ordonner les dictionnaires : $\mathrm{P}(\text{Dictionnaire 2}) > \mathrm{P}(\text{Dictionnaire 4}) > \mathrm{P}(\text{Dictionnaire 3}) > \mathrm{P}(\text{Dictionnaire 1})$.

On peut également calculer les probabilités des analyses données en fonction de ces dictionnaires. Dans l'analyse 1, le dictionnaire contient 4 segments. Si l'on considère que la distribution de probabilité de ces segments est uniforme, alors la probabilité d'observer un de ces segments est égale à $\frac{1}{4}$ et donc la probabilité d'observer la suite de mots considérée, en fonction de ce dictionnaire est $\mathrm{P}(\text{Données}|\text{Dictionnaire 1}) = \left(\frac{1}{4}\right)^{6} \approx 2,4.10^{-4}$ car il y a 6 occurrences de segments dans la séquence de mots. Pour l'analyse 2, $\mathrm{P}(\text{Données}|\text{Dictionnaire 2}) =$

[1] Les explications données dans cette section sont inspirées de [Creutz, 2006, Section 2.2.3]

$\left(\frac{1}{4}\right)^{11} \approx 2,4.10^{-7}$. Pour l'analyse 3, P(Données|Dictionnaire 3) $= \left(\frac{1}{4}\right)^{8} \approx 1,5.10^{-5}$. Et enfin, pour l'analyse 4, P(Données|Dictionnaire 4) $= \left(\frac{1}{11}\right)^{53} \approx 6,4.10^{-56}$. Ces analyses peuvent être ordonnées en fonction de leur probabilité : P(Données|Dictionnaire 1) > P(Données|Dictionnaire 3) > P(Données|Dictionnaire 2) > P(Données|Dictionnaire 4). Dans ce cas, c'est l'analyse 1 qui obtient la plus grande probabilité par rapport aux données, c'est-à-dire le maximum de vraisemblance.

Cependant, il faut également que le dictionnaire ou modèle sélectionné puisse servir à faire des prédictions en présence de données inconnues. En effet, l'analyse 1 prédira une probabilité nulle pour le mot *rebrancher* alors même qu'il est vraisemblable. À l'inverse, l'analyse 2 va donner la même probabilité au mot *rebrancher* qu'au mot *débrancher* : cette analyse est un bon compromis entre complexité du dictionnaire et adéquation aux données. Pour trouver la meilleure analyse, il faut combiner la probabilité du dictionnaire (probabilité a priori) et celle de l'analyse en fonction du dictionnaire. Il s'agit donc de maximiser la probabilité postérieure du modèle, qui se calcule comme suit : $P(\text{Dictionnaire } X|\text{Données}) = P(\text{Dictionnaire } X) \cdot P(\text{Données}|\text{Dictionnaire } X)/P(\text{Données})$. En pratique, on ignore la probabilité des données, car celle-ci est constante pour tous les modèles et n'affecte donc pas le résultat de la comparaison. La probabilité postérieure se calcule donc de la manière suivante : $P(\text{Dictionnaire } X|\text{Données}) \propto P(\text{Dictionnaire } X) \cdot P(\text{Données}|\text{Dictionnaire } X)$

On trouve les valeurs de probabilité suivantes pour les différents modèles :
P(Dictionnaire 1|Données) $= \left(\frac{1}{40}\right)^{36} \cdot \left(\frac{1}{4}\right)^{6} \approx 5,2.10^{-62}$,
P(Dictionnaire 2|Données) $= \left(\frac{1}{40}\right)^{21} \cdot \left(\frac{1}{4}\right)^{11} \approx 5,4.10^{-41}$,
P(Dictionnaire 3|Données) $= \left(\frac{1}{40}\right)^{28} \cdot \left(\frac{1}{4}\right)^{8} \approx 2,1.10^{-50}$,
P(Dictionnaire 4|Données) $= \left(\frac{1}{40}\right)^{22} \cdot \left(\frac{1}{11}\right)^{53} \approx 3,6.10^{-91}$.
Les probabilités postérieures de chaque modèle s'ordonnent comme suit :
P(Dictionnaire 2|Données) > P(Dictionnaire 1|Données) > P(Dictionnaire 3|Données) > P(Dictionnaire 4|Données).

Le dictionnaire qui permet la meilleure analyse est donc le dictionnaire 2 : ce résultat est identique à celui obtenu par la méthode de la longueur minimale de description. Cette manière de procéder pour trouver la meilleure analyse est appelée inférence Bayésienne.

L'inférence Bayésienne est utilisée dans les versions les plus récentes du système Morfessor développé par M. Creutz, en collaboration avec K. Lagus [Creutz, 2003, Creutz et Lagus, 2004, Creutz et Lagus, 2005]. La première version du système [Creutz et Lagus, 2002] repose sur le principe de la description de longueur minimale et présente quelques défauts. En effet, les mots les plus fréquents ne sont pas toujours segmentés par une méthode basée sur MDL, car il est plus efficace, pour compresser les données, de stocker directement ces mots dans le dictionnaire. À l'inverse, les mots rares sont parfois sur-segmentés. Le système a donc été amélioré par l'utilisation du principe d'inférence Bayésienne.

Le principe de l'inférence Bayésienne nécessite une estimation de la probabilité a priori du modèle. Dans [Creutz, 2003], les probabilités a priori sont basées sur la distribution des fréquences et des longueurs des segments. La distribution a priori de la fréquence des segments est dérivée de la loi de Zipf, selon laquelle il y a un petit nombre d'éléments très fréquents, comme par exemple les affixes flexionnels (ex : *+ed* ou *+ing*), et un grand nombre d'éléments très peu fréquents (ex : *abacus*). La distribution a priori de la longueur des morphes est quant à elle basée sur la distribution gamma, qui donne une bonne approximation de la distribution des longueurs des morphèmes en finnois et en anglais. L'utilisation d'heuristiques simples basées notamment sur des règles de morphotactique [Creutz et Lagus, 2004] ou la structuration hiérarchique de la

segmentation de chaque mot [Creutz et Lagus, 2005] permettent d'améliorer encore les résultats de la segmentation produite.

Les systèmes Morfessor ont été utilisés pour segmenter des mots en anglais et en finnois, ainsi qu'en turc, suédois, russe et estonien. Dans les versions les plus récentes du système, les segments sont en plus étiquetés par l'une des catégories suivantes : préfixe, base et suffixe.

Utilisation du contexte d'occurrence

Nous avons vu en introduction de ce chapitre que les morphèmes allient forme et sens. Or les méthodes d'apprentissage non supervisé décrites jusqu'à présent reposent uniquement sur la graphie, sans prendre en compte la dimension sémantique inhérente aux morphèmes. Ceci peut conduire à divers problèmes, identifiés par [Schone et Jurafsky, 2000] :

- Des affixes par ailleurs valides peuvent être appliqués de manière inadéquate (*ally* découpé en *all* + *y*).
- Des cas d'ambiguïtés morphologiques peuvent apparaître (*rating* lié à *rat* plutôt qu'à *rate*).
- Des affixes non productifs peuvent ne pas être identifiés (la relation entre *dirty* et *dirt* risque de ne pas apparaître).

Les performances des algorithmes décrits précédemment peuvent être améliorées en utilisant des mesures de similarité sémantique suivant l'idée que les mots morphologiquement reliés sont similaires à la fois orthographiquement et sémantiquement. En effet, il y a beaucoup de mots qui sont orthographiquement similaires mais qui ne sont pas reliés morphologiquement (comme par exemple *bleu* et *creu*). De plus, les mots qui sont reliés sémantiquement ne sont pas forcément reliés morphologiquement (comme par exemple *bleu* et *vert*). Cependant, si deux mots sont reliés sémantiquement et orthographiquement (comme par exemple *vert* et *verdâtre*) ils sont certainement également reliés morphologiquement [Baroni *et al.*, 2002a].

Les systèmes qui intègrent des informations contextuelles, comme approximation de la similarité sémantique, peuvent être analysés à partir des critères suivants :

- **Co-occurrences de premier ou de second ordre**. Les co-occurrences de premier ordre, qui correspondent à une association syntagmatique, sont utilisées par [Xu et Croft, 1998, Baroni *et al.*, 2002a, Zweigenbaum *et al.*, 2003]. [Schone et Jurafsky, 2000, Schone et Jurafsky, 2001, Bordag, 2005] utilisent quant à eux les co-occurrences de second ordre.
- **Mesures de la similarité contextuelle**. Nous avons vu qu'il existe diverses mesures permettant de calculer la force d'association entre mots. [Zweigenbaum *et al.*, 2003, Bordag, 2005] mesurent la force de co-occurrence par le rapport de vraisemblance, [Xu et Croft, 1998] et [Baroni *et al.*, 2002a] utilisent l'information mutuelle.
- **Intégration des informations contextuelles en début ou en fin de processus**. La plus grande difficulté réside dans l'intégration des résultats de l'analyse graphique et de l'analyse contextuelle. Lorsque les informations contextuelles sont utilisées en fin de processus, elles permettent de pondérer les relations morphologiques découvertes ou de rejeter certaines analyses [Schone et Jurafsky, 2000, Schone et Jurafsky, 2001, Baroni *et al.*, 2002a, Zweigenbaum *et al.*, 2003]. Lorsqu'elles sont utilisées en début de processus, elles permettent de contraindre l'analyse morphologique basée sur la comparaison des graphies et donc d'obtenir une plus grande précision. La méthode proposée par [Freitag, 2005] consiste à classer automatiquement les mots dans des classes syntaxiques. Les membres de classes différentes sont ensuite comparés deux à deux afin de découvrir des paires d'affixes. [Bordag, 2005] applique quant à lui la méthode de Harris en restreignant les mots utilisés pour le calcul des transitions aux mots les plus proches du mot cible. La proximité est

calculée par une comparaison des vecteurs de co-occurrence de chaque mot, pondérés par le rapport de vraisemblance.

Le corpus peut être combiné à d'autres informations restreignant encore davantage le contexte d'occurrence. [Jacquemin, 1997] complète l'utilisation d'un corpus par une liste de termes polylexicaux. Dans un premier temps, le système extrait des paires de mots graphiquement similaires qui partagent un préfixe identique. Puis, les termes polylexicaux composés de deux mots sont regroupés avec des suites de mots issues du corpus, de longueur maximale égale à 4 et contenant deux mots graphiquement similaires aux mots du terme. C'est le cas par exemple pour *require-s the use* (suite de mots issue du corpus) et *use-r require-ment* (terme). Les meilleurs résultats sont obtenus pour des suites de mots contiguës, sans permutation et pour une longueur maximale du suffixe égale à 3. Des classes de groupes de termes sont ensuite constituées en fonction de la similarité des suffixes (comme par exemple *active immuniz-ation / active-ly immuniz-ed* et *chemical transform-ation / chemica-ly transform-ed*). Les classes incorrectes sont filtrées à l'aide d'une mesure de la qualité de la classe reposant sur la longueur des suffixes et des chaînes de caractères partagées. Enfin, un algorithme de classification permet de regrouper les classes correspondant à des relations morphologiques similaires, dont les suffixes présentent des alternances.

Utilisation de ressources lexicales

Les informations contextuelles permettent l'acquisition de données morphologiques à partir de mots dont la proximité sémantique est établie et augmentent ainsi la précision du système. Des ressources lexicales et sémantiques, lorsqu'elles sont disponibles, peuvent remplir le même rôle, de manière encore plus efficace car les liens lexicaux et sémantiques qu'elles contiennent ont été validés manuellement.

Par exemple, [Gaussier, 1999] utilise des lexiques flexionnels qui permettent de compléter les suffixes extraits par la méthode du plus long préfixe commun avec leur catégorie syntaxique. Le système DéCor [Hathout, 2005] se base également sur un lexique flexionnel de référence, contenant des lemmes associés à leur catégorie morphosyntaxique pour apprendre un ensemble de schémas de préfixation et de suffixation par analogie. Ces schémas sont ensuite appliqués aux lemmes. Enfin les relations morphologiques induites sont filtrées par des critères statistiques pour ne conserver qu'une seule base par lexème.

Cependant, les informations d'ordre morpho-syntaxique s'avèrent parfois insuffisantes et d'autres méthodes leur préfèrent donc des ressources sémantiques.

[Grabar et Zweigenbaum, 1999, Zweigenbaum et Grabar, 2000] constatent que les éléments d'une terminologie (dans ce cas, le répertoire d'anatomopathologie de la SNOMED) partagent régulièrement des liens morphologiques lorsqu'ils sont également reliés par des liens hiérarchiques ou transversaux. Deux procédés morphologiques majeurs sont distingués :

- Dérivation par affixation. Un affixe (préfixe ou suffixe) est ajouté à une base (mot initial) pour aboutir à un mot construit : *salive / salivaire*, *rayon / rayonnement*.
- Composition. Plusieurs radicaux, et éventuellement des affixes, sont combinés pour obtenir un mot construit : *sarcome / chondrosarcome*, *névrose / aponévrose / aponévrosite*.

Dans un premier temps, les termes de la SNOMED sont étiquetés et lemmatisés. Puis une liste de référence de près de 9 000 formes est constituée à partir de la SNOMED et de la Classification Internationale des Maladies. Les couples de termes sémantiquement liés sont ensuite comparés afin d'identifier les liens morphologiques. Deux mots sont considérés comme possédant un lien morphologique hypothétique s'ils possèdent une chaîne de caractères initiale commune assez longue (cette longueur est fixée par un paramètre). Chaque couple de mots ainsi liés donne lieu à une règle, consistant en une paire de chaînes finales (par exemple *sinus, sinusite*$\rightarrow \varepsilon|ite$).

Toutes les règles sont conservées et sont ensuite appliquées aux mots de la liste de référence pour repérer des couples de mots morphologiquement liés. Par transitivité, il est possible d'obtenir des familles morphologiques à partir de ces couples.

[Hathout, 2002] se base quant à lui sur une ressource générale, WordNet, pour acquérir des analogies « morpho-synonymiques » telles que *abandon/V : abandonment/N = desert/V : desertion/N* où (*abandon/V,desert/V*) et (*abandonment/N,desertion/N*) sont des couples de synonymes. L'utilisation de la relation de synonymie permet ici de raffiner les suffixes acquis par la méthode du plus long préfixe commun.

2.3.4 Évaluation

L'évaluation des résultats de l'apprentissage peut être effectuée de diverses manières, parmi lesquelles on trouve bien sûr l'évaluation manuelle des résultats. Lorsque l'évaluation repose sur des ressources standard, il convient de distinguer deux types de méthodes : l'évaluation directe des segments produits et l'évaluation des liens morphologiques induits. Dans le premier cas, ce sont les positions des frontières morphologiques identifiées par la méthode qui sont évaluées. Dans le second cas, la pertinence des liens morphologiques induits entre mots partageant un même radical est mesurée. Ces deux méthodes peuvent être complétées par une évaluation des apports de l'analyse morphologique dans une application réelle comme la recherche d'informations par exemple.

Évaluation des segments

Une des ressources les plus utilisées pour l'évaluation des résultats de la segmentation morphologique est la base CELEX [Baayen *et al.*, 1995], disponible pour l'anglais, l'allemand et le néerlandais. La base contient notamment une analyse de la structure morphologique des lemmes, comme par exemple `((dis)[N|.N],((engage)[V],(ment)[N|V.])[N])[N]`.

La base CELEX est notamment utilisée pour l'évaluation des systèmes décrits dans les articles suivants : [Bordag, 2005] et [Freitag, 2005]. [Creutz et Lindén, 2004] décrivent le système *Hutmegs* qui contient les segmentations morphologiques de 1,4 million de mots finnois et 120 000 mots anglais. Ces segmentations standard ont été obtenues à partir de CELEX pour l'anglais et l'analyseur morphologique à deux niveaux FINTWOL pour le finnois. Hutmegs fournit également des scripts pour l'évaluation automatique des segmentations produites par un algorithme quelconque.

Évaluation des liens morphologiques

Les ressources existantes peuvent également servir à l'évaluation des liens morphologiques. [Baroni *et al.*, 2002a] utilisent l'analyseur de XEROX pour l'anglais et l'allemand. Dans la mesure où cet outil effectue uniquement une analyse de la morphologie flexionnelle, les auteurs ont également analysé manuellement une partie des résultats afin d'obtenir une meilleure estimation de la précision de leur système. La base CELEX est quant à elle utilisée par [Hathout, 2002, Schone et Jurafsky, 2000, Schone et Jurafsky, 2001].

Évaluation par utilisation dans une application

L'une des applications les plus anciennes de l'analyse morphologique, et particulièrement la racinisation, est la recherche d'informations. Le regroupement des mots en familles morphologiques sémantiquement homogènes doit permettre l'augmentation du rappel, c'est-à-dire le

nombre de documents pertinents retrouvés. Les tâches de recherche d'informations constituent donc un moyen d'évaluer les résultats de l'analyse morphologique [Krovetz, 1993]. Il existe deux moyens d'incorporer les résultats d'une analyse morphologique à la recherche d'information [Bilotti *et al.*, 2004]. La première manière de procéder consiste à étendre la requête avec les mots identifiés comme morphologiquement liés aux mots de la requête [Moreau et Claveau, 2006]. La seconde consiste à indexer les documents par les racines ou les segments morphémiques et à procéder de même pour les requêtes pour pouvoir comparer requête et documents [Hahn *et al.*, 2003].

2.4 Conclusion

Nous venons de présenter le domaine de la morphologie à travers le point de vue de trois disciplines. Chacune d'entre elles apporte un éclairage différent :

- La linguistique fournit le vocabulaire et les outils de description nécessaires au travail sur la morphologie.
- La psychologie cognitive donne un aperçu des indices utilisés pour le traitement cognitif de la morphologie. Ces indices, comme la fréquence, la similarité graphique, la productivité ou les probabilités transitionnelles, peuvent être extraits des corpus textuels et donc être utilisés pour élaborer des systèmes d'apprentissage de connaissances morphologiques.
- Enfin, le traitement automatique des langues démontre par la diversité des approches que l'analyse morphologique est une tâche complexe. La complexité est bien sûr dépendante des objectifs fixés au système. La finesse de l'analyse produite est un premier paramètre. Les systèmes les plus simples ne traitent que de la flexion, les plus complets intègrent également la composition, voire une analyse morphosémantique. Les langues traitées diffèrent également par leur niveau de complexité morphologique : la morphologie de l'anglais est moins complexe que celle du finnois.

Rappelons brièvement les objectifs énoncés dans l'introduction générale : analyse non supervisée, à partir de corpus, indépendance aux langues et traitement du vocabulaire technique. Les systèmes capables de réaliser ce genre de tâche à l'heure actuelle sont rares : la grande majorité de ceux que nous avons présentés sont limités soit dans les langues ou les domaines qu'ils sont capables de traiter, soit dans les procédés morphologiques qu'ils peuvent analyser.

Nous avons donc élaboré notre propre système d'analyse morphologique non supervisée. Nous avons en fait testé deux approches différentes, implémentées dans deux systèmes. Le premier procède par segmentation et se rapproche ainsi du système Morfessor [Creutz et Lagus, 2005]. Le second procède par classification et groupe les mots en classes similaires aux classes d'équivalences produites par les méthodes de désuffixation. Nous allons décrire ces deux systèmes dans la partie suivante, après avoir présenté la méthode d'acquisition des données d'apprentissage.

Deuxième partie

Apprentissage de connaissances morphologiques

Introduction

L'intégration de la morphologie au processus global d'acquisition de ressources lexicales à partir de textes suppose la disponibilité d'un système d'analyse morphologique complet, capable de traiter la langue et le domaine cibles, sans pour autant en être dépendant, afin de garantir la réutilisabilité de la méthode.

Afin de produire un système capable de traiter les données issues de corpus de textes techniques et spécialisés, il faut tenir compte des spécificités du vocabulaire technique. Un tel système doit donc posséder les propriétés suivantes :

1. Traitement de la flexion, de la dérivation et de la composition. Ces différents procédés présentent un degré de complexité croissant.
2. Indépendance au domaine traité. Il existe des systèmes d'analyse de la langue médicale, basés sur des dictionnaires d'éléments morphologiques. Ces dictionnaires sont donc limités au domaine de la médecine et les analyses ne sont donc pas transposables à d'autres domaines.
3. Indépendance à la langue. Il doit être possible de traiter, au minimum, des données en anglais, en français, et éventuellement d'autres langues.

Ces besoins imposent une approche par apprentissage non supervisé, c'est-à-dire que les seules données d'apprentissage sont issues d'un corpus de textes bruts, sans aucune règle ou lexique spécifique à la langue ou au domaine.

Nous allons tout d'abord décrire la méthodologie de construction de corpus que nous avons employée. Celle-ci procède par acquisition automatique de textes spécifiques aux domaines traités à partir d'Internet. La méthode est totalement automatisée.

Nous allons ensuite décrire les systèmes d'analyse morphologique que nous avons élaborés : le premier, présenté dans le Chapitre 4, découpe les mots en segments étiquetés. Le second, présenté dans le Chapitre 5, regroupe les mots dans des familles morphologiques.

Chapitre 3

Préliminaire : construction de corpus

3.1 Un corpus : oui, mais lequel ? ...

Les corpus disponibles, que ce soit pour l'anglais ou le français, sont généralement à vocation généraliste. Ainsi, pour l'anglais, on dispose du corpus de référence British National Corpus (BNC). En français, on trouve diverses bases de textes, regroupant essentiellement des textes littéraires, des articles de journaux et des œuvres libres de droit, comme par exemple Frantext qui contient plus de 217 millions d'occurrences, avec une majorité d'œuvres littéraires ou la base des articles du quotidien *Le Monde*. Ces corpus permettent un travail sur la langue générale mais sont malheureusement inadaptés au travail sur des domaines spécialisés que nous voulons réaliser. En effet, nous avions pour objectif de concevoir des méthodes d'acquisition automatique de connaissances valables pour plusieurs langues (français, anglais, ...) et, à l'intérieur même de ces langues, pour des domaines spécialisés variés (médecine, sciences de la terre, ...). Nous ne pouvions donc utiliser des corpus généralistes existant comme le BNC pour l'anglais ou le corpus de *Le Monde* pour le français. De plus, nous voulions également tester des méthodes utilisant peu de connaissances préalables et reposant essentiellement sur les données présentes dans le corpus (approche endogène). Ce type d'approche nécessite des corpus à l'échelle du million de mots, voire plus. Compte tenu de ces objectifs et des moyens dont nous disposions, nous avons choisi de construire automatiquement nos corpus de textes à partir du Web.

3.2 ... Pourquoi pas le Web ?

La construction automatique de corpus à partir du Web est un domaine de recherche très actif à l'heure actuelle, comme en témoignent les nombreux ouvrages [Kilgarriff et Grefenstette, 2003], ateliers et colloques organisés à ce sujet. Nous pouvons notamment citer les ateliers « Web as Corpus » organisés en juillet 2005 à Birmingham au Royaume-Uni (conférence *Corpus Linguistics*) et avril 2006 à Trento en Italie (conférence *EACL*), et le projet WaCky pour *Web-as-Corpus kool ynitiative* (`http://wacky.sslmit.unibo.it/`).

3.2.1 Avantages et inconvénients

Le formidable essor de la thématique du « Web comme corpus » s'explique aisément. En effet, le Web offre l'accès à une énorme quantité de documents (qu'il s'agisse d'ailleurs de documents textuels, d'images, de vidéos ou de sons), dans un nombre sans cesse croissant de langues. Ces documents sont gratuits, directement et immédiatement accessibles dans un format électronique

utilisable pour des traitements automatiques et de genres très divers : littérature, articles scientifiques ou de vulgarisation, journaux, ainsi que des genres qui ne sont pas traditionnellement représentés dans les corpus comme les blogs ou les forums de discussion plus proches de la langue orale. Le Web permet donc de constituer rapidement des corpus ad hoc pour des applications spécifiques ou des domaines spécialisés [Baroni et Bernardini, 2004] ou encore des corpus pour des langues minoritaires [Ghani *et al.*, 2001, Naets, 2005].

L'utilisation du Web comme corpus s'accompagne bien sûr également d'un questionnement sur la manière d'utiliser le Web pour cette application et sur la pertinence même d'une telle utilisation. Les plus réfractaires considéreront d'ailleurs que le Web n'est pas un corpus à proprement parler.

Les problèmes les plus importants sont peut-être les suivants :

- **Problèmes juridiques :** Il est difficile de trouver des réponses claires aux questions juridiques de droit d'auteur. [Kilgarriff et Grefenstette, 2003] éludent le problème en argumentant que n'importe quel corpus constitué à partir du Web n'est qu'une sous partie de tous les index et autres pages sauvegardées en cache par les moteurs de recherche comme Yahoo ou Google. Ces moteurs de recherche commerciaux enfreindraient donc les lois à une plus grande échelle. Certaines initiatives comme les licences Creative Commons pour les œuvres intellectuelles (`http://fr.creativecommons.org`) tentent de mettre un peu d'ordre dans le flou qui semble régner sur le Web afin d'encadrer juridiquement la diffusion d'œuvres en ligne. Cette initiative est désormais relayée par le moteur de recherche Yahoo via le site `http://search.yahoo.com/cc` qui permet la collecte de documents accompagnés d'une licence Creative Commons. D'une manière générale, pour un corpus constitué à partir d'Internet, seule la liste des URLs des documents constituant le corpus est libre de droits et donc distribuable.
- **Problèmes techniques :** Les données extraites à partir du Web sont parfois fortement bruitées. Ce bruit se retrouve à différents niveaux de la structure linguistique et documentaire. Ainsi, on trouve nombre de fautes d'orthographe ou mots à l'orthographe adaptée pour une écriture rapide (langage SMS ou langage tchaté), même si comme le remarquent [Kilgarriff et Grefenstette, 2003, Liu et Curran, 2006], la proportion de formes erronées est négligeable par rapport à la proportion de formes correctes. De plus, le contenu d'une page Web s'accompagne également d'éléments de navigation ou de publicités qu'il faut éliminer. Il est donc nécessaire d'effectuer un certain nombre de traitement sur les textes pour les rendre utilisables. Ces différents traitements seront détaillés dans la section 3.3.

3.2.2 Approches du Web comme corpus

Il existe diverses manières d'utiliser le Web comme source de données linguistiques. [Baroni et Ueyama, 2006] distinguent trois approches principales, selon qu'elles utilisent ou non des moteurs de recherche :

1. Utilisation directe d'un moteur de recherche et estimation de la fréquence d'occurrence de l'expression recherchée par le nombre de pages retournées par le moteur de recherche [Turney, 2001, Léon et Millon, 2005]. Cette approche présente de sérieuses limites. En effet, il est impossible d'effectuer des requêtes complexes, utilisant les expressions régulières par exemple. De plus, Jean Véronis a montré sur son blog[1] que les décomptes du nombre de résultats donnés par le moteur de recherche Google sont plus qu'approximatifs.

[1] `http://aixtal.blogspot.com/`

2. Construction de corpus à partir de requêtes sur des moteurs de recherche : cette approche sera détaillée dans la section 3.3.
3. Utilisation de robots de parcours du Web [Vaufreydaz, 2002, Baroni et Ueyama, 2006, Liu et Curran, 2006]. L'avantage de cette approche est qu'elle est indépendante des moteurs de recherche commerciaux.

3.3 Constitution automatique de corpus à partir du Web

Dans les sections suivantes, nous détaillons les différentes étapes de construction de corpus à partir du Web en passant par des requêtes sur des moteurs de recherche. Cette approche a notamment été utilisée par [Ghani *et al.*, 2001, Baroni et Bernardini, 2004]. La méthodologie décrite s'inspire fortement de la méthode BootCat [Baroni et Bernardini, 2004], notamment en ce qui concerne les étapes de collecte d'URLs et d'extraction du contenu des documents HTML. L'ensemble des fonctionnalités décrites a été implémenté en Python (voir Annexe B.2.1, p. 153).

3.3.1 Collecte d'URLs

La première étape de la construction d'un corpus à partir du Web est la collecte d'un ensemble d'URLs. Il est certes possible d'effectuer ce travail manuellement, en utilisant des moteurs de recherche via un navigateur Web, mais ce travail est fastidieux si l'on veut constituer un corpus de taille importante. Il existe maintenant des APIs permettant d'interroger automatiquement divers moteurs de recherche, sans passer par un navigateur Web. Nous pouvons citer les APIs suivantes :

– Google Web APIs : `http://www.google.com/apis/index.html`
– Yahoo ! Search Web Services : `http://developer.yahoo.net/search/index.html`

Ces APIs retournent une liste d'URLs correspondant à une requête. Il est également possible de paramétrer les requêtes pour que le moteur de recherche ne renvoie que des documents écrits dans une des langues qu'il est capable de reconnaître (ce qui exclut bien sûr les langues minoritaires), ainsi que des formats de fichiers spécifiques (HTML, Microsoft Word, pdf, texte, xls, etc.).

Les requêtes utilisées pour l'acquisition d'URLs ont une grande influence sur la qualité et l'adéquation des documents retournés par le moteur de recherche. Que ce soit pour construire des corpus spécialisés, des corpus de langues minoritaires ou des corpus généraux, plusieurs mots sont généralement combinés au sein d'une même requête afin d'augmenter la précision. La sélection de ces mots est dépendante du type de corpus souhaité :

– Corpus spécialisé : liste de mots caractéristiques du domaine. Selon [Baroni et Bernardini, 2004], une liste de 5 à 15 mots est suffisante.
– Corpus de langue minoritaire : mots spécifiques à la langue et exclusion de mots spécifiques à une langue proche [Naets, 2005] ou appartenant à des documents non pertinents [Ghani *et al.*, 2001].
– Corpus de langue générale : mots fréquents dans la langue cible, qui ne sont ni des mots outils, ni des mots spécifiques [Baroni et Sharoff, 2005].

Les mots ainsi sélectionnés peuvent être automatiquement combinés pour générer des requêtes de taille variable, éventuellement de manière aléatoire. Le choix de la taille des requêtes est crucial car il s'agit en effet de trouver le meilleur compromis entre rappel et précision. En effet, une requête courte d'un mot voire deux permettra la récupération d'un grand nombre

d'URLs, parfois non pertinentes, compte tenu de la polysémie éventuelle des mots de la requête, tandis qu'une requête longue, trop précise, risque de diminuer le nombre d'URLs retournées.

Les documents correspondant aux URLs retournées par les requêtes peuvent être de divers formats : HTML, PDF, Texte, XML, formats propriétaires (.doc, .xls ou .ppt). Pour la construction d'un corpus, il est nécessaire de normaliser ces formats afin de ne disposer que de documents au format texte simple voire XML. Les formats comme le PDF sont assez difficiles à traiter, nous avons donc uniquement traité les fichiers au format HTML. Les sections suivantes détaillent le processus de transformation des fichiers HTML en fichiers textes exploitables pour des traitements automatiques. Nous avons identifié trois problèmes principaux : détection des cadres HTML, encodage des caractères et extraction du contenu.

3.3.2 Cas particulier des cadres HTML

Certains documents HTML sont vides de contenu et pointent vers un ensemble d'autres documents HTML via le système des cadres HTML ou *frames*. Les cadres permettent de diviser l'écran en plusieurs zones et chaque zone contient une page HTML. Dans le Listing 3.1, le document HTML fait référence à trois autres pages : frame_a.html, frame_b.html et frame_c.html.

```
1 <html>
2   <frameset rows="50%,50%">
3     <frame src="frame_a.html">
4     <frameset cols="25%,75%">
5       <frame src="frame_b.html">
6       <frame src="frame_c.html">
7     </frameset>
8   </frameset>
9 </html>
```

Listing 3.1: Exemple de cadres HTML.

Pour obtenir le contenu de la page, il est donc nécessaire de récupérer également les fichiers contenus dans les divers cadres. Dans ce cas, nous avons considéré que le fichier contenant les informations d'intérêt de la page était celui qui avait la plus grande taille.

3.3.3 Recodage

Dans un document HTML, les caractères peuvent être encodés dans divers formats : ASCII, Windows-1252, ISO 8859-1, UTF-8, etc. Ces formats de codage sont indiqués soit dans les informations de protocole HTTP (En-tête HTTP), soit dans les métadonnées du fichier HTML (voir Listing 3.2, ligne 4). Lors de la constitution d'un corpus, il est souhaitable de convertir tous les documents dans le même format de codage, comme UTF-8 par exemple, afin de faciliter la réutilisation ultérieure du corpus. Les documents HTML peuvent également inclure diverses entités codant les caractères par un nom symbolique, comme « &copy ; » pour le caractère '©' et des références de caractères numériques décimales (comme '&#226 ;') ou hexadécimales (comme '&#xE0 ;') qui correspondent elles aussi à des caractères spécifiques absents du code ASCII (voir Listing 3.2, ligne 18). Ces entités doivent être converties dans le caractère correspondant.

```
1 <html>
2   <head>
3     <meta name="description" content="Extrait de Au Bonheur des
          Dames de Émile Zola">
4     <meta http-equiv="Content-Type" content="text/html; charset=
          cp1252" />
5     <meta http-equiv="pragma" content="no-cache" />
6     <title>Extrait de "Au Bonheur des Dames" de Émile Zola</title>
7     <script language="JavaScript">
8     <!--
9     window.name="Au Bonheur des Dames"
10    // -->
11    </script>
12    <style type="text/css">
13    h1 {color: red}
14    h3 {color: blue}
15    </style>
16  </head>
17  <body>
18    <p>C'&eacute;tait , &#xE0; l'encoignure de la rue de la
          Michodière et de la rue Neuve-Saint-Augustin , un magasin de
          nouveaut&eacute;s dont les &eacute;talages &eacute;clataient
          en notes vives , dans la douce et p&#226;le journ&eacute;e d'
          octobre .</p>
19  </body>
20 </html>
```

Listing 3.2: Exemple de code HTML.

3.3.4 Extraction du contenu d'un document HTML

Un document HTML contient du texte formaté par des balises comme **<p>**, qui marque un début de paragraphe (voir Listing 3.2, ligne 18). Pour extraire le contenu de l'un de ces documents il est nécessaire non seulement d'éliminer les balises HTML ainsi que les lignes de script (voir Listing 3.2, ligne 7) et de style (voir Listing 3.2, ligne 12) mais aussi les informations textuelles non pertinentes comme les éléments de navigation (menus, liens vers les pages précédentes et suivantes du site), les publicités, les informations légales (copyright), etc. qui se répètent généralement sur toutes les pages d'un même site (ce que l'on appelle *boilerplate* en anglais). La Figure 3.1 montre un tel document : le contenu pertinent, qui est encadré, ne représente qu'une faible proportion du texte, le reste correspondant à des publicités, à des informations légales ou des liens vers d'autres sections du site.

Lorsque l'on collecte un corpus de documents à partir de nombreux sites Web différents, la structure des documents est très variable : par exemple, le menu peut se trouver en-haut, en-bas, à gauche ou à droite. Il est donc difficile d'utiliser des méthodes de filtrage se basant sur une structure de document spécifique. [Finn *et al.*, 2001] proposent une méthode générique d'extraction du contenu d'un document HTML. Il s'agit d'extraire la sous-partie du document dans laquelle la densité des mots est importante et d'éliminer les sous-parties caractérisées par

FIG. 3.1: Exemple de document HTML.

une forte densité de balises. Le document HTML est tout d'abord segmenté en deux types d'unités : les mots et les balises HTML. La Figure 3.2, p. 67 représente graphiquement l'effectif cumulé des balises par rapport au nombre d'unités (mots et balises) dans le document. La zone de contenu du document, c'est-à-dire celle qui contient le plus de mots, correspond à un plateau sur la courbe : dans cette zone, l'effectif cumulé des balises progresse peu ; sur la Figure 3.2, ce plateau est délimité par deux barres verticales. L'objectif est donc de délimiter cette zone de contenu informatif.

Soit D un document contenant n unités (somme du nombre de mots et de balises). Soient $B(i,j)$ le nombre de balises se trouvant entre les positions i et j du document D et $M(i,j)$ le nombre de mots se trouvant entre les positions i et j du document D. Les positions de début d et de fin f du plateau sont telles que[1] :

$$(d,f) = \arg\max_{(d_i,f_j)} \left[B\left(0,d_i\right) + M\left(d_i,f_j\right) - B\left(d_i,f_j\right) + B\left(f_j,n\right)\right]$$

Le résultat de l'extraction de contenu du document de la Figure 3.1 se trouve Figure 3.3. On constate que les publicités, liens vers d'autres rubriques et autre informations non pertinentes ont bien été supprimées. Seule subsiste la définition de la notion d'encyclopédie, s'apparentant à une zone de contenu et correspondant au premier plateau sur la courbe. Les deux premiers mots de cette définition (*Une encyclopédie*) ont été supprimés car le mot *encyclopédie* est entouré de

[1]La fonction que nous utilisons est légèrement différente de celle proposée par [Finn *et al.*, 2001] qui ne prend pas en compte le nombre de balises se trouvant entre d et f.

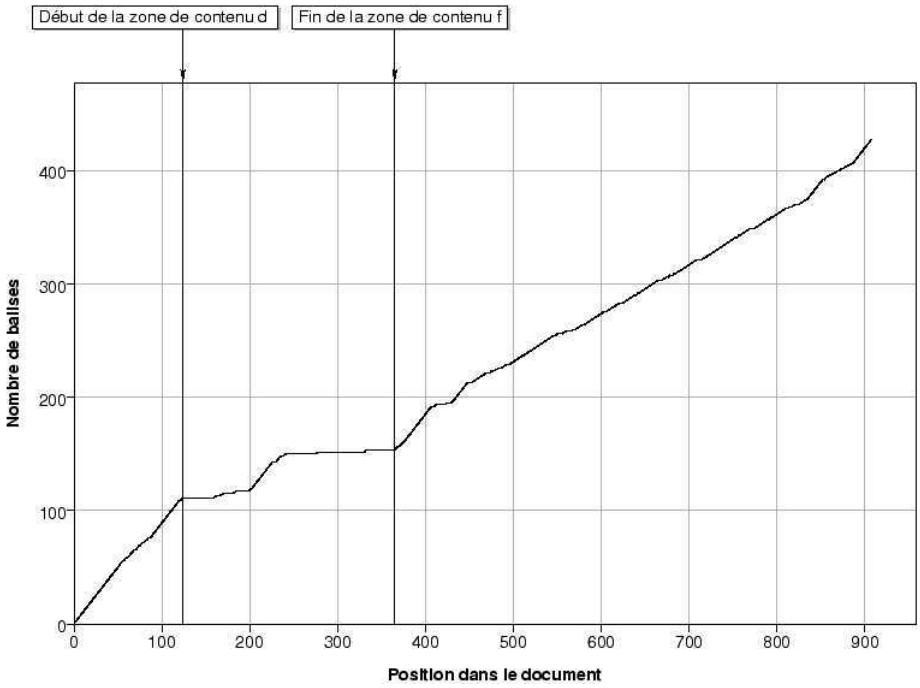


FIG. 3.2: Effectif cumulé des balises par rapport à la position pour le document de la Figure 3.1.

balises <*strong*> dans le fichier source. La position optimale du début de la zone de contenu a donc été identifiée après la balise </*strong*> fermante.

Comme l'observe [Baroni, 2005], cette méthode est bien adaptée à l'extraction de contenu pour des pages provenant de sites différents car elle ne repose pas sur des informations de structure dépendantes de sites précis. De plus, comme nous l'avons vu, elle est assez précise même s'il arrive que cette méthode supprime des éléments pertinents comme des mots ou des phrases situés en début et fin de texte ou contenus dans des tableaux HTML. Les documents comportant des zones de contenus entre-coupées de publicités ou d'éléments de navigation constituent une autre pierre d'achoppement de cette méthode qui ne permet de détecter qu'une zone centrale de contenu par fichier.

3.4 Traitement des corpus

Afin de pouvoir utiliser les textes constituant un corpus, il est nécessaire de les soumettre à un certain nombre de traitements. Nous détaillons ici les traitements minimaux (découpage en mots et vérification de la langue) que nous avons utilisés. À ces traitements peuvent s'ajouter des analyses plus sophistiquées comme l'étiquetage morphosyntaxique ou l'analyse grammaticale, que nous ne mentionnerons pas car nous ne les avons pas appliquées à nos corpus.

3.4.1 Découpage des textes en mots, phrases et paragraphes

Les textes sont représentés dans un fichier sous forme de séquences de caractères : lettres, nombres et caractères de ponctuation. Afin de pouvoir traiter ces données, il est nécessaire d'identifier différents types d'unités : paragraphes, phrases et mots. Cette tâche est loin d'être

est un ouvrage où l'on tente de traiter de façon exhaustive de l'ensemble du savoir humain. Par extension, le mot désigne également un ouvrage qui traite systématiquement d'un domaine de connaissance en particulier. Le mot encyclopédie vient du grec ancien enkuklios paideia qui désignait l'éducation comprenant l'étude de toutes les sciences. Il devint enkuklopaideia par une erreur de transcription, puis au XVIe siècle les savants le latinisèrent en encyclopaedia .
Cet article chez Wikipedia
Ejecta
Par les ejecta , les scientifiques définnisent les objets rejeté par les volcans qui ne sont pas de la lave pendant certaines éruptions explosives qui fragmentent le magma et qui retombent plus ou moins verticalement hors du cratère.
Ces ejecta sont classés suivant leur taille dans un ordre croissant, cendres, lapilli, blocs et bombes sont des ejecta. La majorité de ceux-ci retombent alors en épousant une trajectoire parabolique. Par ailleurs, les particules les plus fines peuvent atteindre une altitude suffisante pour être entraînées par les courants aériens sur des distances considérables.
Chaque éruption volcanique est caractérisée par une catégorie particulière d'ejecta.Grace à cela, les vulcanologues peuvent définir les types de volcans. Ce sont ces produits qui, par leur accumulation, forment progressivement les cônes volcaniques.

FIG. 3.3: Résultat de l'extraction de contenu pour le document de la Figure 3.1.

triviale. En effet, il n'est pas toujours aisé de définir ce qu'est un mot, voire une phrase et donc, à plus forte raison, de les identifier de manière automatique [Grefenstette et Tapanainen, 1994, Manning et Schütze, 1999]. Le mot peut être défini, de manière un peu simpliste, comme une suite de caractères séparés par des espaces ou des signes de ponctuation. Cependant, la solution triviale qui consiste à déterminer les frontières de mots par les espaces et les signes de ponctuation est loin d'être satisfaisante. Considérons les exemples suivants :

(1) L'edmontosaurus mesurait **3,5** mètres**.**

(2) Des chercheurs ont analysé **10.000** vertèbres de dinosaures**.**

(3) Cette nouvelle technique a été développée par **M.** Dupond**.**

(4) For additional information see also **http ://www.fda.gov/cder/approval/index.htm**

(5) **Pyroclastic-fall** deposits**,** referred to as tephra**,** consist of combinations of pumice**,** scoria**, dense-rock** material**,** and crystals**.**

(6) Recent geophysical data revealed that Mauna **Loa's** summit caldera has begun to swell and stretch at a rate of 2 to **2.5** inches a year**.**

Ces exemples permettent de faire les observations suivantes :

- Le **point** ne marque pas toujours une fin de phrase. Il peut être contenu dans divers types d'unités comme les nombres (voir exemples 2 et 6), les abréviations (exemple 3) ou les URLs (exemple 4). Il faut donc faire la différence entre ces occurrences et les occurrences du point comme marqueur de fin de phrase.
- La **virgule** doit dans certains cas, lorsqu'elle apparaît en fin de mot comme dans la phrase 5, être séparée de la chaîne de caractères qui la précède. Dans d'autres cas, elle peut faire partie d'un nombre comme dans l'exemple 1.
- L'**apostrophe** marque généralement une contraction comme dans *c'est* ou *l'utilisation.* Elle est également utilisée pour marquer le possessif comme dans l'exemple 6. Nous avons

choisi de toujours séparer les clitiques accolés au mot suivant ou précédent par une apostrophe.
- Le **tiret**. Nous avons conservé les tirets contenus à l'intérieur de mots (voir exemple 5), sauf dans le cas des traits d'union marquant une liaison avec un pronom comme dans *peut-on* ou *puis-je*. En effet, comme nous le verrons dans les chapitres suivants, les tirets fournissent des informations précieuses sur les frontières morphémiques et il est donc utile de les préserver.

Ces exemples montrent également qu'au cours du découpage d'un texte, il est utile de repérer certaines unités qui ont une structure spécifique comme les dates, les nombres, les abréviations, les URLs ou les adresses e-mail. Leur structure peut généralement être décrite à l'aide d'expressions régulières : certaines de ces expressions régulières sont identiques pour toutes les langues (URL, adresses e-mail), d'autres sont spécifiques aux langues comme les nombres ou les abréviations [Grefenstette, 1998]. Ces unités contiennent souvent des signes de ponctuation ambigus comme le point ou la virgule. Une fois ces unités identifiées, les signes de ponctuation restant peuvent être utilisés de manière non ambiguë pour délimiter les mots et les phrases.

Nous avons implémenté l'ensemble des opérations permettant la tokenisation d'un texte en Python (voir Annexe B.2.2, p. 155).

3.4.2 Vérification de la langue

Compte tenu du caractère multilingue du Web, il est possible de trouver des portions de texte de langues différentes au sein du même document. Sauf utilisation particulière, un corpus est généralement construit pour une seule langue cible. Il peut donc s'avérer nécessaire de filtrer les phrases ou paragraphes écrits dans une langue différente de la langue cible. On distingue deux types de méthodes pour identifier automatiquement la langue d'un texte : les méthodes basées sur les n-grammes (suites de n caractères) et les méthodes basées sur les mots fréquents [Grefenstette, 1995]. C'est cette dernière technique que nous avons utilisée. Les mots les plus fréquents d'une langue (généralement les mots outils) peuvent être identifiés en comparant leur fréquence dans un corpus d'apprentissage et leur fréquence dans le texte [Dunning, 1994, McNamee, 2005]. Cette méthode est relativement efficace, sauf bien sûr pour des passages très courts de quelques mots. Nous l'avons utilisée à l'échelle du paragraphe.

Nous avons développé un outil de reconnaissance de langue basé sur la méthode décrite dans [McNamee, 2005]. La méthode se décompose en une étape d'apprentissage et une étape d'utilisation. Dans un premier temps, des textes d'apprentissage sont collectés pour chaque langue. Les textes choisis par [McNamee, 2005] proviennent du corpus de textes du Projet Gutenberg. Nous avons également utilisés des textes du Projet Gutenberg pour notre corpus d'apprentissage (voir Tableau B.1, p. 156). Puis, les mots sont extraits de ces textes et un fichier contenant le profil de chaque langue est généré. Un profil correspond aux 1 000 mots les plus fréquents de chaque langue, associés au pourcentage d'occurrences du mot dans le texte d'apprentissage (voir Tableau B.2, p. 156). Ces profils sont ensuite utilisés pour identifier la langue d'une phrase ou d'un paragraphe. La langue d'un texte sera déterminée en calculant un score pour chaque langue afin de rechercher la langue de score maximal. Pour donner un score à une langue, il suffit d'additionner les poids des mots du texte dans cette langue.

Prenons l'exemple de la séquence « in die ». Les scores pour les différentes langues sont les suivants :

- allemand : 1,61 + 2,46 = 4,07
- anglais : 2,59 + 0 = 2,59
- français : 0 + 0 = 0
- italien : 1,09 + 0 = 1,09
- néerlandais : 2,07 + 1,23 = 3,30

La langue de la séquence « in die » est donc l'allemand. Si aucun score ne dépasse 0, alors la langue du texte est inconnue.

Les caractéristiques de l'outil ainsi que les corpus d'apprentissage utilisés pour les différentes langues figurent dans l'Annexe B.2.3, p. 155.

3.4.3 Extraction des mots d'un corpus

Les outils de tokenisation et de vérification de la langue peuvent être mis à profit pour extraire les mots d'un corpus et procéder à un premier filtrage. En effet, la tokenisation utilise des expressions régulières représentant des entités spécifiques comme les nombres, les dates, les URLs ou les adresses e-mail. Celles-ci pourront donc être exclues de la liste des mots du corpus. De plus, le module d'identification de langue permet d'éliminer les mots qui se trouvent dans des paragraphes qui ne sont pas écrits dans la langue cible du corpus. Ce second filtrage élimine une grande partie des mots étrangers. Restent les entités nommées telles que les lieux et personnes qui sont plus difficiles à identifier et que nous n'avons pas traitées (voir par exemple [Elkateb-Gara, 2005] pour un exemple de traitement des entités nommées).

3.5 Corpus collectés

Nous avons collecté 4 corpus en français et en anglais, couvrant deux domaines spécialisés distincts : la volcanologie et le cancer du sein. Dans la suite de ce rapport, ils seront désignés respectivement par « cancer-fr », « cancer-en », « volcano-fr » et « volcano-en ». Les caractéristiques des corpus ainsi que les amorces utilisées sont détaillés dans l'annexe A, p. 149.

La construction automatique de tels corpus est très rapide, dans la mesure où les différents programmes utilisent des processus légers (threads) s'exécutant en parallèle. L'étape de collecte d'URL pour 20 amorces différentes et 3 mots par requête dure moins d'une minute. Puis, l'étape de collecte et de nettoyage des fichiers HTML correspondant aux URLs dure 35 minutes pour environ 4 000 fichiers, d'une taille totale après nettoyage de 36 Mo. Ces temps de traitement ont été obtenus avec une machine comportant un processeur Intel 1,73 GHz avec 1 Go de RAM.

Les mots issus de ces corpus ont été utilisés comme données d'apprentissage pour les systèmes d'analyse morphologique non supervisée que nous avons élaborés. Dans le prochain chapitre, nous allons tout d'abord décrire le système d'analyse morphologique par segmentation.

Chapitre 4

Analyse morphologique par segmentation

4.1 Introduction

Nous allons décrire dans ce chapitre une première méthode d'analyse morphologique non supervisée. Elle prend pour entrée une liste des mots d'un corpus et produit en sortie une segmentation morphologique de ces mots. La segmentation est un découpage des mots en segments morphémiques comme par exemple : *segment + ation + s* ou *hormon + o + thérap + ie*. Ces segments ne sont pas tout à fait des morphèmes car aucune tentative n'est faite pour identifier les allomorphes. Les segments sont toutefois étiquetés par une des catégories suivantes : préfixe, radical, suffixe et segment de liaison.

Avant de décrire cette méthode, faisons un bref récapitulatif des indices utilisables pour l'acquisition automatique de connaissances morphologiques à partir d'une liste de mots :

- **Prévisibilité.** L'utilisation de la prévisibilité d'un segment en fonction de la chaîne de caractères précédente a été proposée par [Harris, 1955] et reprise par [Hafer et Weiss, 1974] et [Déjean, 1998]. De manière similaire, [Saffran *et al.*, 1996] suggèrent que les chutes de probabilités transitionnelles entre les syllabes permettent l'identification des frontières de mots. Suivant cette suggestion, nous avons utilisé les variations des probabilités transitionnelles entre les sous-chaînes d'un mot pour identifier des segments de mots.
- **Comparaison des graphies.** La comparaison orthographique des mots permet d'identifier les segments partagés et différents. [Neuvel et Fulop, 2002] procèdent par alignement des mots à partir de leur frontière gauche ou droite, tandis que [Schone et Jurafsky, 2001] insèrent les mots dans un arbre de recherche, à l'endroit ou à l'envers, pour facilement reconnaître les positions des différences entre mots. Ces méthodes, de même que les méthodes du plus long préfixe ou suffixe commun, sont adaptées à l'identification de préfixes et de suffixes mais sont inadaptées au traitement des mots composés, contenant plusieurs radicaux. Nous avons donc eu recours à des alignements multiples, résultant en diverses analyses possibles pour chaque mot.
- **Longueur et fréquence.** Selon Zipf [Zipf, 1968, page 173], plus un morphème est fréquent, plus il est court[1]. De plus, [Vergne, 2005] montre qu'il est possible de distinguer mots informatifs et mots non informatifs sur la base de leur longueur et de leur fréquence. Ainsi, si nous faisons un parallèle entre mots et morphèmes, alors les radicaux, qui ont un

[1] « the length of a morpheme tends to bear an inverse ratio to its relative frequency of occurrence »

sens plus marqué que les affixes, sont plus longs et moins fréquents, tandis que les affixes sont généralement fréquents et courts. Les informations de longueur et de fréquence sont également utilisées dans le modèle probabiliste proposé par [Creutz et Lagus, 2004] : la probabilité qu'un segment est un radical est fonction de sa longueur. Plutôt que d'utiliser des propriétés de longueur et de fréquence de manière absolue, et donc recourir à des seuils, elles peuvent être mises en œuvre dans le cadre de la théorie sausurienne selon laquelle les éléments reliés sur l'axe syntagmatique sont définis par leurs différences [Vergne, 2005]. Nous utilisons les différences de fréquence et de longueur pour (1) distinguer les radicaux des affixes et (2) imposer des contraintes sur les segments identifiés dans un mot : les affixes doivent être plus courts et plus fréquents que les radicaux.

Après ce bref rappel des indices disponibles pour la segmentation morphologique non supervisée, nous allons maintenant décrire la méthode que nous avons élaborée.

4.2 Description de la méthode

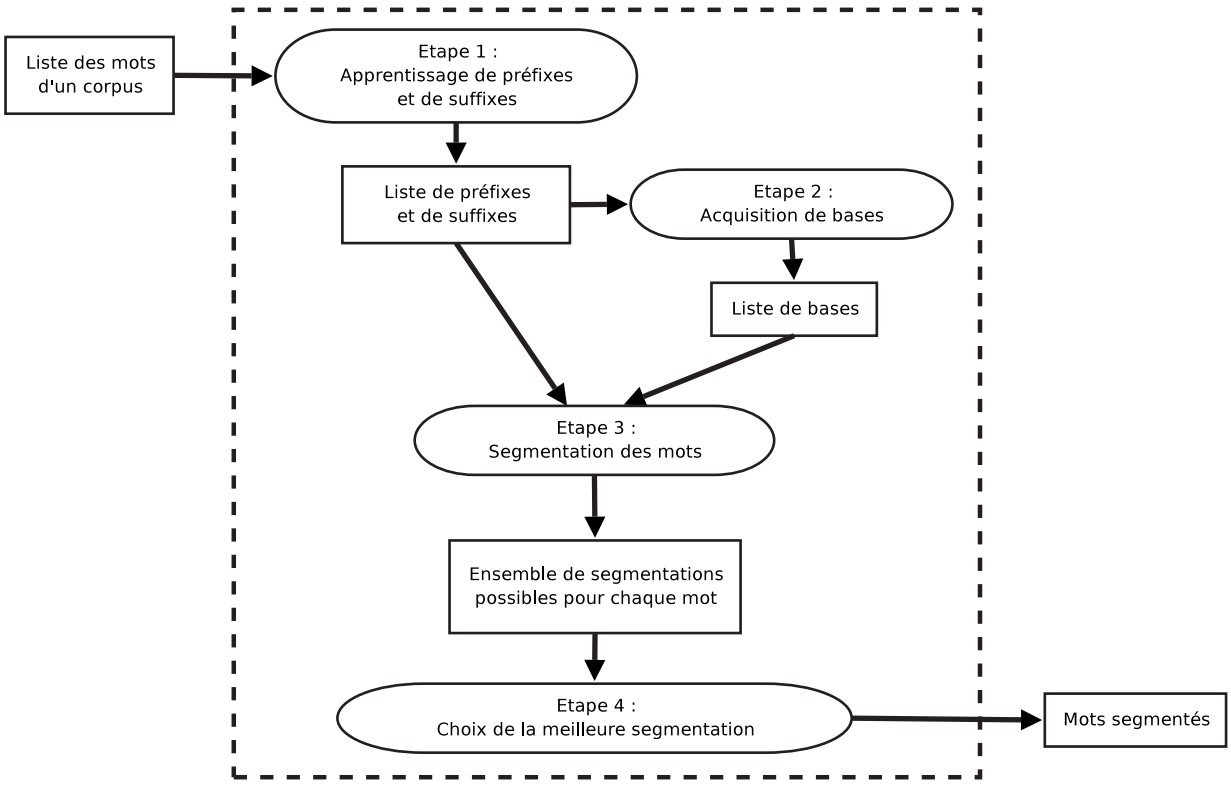


FIG. 4.1: Architecture globale du système d'analyse morphologique par segmentation

L'objectif de la méthode d'analyse morphologique par segmentation est de découper les mots en segments étiquetés. Nous ne considérons que la morphologie concaténative linéaire, sans prendre en compte les cas d'alternance dans les radicaux. Les segments découverts sont étiquetés par l'une des catégories suivantes : radical, préfixe, suffixe et segment de liaison. Cette dernière catégorie n'est généralement pas prise en compte par les méthodes d'acquisition automatique de connaissances morphologiques, mais nous pensons qu'elle est justifiée car les définitions classiques

des préfixes et suffixes n'englobent pas cette catégorie. En effet, les suffixes se trouvent après les radicaux, à la fin des mots, tandis que les préfixes se trouvent en tête de mot, avant les radicaux. Par exemple, nous considérons que "-**o**-" dans "hormon**o**thérapie" ou que le tiret dans "sous**-**estimation" sont des segments de liaison, car ils se trouvent toujours liés à deux autres segments, à gauche et à droite et ne peuvent jamais apparaître seuls en début ou en fin de mot. De plus, similairement à [Creutz et Lagus, 2004], nous utilisons la définition syntagmatique des catégories morphologiques pour contraindre l'ordre possible des segments morphémiques.

Dans les sections suivantes, nous détaillons le système de segmentation morphologique. La Figure 4.1 donne une vue globale de l'architecture du système et de ses différentes étapes. Dans un premier temps, le système extrait un ensemble de préfixes et de suffixes. Ces derniers sont ensuite utilisés pour identifier des bases, autrement dit des radicaux. Puis, les mots contenant chacune des bases ainsi identifiées sont alignés, afin de les segmenter en fonction de leurs points communs et de leurs différences. Étant donné qu'un mot peut contenir plusieurs bases différentes, il peut être segmenté plusieurs fois. Toutes ces segmentations possibles sont conservées et combinées, et la meilleure d'entre elles est sélectionnée.

Nous allons d'abord présenter chacune de ces étapes, ainsi qu'une solution possible pour la segmentation de mots absents du corpus d'apprentissage et la construction de familles morphologiques à partir des segmentations de chaque mot. Puis nous détaillons les différentes évaluations auxquelles le système a été soumis.

4.2.1 Extraction de préfixes et de suffixes

L'unique entrée du système est une liste de mots L. La méthode n'utilise pas la fréquence des mots. La première étape consiste en l'extraction d'un ensemble de préfixes P et de suffixes S. L'acquisition de ces affixes repose sur le calcul des probabilités de transition entre les sous-chaînes d'un mot. Rappelons que plus les probabilités de transitions sont faibles, plus il est difficile de prédire la co-occurrence des sous-chaînes. Afin d'augmenter la précision des affixes extraits à cette étape, seuls les mots les plus longs sont segmentés de cette manière. En effet, plus un mot est long, plus il est susceptible d'être préfixé et suffixé, ce qui réduit les risques d'erreur à l'extraction. Pour ce faire, les mots de la liste L sont tout d'abord triés par ordre de longueur décroissante puis segmentés en utilisant la variation des probabilités de transition entre sous-chaînes.

Segmentation basée sur les probabilités transitionnelles

La procédure de segmentation utilisée à cette étape repose sur l'étude des variations des probabilités transitionnelles entre sous-chaînes à toute position dans le mot. Plusieurs sous-chaînes possibles se terminent et débutent à une position donnée dans le mot. Par exemple, pour le mot *soignant*, plusieurs chaînes de caractères se terminent par g : *ig*, *oig*, etc. De même, plusieurs chaînes commencent après g : *na*, *nan*, etc. Afin de prendre en compte toutes les mesures de probabilité transitionnelle qu'il est possible de calculer à une position du mot, nous faisons la moyenne du maximum des probabilités transitionnelles à toute position dans le mot.

La fonction t décrite ci-après, détaille le calcul utilisé pour obtenir un profil des variations des probabilités transitionnelles entre les sous-chaînes d'un mot w :

Soient :
- w un mot dont les frontières sont explicitement marquées par le symbole # ;
- n la longueur de w, marqueurs de frontières inclus ;
- $s_{i,j}$ une sous-chaîne de w commençant à la position i et se terminant à la position j ;
- k une position inter-caractères dans le mot w telle que $1 \leq k \leq n-1$.

$$t(k) = \frac{\sum_{i=0}^{k-1} \sum_{j=k+1}^{n} \max[p(s_{i,k}|s_{k,j}), p(s_{k,j}|s_{i,k})]}{k \cdot (n-k)}$$

où :

$$p(s_{i,k}|s_{k,j}) = \frac{f(s_{i,j})}{f(s_{k,j})} \quad \text{et} \quad p(s_{k,j}|s_{i,k}) = \frac{f(s_{i,j})}{f(s_{i,k})}$$

La fonction $t(k)$ correspond à la moyenne du maximum des probabilités transitionnelles[1] pour toutes les sous-chaînes de w de longueur supérieure ou égale à 1 se terminant et commençant à la position k. Les probabilités transitionnelles étant asymétriques, nous calculons à la fois $p(s_{i,k}|s_{k,j})$ et $p(s_{k,j}|s_{i,k})$ et ne conservons que le maximum des deux valeurs[2]. La fréquence f d'une sous-chaîne $s_{i,j}$, notée $f(s_{i,j})$ correspond à son nombre d'occurrences dans L.

Les valeurs de cette fonction permettent d'obtenir un profil général des variations des probabilités transitionnelles pour le mot w. Les minimums locaux indiquent des frontières morphémiques potentielles. En effet, les chutes de la valeur des probabilités transitionnelles indiquent qu'il est difficile de prévoir le segment suivant en fonction du segment précédent. Cette remarque vaut également en sens inverse, en partant de la fin du mot : il est difficile de prévoir le segment précédent en fonction du segment suivant.

Un minimum local est validé si sa différence avec le maximum précédent et le maximum suivant est au moins égale à un écart-type des valeurs prises par la fonction t pour w. La figure 4.2 représente ces profils pour divers mots en anglais et en français. Les frontières morphémiques valides sont indiquées par une ligne verticale en gras, ce qui correspond aux segmentations suivantes : *myo + fibro + blastique + s*, *paléo + climat + olog + ue*, *post + transplant + ation* et *dis + integrat + ing*. Certains minimums ne sont pas validés, du fait de l'utilisation d'un seuil.

Nous rappelons que l'objectif de cette première étape n'est pas la segmentation de l'ensemble des mots. En effet, la méthode basée sur la variation des probabilités transitionnelles n'est pas suffisamment efficace, surtout pour les mots les plus courts, comme le montrent les exemples de la Figure 4.3. Selon cette méthode, la segmentation du mot *soignant* serait *so + ignant* et celle de *adultes ad + ultes*.

[1] Les probabilités transitionnelles sont équivalentes aux probabilités conditionnelles.

[2] Voir [Véronis, 2003] pour une utilisation des probabilités conditionnelles pour estimer la force d'association entre mots dans un contexte donné.

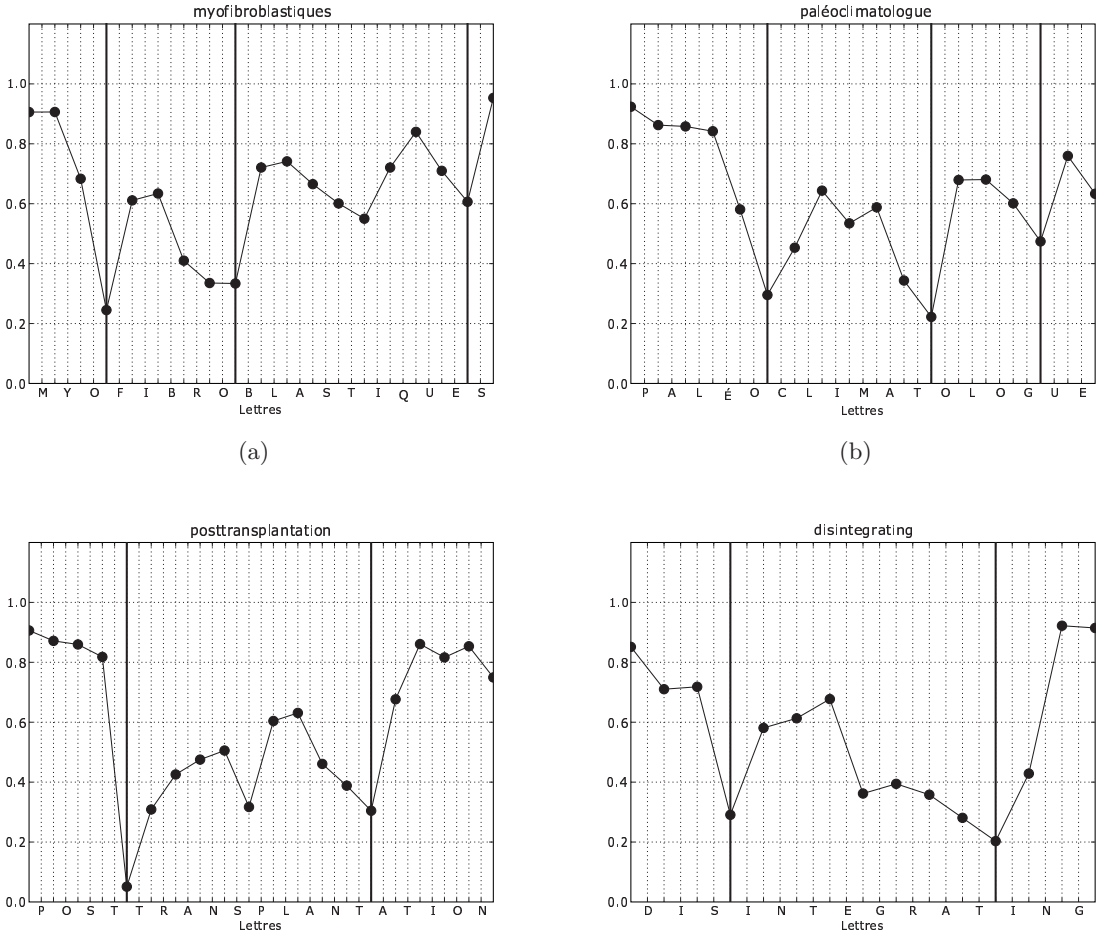


FIG. 4.2: Exemples de profils de variation des probabilités transitionnelles pour des mots issus des corpus suivants : (a) cancer-fr ; (b) cancer-en ; (c) volcano-fr ; (d) volcano-en.

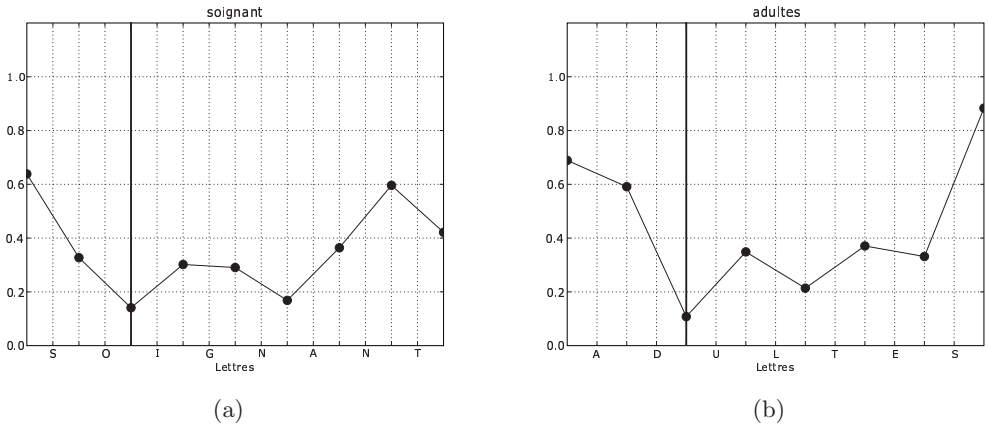


FIG. 4.3: Exemples de segmentations erronées effectuées par la méthode basée sur la variation des probabilités transitionnelles.

Identification d'une base parmi les segments et acquisition de préfixes et de suffixes

Lorsqu'un mot a été segmenté de cette manière, une base[1] est identifiée parmi les segments découverts. Dans la mesure où la segmentation met à jour des segments de tout type (préfixe, suffixe, base), il n'est pas possible d'utiliser uniquement des critères positionnels pour étiqueter ces segments. Nous considérons que la base est le segment le plus long et le moins fréquent, avec les contraintes supplémentaires suivantes :

– **contrainte de fréquence** : le segment doit apparaître au moins deux fois dans L.

– **contrainte de position** : le segment doit apparaître au moins une fois en début de mot.

Les Tableaux 4.1 donnent divers exemples de mise en œuvre de cette méthode. Ainsi, parmi les segments identifiés dans le mot *paléoclimatologue*, *climat* sera considéré comme étant le radical car il s'agit du segment le plus long (6 lettres) et le moins fréquent (17 occurrences). Il faut noter que la base identifiée n'est pas forcément minimale : par exemple, la base *blastique* contient le suffixe *+ique* et n'est donc pas totalement décomposée à cette étape.

Segments	myo	fibro	blastique	s
Fréquence	34	74	17	11 964
Longueur	3	5	9	1

(a)

paléo	climat	olog	ue
68	17	288	1 348
5	6	4	2

(b)

Segments	post	transplant	ation
Fréquence	278	42	1 163
Longueur	4	10	5

(c)

dis	integrat	ing
295	13	2 793
3	8	3

(d)

TAB. 4.1: Exemple d'identification du radical parmi les segments obtenus pour les mots de la Figure 4.2.

Afin d'augmenter le nombre d'affixes extraits, nous recherchons la base identifiée dans l'ensemble des mots de L et ajoutons les sous-chaînes qui précèdent et suivent immédiatement la base dans ces mots aux listes P et S si elles sont plus courtes et plus fréquentes que la base. De plus, nous éliminons les préfixes de longueur 1 car nous avons remarqué qu'ils conduisent à des segmentations erronées dans les étapes ultérieures de la segmentation. Les préfixes et suffixes acquis à partir de la base *climat* sont listés dans la Table 4.2. Les préfixes extraits sont *paléo+*, *ac+* et *réa+*. Le préfixe *dendro+* n'est pas validé car il est moins fréquent que *climat* : il n'y a que 4 occurrences de *dendro* en début de mot dans L tandis que *climat* compte 17 occurrences. Les suffixes identifiés sont *+s*, *+isés*, *+ologue*, *+ation*, *+ologie*, *+iques*, *+er*, *+ique*, *+é* et *+isé*. Les suffixes *+isation* et *+ologues* ne sont pas validés car il sont plus longs que *climat*.

[1]Dans la suite du chapitre, nous considérons les termes « base »et « radical »comme synonymes.

préfixe	**base**	**suffixe**
paléo	climat	s
	climat	~~isation~~
	climat	isés
	climat	s
paléo	climat	ologue
	climat	
ac	climat	ation
~~dendro~~	climat	ologie
	climat	~~ologues~~
	climat	ologie
	climat	iques
réa	climat	er
	climat	ique
ac	climat	é
	climat	isé
paléo	climat	iques

TAB. 4.2: Préfixes et suffixes acquis à partir de la base *climat*. Les affixes non validés sont barrés.

Cette méthode est appliquée uniquement aux mots les plus longs du corpus. Il reste donc à déterminer le nombre de mots longs nécessaires pour cet apprentissage. Dans la première version du système, ce nombre était déterminé par un paramètre fixé avant l'apprentissage, compris entre 100 et 500 mots [Bernhard, 2005]. Nous avons par la suite utilisé un paramètre différent, qui mesure la stabilité de la liste des affixes extraits. En effet, le nombre de nouveaux affixes extraits diminue lorsque le nombre de mots segmentés augmente. Nous mettons donc fin à la procédure lorsque pour N mots successifs au moins la moitié des affixes extraits sont connus, c'est-à-dire appartiennent aux listes P et S. Les Tableaux 4.4 et 4.6 listent les préfixes et suffixes les plus fréquents pour chacun de nos corpus avec N=5. Le tableau 4.7 donne le nombre de préfixes et de suffixes extraits pour chaque corpus pour N=5 et N=10.

cancer-fr			
Préfixes		**Suffixes**	
extra-	14	s	102
hyper	13	e	50
intra	12	es	26
immuno-	11	ation	16
radio	11	ie	16
radio-	11	ique	16
immuno	9	ale	15
intra-	9	que	14
neuro	9	iques	12
ex	8	ée	11

volcano-fr			
Préfixes		**Suffixes**	
volcano-	8	s	41
thermo	7	e	12
re	6	es	12
inter	5	ées	7
micro	5	és	6
dis	3	ment	6
non-	3	ne	6
paléo	3	ée	5
poly	3	ique	5
pyro	3	me	5

TAB. 4.4: Préfixes et suffixes les plus fréquents dans les corpus français pour N=5.

cancer-en			
Préfixes		**Suffixes**	
non-	14	s	35
immuno	13	al	18
ultra	8	e	17
hyper	6	y	12
non	6	-related	11
pre	6	-based	10
re	6	-induced	8
re-	6	-containing	8
anti	5	ic	6
post	5	ly	6

volcano-en			
Préfixes		**Suffixes**	
micro	10	s	76
dis	9	al	22
inter	6	ly	18
non-	6	ed	17
co	5	ion	14
hyper	5	ic	13
multi-	5	ing	12
re	5	d	9
cross-	4	es	7
magma-	4	ally	6

TAB. 4.6: Préfixes et suffixes les plus fréquents dans les corpus anglais pour N=5.

	Nombre de préfixes		Nombre de suffixes	
	N = 5	N = 10	N = 5	N = 10
cancer-fr	124	150	157	211
cancer-en	96	247	186	367
volcano-fr	36	83	84	144
volcano-en	118	311	207	549

TAB. 4.7: Nombres de préfixes et de suffixes acquis à partir des mots de chaque corpus.

4.2.2 Acquisition de bases

Les bases sont obtenues en retranchant des mots de L toutes les combinaisons possibles des affixes acquis précédemment et de la chaîne vide. Bien sûr, la liste de bases obtenues de cette manière contient beaucoup de bases non plausibles. Nous appliquons donc les contraintes suivantes à chaque base extraite b :

- **contrainte de longueur** : elle doit avoir une longueur minimale de 3.
- **contrainte de non inclusion** : elle doit pouvoir être suivie d'au moins deux lettres différentes (y compris la marque de fin de mot) ; dans le cas contraire, cela voudrait dire que la base est incluse dans une autre base.
- **contrainte typographique** : elle ne peut pas contenir de tiret, car les tirets marquent une frontière morphémique.
- **contrainte de position** : au moins un mot de L doit commencer par b.

4.2.3 Segmentation des mots

Les mots sont segmentés par comparaison des graphies des mots contenant la même base afin de détecter les frontières entre segments partagés et segments différents. Lors de cette étape, toutes les bases acquises précédemment sont examinées l'une après l'autre et les mots dans lesquels elles figurent sont comparés, quelle que soit la position de la base dans ces mots. Il est

ainsi possible d'identifier plusieurs segments dans chaque mot qui peuvent être aussi bien des préfixes que des suffixes.

Les Figures 4.4 présentent les segmentations proposées pour les mots contenant la base *océan*, à diverses étapes de la segmentation. Avant segmentation, chaque mot comprend au plus 3 segments : un préfixe, la base (*océan*) et un suffixe (voir Figure 4.4a).

Les critères suivants sont utilisés pour effectuer la segmentation :

1. Inclusion d'un tiret : les tirets sont considérés comme marquant des frontières de segments. Ainsi, comme le montre la Figure 4.4b, le segment *médio-* est segmenté en *médio* + -.
2. Inclusion d'un autre affixe lié à la même base. Si un préfixe p_1 se termine par un autre préfixe p_2 également lié à b, alors il est découpé en deux segments. De même, si un suffixe s_1 débute par un autre suffixe s_2 également lié à b, alors il est découpé en deux segments. C'est le cas par exemple pour le suffixe *iques* qui est découpé en *ique* + *s* (voir Figure 4.4c). Cette segmentation est récursive. En effet, le segment *iennes* est découpé en *ienne* + *s*, puis *ienne* est découpé en *ien* + *ne* et enfin *ien* est découpé en *ie* + *n*.
3. Inclusion d'un autre affixe des listes P ou S. Par exemple si un préfixe p_1 débute par un préfixe p_2 appartenant à la liste P, alors il peut être découpé en deux segments. De la même manière, si un suffixe s_1 se termine par un suffixe s_2 appartenant à la liste S, alors il peut être découpé en deux segments. C'est pour cette raison que le suffixe *ographie* est segmenté en *ograph* + *ie* car le suffixe *ie* appartient à la liste S (voir Figure 4.4d).

Les segments ainsi obtenus sont étiquetés par l'une des trois catégories d'affixes suivantes : préfixe, suffixe et segment de liaison, en fonction de leur position dans le mot par rapport à la base. Ainsi, les tirets, compte tenu de leur distribution (ils apparaissent toujours entre deux autres segments), sont systématiquement étiquetés comme des segments de liaison. Les segments qui peuvent apparaître en toute fin de mots prennent l'étiquette suffixe. C'est le cas par exemple des segments *s* ou *ique*. À l'inverse, les segments qui peuvent apparaître en début de mot sont étiquetés comme préfixes. C'est le cas du segment *sub*. De plus, afin de prendre en compte les mots composés, nous donnons une étiquette temporaire aux segments qui contiennent une autre base. Ces segments ont pour étiquette « base potentielle ». C'est le cas par exemple du segment *ograph*.

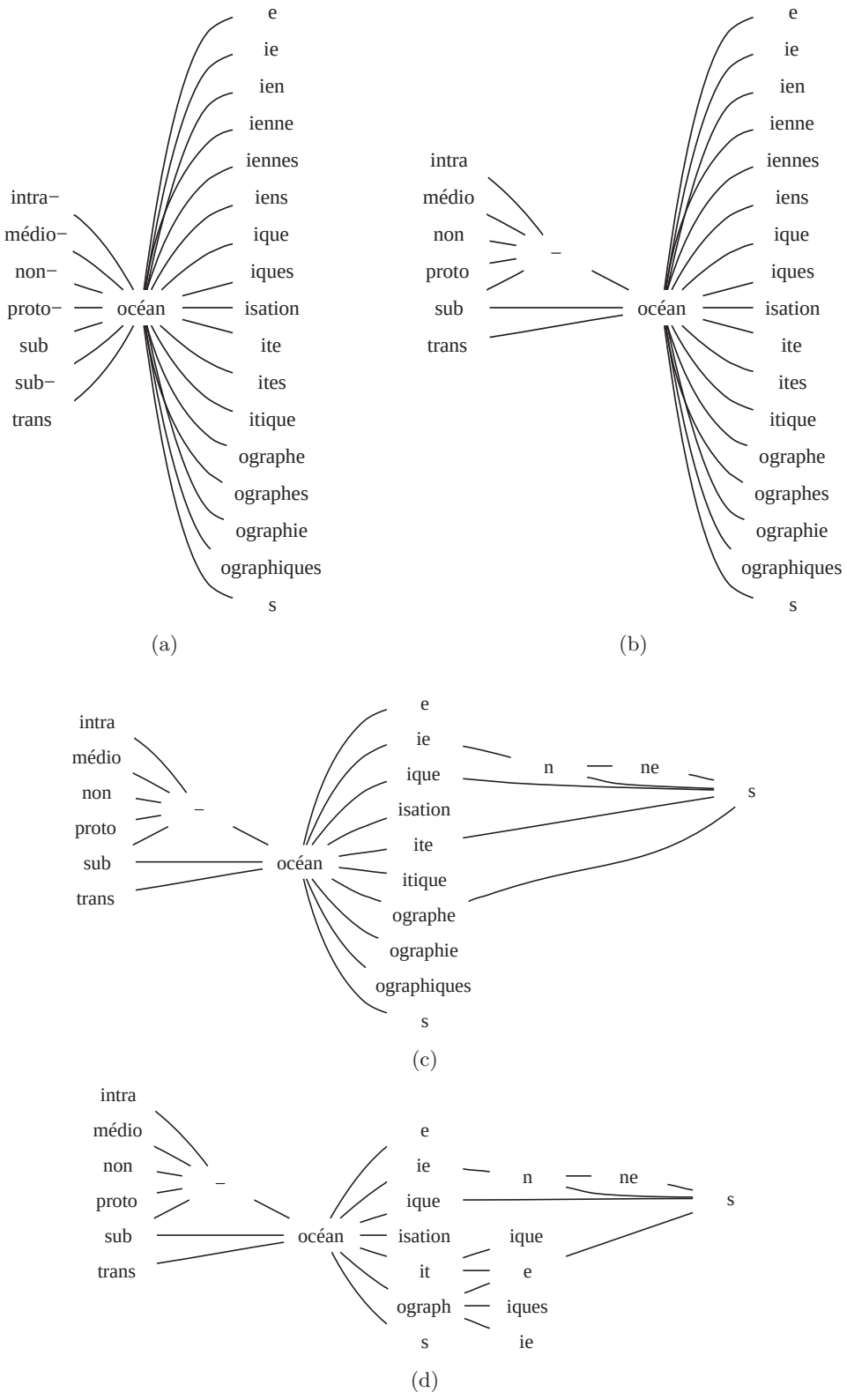


FIG. 4.4: Étapes de la segmentation des mots contenant la même base : (a) avant segmentation ; (b) après segmentation basée sur les tirets ; (c) après segmentation par comparaison aux autres affixes liés à la base ; (d) après segmentation par recherche d'affixes des listes P ou S.

A la suite de la comparaison des mots contenant la même base, de nouveaux affixes, qui n'appartiennent pas aux listes P et S peuvent être découverts et ces derniers doivent être validés, afin d'éviter une baisse trop importante de la précision des segmentations. La procédure de validation que nous appliquons est similaire à la méthode de validation des nouveaux morphèmes proposée par [Déjean, 1998] et consiste à prendre en compte la proportion d'affixes déjà connus, c'est-à-dire appartenant aux listes P et S, par rapport aux nouveaux affixes.

La méthode de H. Déjean permet la découverte de nouveaux morphèmes en vérifiant si les suffixes découverts correspondent à des morphèmes déjà identifiés auparavant. Si la moitié de ces éléments appartiennent à la liste des morphèmes déjà découverts, alors les autres morphèmes sont ajoutés à la liste. Prenons l'exemple du Tableau 4.8. Dans cet exemple, la chaîne de caractères *consider* peut être suivie de 5 séquences différentes : *+able*, *+ably*, *+ation*, *+ed* et *+ing*. Trois de ces séquences appartiennent à la liste des morphèmes déjà identifiés, ce qui correspond à plus de la moitié. Les segments *+able* et *+ably* sont donc validés.

Morphèmes trouvés	Mots	Nouveaux morphèmes
	considerable	+able
	considerably	+ably
+ation	consideration	
+ed	considered	
+ing	considering	

Tab. 4.8: Exemple de validation des nouveaux morphèmes tiré de [Déjean, 1998, p. 70].

Dans la mesure où les alignements que notre système produit par comparaison des mots permettent aussi bien de découvrir de nouveaux préfixes que de nouveaux suffixes, nous avons dû adapter cette méthode de validation. Nous validons les suffixes en fonction des préfixes avec lesquels ils apparaissent, tandis que les préfixes sont validés en fonction des suffixes avec lesquels ils apparaissent.

Prenons l'exemple de la Table 4.9, qui présente l'ensemble des mots non préfixés contenant la base *hous*.

Mots	Suffixes de la liste S	Bases potentielles	Suffixes inconnus
housekeeping		+ekeeping	
housing	+ing		
household		+ehold	
house's			+e's
house	+e		
housed	+ed		

Tab. 4.9: Validation des suffixes pour les mots contenant la base *hous* et commençant par la chaîne vide.

Soit $|A_1|$ le nombre de suffixes appartenant à la liste S, $|A_2|$ le nombre de bases potentielles et $|A_3|$ le nombre de suffixes inconnus. Pour les exemples de la Table 4.9, $|A_1|$=3, $|A_2|$= 2 et $|A_3|$=1. Les suffixes inconnus, ainsi que les bases potentielles ne sont validés que si les conditions suivantes sont remplies :

$$\frac{|A_1| + |A_2|}{|A_1| + |A_2| + |A_3|} \geq a \quad \text{et} \quad \frac{|A_1|}{|A_1| + |A_2|} \geq b$$

La première inégalité permet de mesurer la proportion d'affixes déjà connus et de bases potentielles par rapport à tous les affixes. La seconde inégalité complète la première et permet d'éviter la validation de suffixes inconnus si le nombre de bases potentielles est très important (ce qui peut arriver si la base correspond à un préfixe de la langue par exemple).

a et b sont des paramètres fixés manuellement. D'après nos expériences, les valeurs par défaut suivantes permettent généralement d'obtenir de bons résultats : $a \geq 0,8$ et $b = 0,1$.

Pour les exemples de la Table 4.9, et pour les valeurs suivantes : $a = 0,8$ et $b = 0,1$, les bases potentielles *ekeeping* et *ehold* ainsi que le nouveau suffixe *e's* sont validés ($\frac{3+2}{3+2+1} > 0,8$ et $\frac{3}{3+2} > 0,1$).

La validation des préfixes se fait de manière totalement similaire. Considérons les exemples de la Table 4.10. Il s'agit de l'ensemble de mots contenant la base *hous* et se terminant par le suffixe *+e*. Dans ce cas, $|A_1|$=2 (la chaîne vide est toujours considérée comme un préfixe valide), $|A_2|$= 4 et $|A_3|$=0. Les bases potentielles *glass*, *green*, *light* et *ware* sont validées dans le cas où $a = 0,8$ et $b = 0,1$ car $\frac{2+4}{2+4+0} > 0,8$ et $\frac{2}{2+4} > 0,1$

Mots	Préfixes de la liste P	Bases potentielles	Préfixes inconnus
glasshouse		glass+	
greenhouse		green+	
lighthouse		light+	
rehouse	re+		
warehouse		ware+	
house	ϵ+		

TAB. 4.10: Validation des préfixes pour les mots contenant la base "hous" et se terminant par le suffixe 'e'.

Nous validons de cette manière les préfixes apparaissant avec tous les suffixes possibles. De la même manière, nous faisons une itération sur la liste des préfixes possibles, y compris la chaîne vide, pour valider les suffixes.

Les segmentations valides de chaque mot sont stockées. Nous conservons ainsi tous les segments proposés car un mot peut contenir plusieurs bases différentes et donc être aligné et segmenté plus d'une fois. Quand toutes les bases ont été analysées, nous examinons les segments stockés pour chaque mot et supprimons les bases potentielles. Cette étape a pour objectif de vérifier que la base contenue dans un segment identifié comme base potentielle a bien été validée lors de l'alignement du mot en fonction de cette base. Les bases potentielles sont donc soit remplacées par d'autres segments, découverts au cours du processus (en entier ou seulement en partie) ou alors étiquetées en fonction de leur position dans le mot par une des catégories d'affixes (préfixe, suffixe ou segment de liaison) si aucun remplacement n'est possible. On rencontre ce dernier cas lorsque la segmentation effectuée à partir de la base potentielle n'est pas vali-

dée. Enfin, nous calculons la fréquence d'occurrence de chaque segment étiqueté. La fréquence d'occurrence correspond au nombre de mots différents dont l'analyse inclut le segment.

4.2.4 Sélection de la meilleure segmentation

Pour chaque mot, nous avons conservé l'ensemble des segments étiquetés résultant de ses segmentations successives (une segmentation par base). Ceci va nous permettre de sélectionner la meilleure des segmentations possibles, en fonction des fréquences observées pour les différents segments découverts. La meilleure segmentation d'un mot sera celle qui utilise des segments les plus fréquents.

Afin de choisir la meilleure segmentation, nous appliquons une stratégie de recherche partant du début du mot et privilégiant le segment le plus fréquent, en cas de choix (best-first search). La segmentation finale doit également respecter des contraintes sur la succession possible des segments morphémiques :

- au moins une base dans la segmentation finale.
- un préfixe ne peut pas être directement suivi par un suffixe, mais un suffixe peut être suivi par un préfixe (comme par exemple dans le mot allemand *Terminologieübersetzung* où le suffixe *+ie* est directement suivi par le préfixe *über+*).
- le nombre de préfixes successifs, éventuellement séparés par un segment de liaison, est limité à 3.
- le nombre de suffixes successifs, éventuellement séparés par un segment de liaison, est limité à 3.

Ces contraintes sont implémentées sous forme d'un automate, permettant de vérifier au fur et à mesure de la recherche que la succession de segments est bien valide. La fréquence des segments est également prise en compte, dans la mesure où les bases sélectionnées doivent être moins fréquentes que les autres segments.

Considérons les exemples de la Figure 4.5. Plusieurs analyses sont possibles pour le mot *déformée* : *déformée, déformé + e, déform + é + e et dé + form + é + e*. La dernière de ces analyses est sélectionnée. En effet, le segment *dé* est plus fréquent que les autres segments potentiels situés en début de mot. Il sera donc sélectionné. La suite de l'analyse est validée car la base *form* est moins fréquente que les autres segments.

Le mot *intra-océanique* peut également être segmenté de diverses manières. Le préfixe *intra+* et le tiret de liaison sont forcément sélectionnés mais la suite de l'analyse est soumise à un choix. Là encore, la base la plus fréquente, à savoir *océan* est sélectionnée, ce qui conduit à l'analyse suivante : *intra + – + océan + ique*.

À la fin de cette étape, chaque mot de la liste L est segmenté et tous les segments morphémiques qu'il contient sont étiquetés. De plus, nous obtenons également une liste de tous les segments sélectionnés, étiquetés par leur catégorie ainsi que le nombre de fois où ils ont été sélectionnées (ceci correspond à la fréquence du segment). Cette liste de segments peut être utilisée pour segmenter tout mot dans la même langue, selon une procédure détaillée dans la prochaine section.

La Table 4.11 présente quelques exemples de segmentations obtenues pour le corpus cancer-en, avec N=5, a=0,9 et b=0,1.

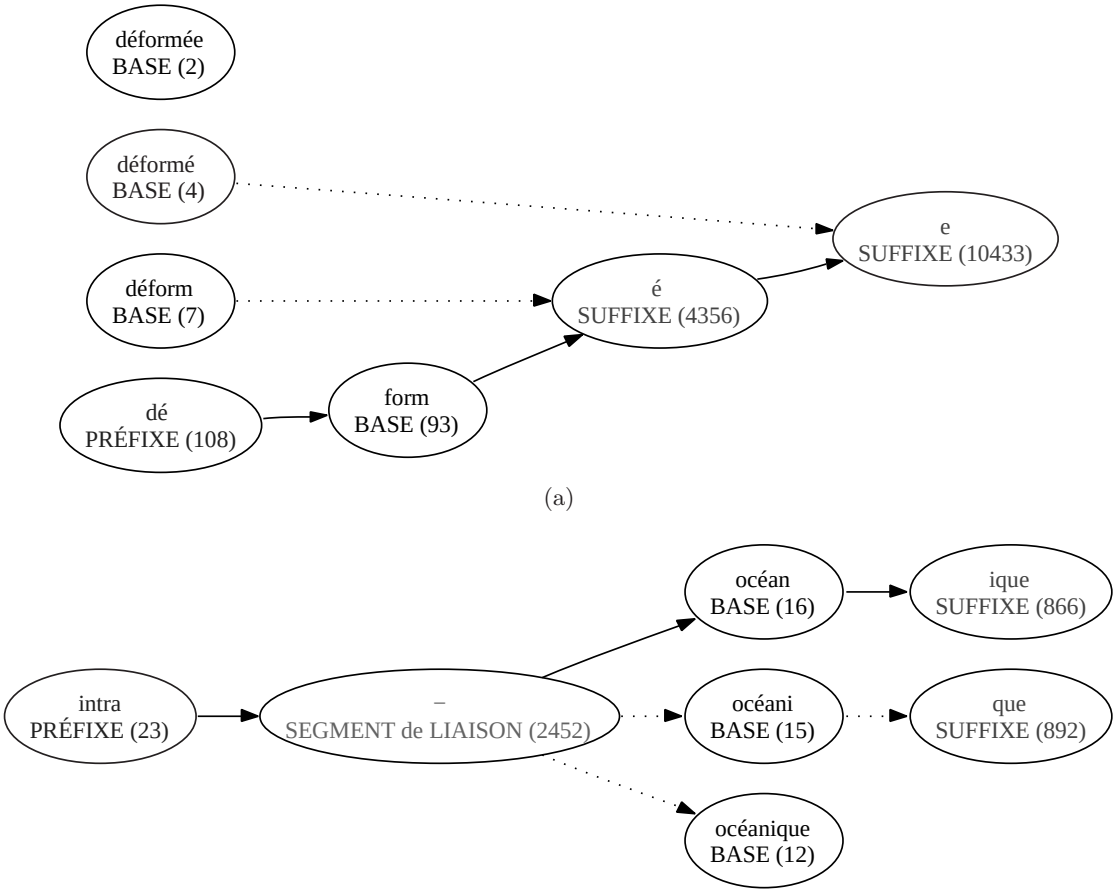


FIG. 4.5: Choix de la meilleure segmentation. Les choix possibles sont figurés par des lignes en pointillés, les choix validés sont représentés par une ligne continue. Chaque nœud du graphe correspond à un segment. Le type du segment est indiqué en lettres majuscules et sa fréquence est indiquée entre parenthèses. Les résultats sont donnés pour deux mots issus du corpus volcano-fr : (a) déformée ; (b) intra-océanique.

Segmentations	Morfessor	Segmentations de référence
$accomplish_b + es_s$	$accomplish_b + es_s$	$accomplish_b + es_s$
$accomplish_b + ment_s$	$accomplish_b + ment_b$	$accomplish_b + ment_s$
$beautiful_b + ly_s$	$beautiful_b + ly_s$	$beauti_b + ful_s + ly_s$
$beautiful_b + ly_s$	$beautiful_b + ly_s$	$beauti_b + ful_s + ly_s$
$con_p + figur_b + ation_s$	$configur_b + ation_b$	$con_p + figur_b + ation_s$
$insur_b + e_s$	$insure_b$	$in_p + sure_b$
$insur_b + e_s + d_s$	$insur_b + ed_s$	$in_p + sur_b + ed_s$
$insur_b + ing_s$	$insur_b + ing_s$	$in_p + sur_b + ing_s$
$micro_p + organism_b + s_s$	$micro_b + organism_b + s_s$	$micro_p + organ_b + ism_s + s_s$
$photograph_b + e_s + r_s + s_s$	$photo_b + graph_b + er_s + s_s$	$photo_p + graph_b + er_s + s_s$
$re_p + sid_b + e_s + d_s$	$resided_b$	$resid_b + ed_s$
$re_p + sid_b + e_s + s_s$	$reside_b + s_s$	$reside_b + s_s$
$re_p + sid_b + ing_s$	$re_p + siding_b$	$resid_b + ing_s$
$un_p + expected_b + ly_s$	$un_p + expect_b + ed_s + ly_s$	–

TAB. 4.11: Exemples de segmentations obtenues pour le corpus cancer-en, avec N=5, a=0,9 et b=0,1. Les catégories des différents segments sont marquées par un indice : b pour les bases, p pour les préfixes, s pour les suffixes et l pour les segments de liaison. Les résultats sont comparés avec ceux du système Morfessor-Categories-MAP et les segmentations de références du système Hutmegs basé sur CELEX [Creutz, 2006, p. 74].

4.2.5 Réutilisation de la liste des segments obtenus

Les corpus utilisés pour l'apprentissage ne contiennent pas tous les mots de la langue. Il est donc souhaitable de pouvoir réutiliser les segments appris pour segmenter des mots absents de la liste utilisée pour l'apprentissage.

Cette étape est bien sûr optionnelle et est proposée comme solution pour la segmentation des mots qui n'appartiennent pas à la liste L utilisée pour l'apprentissage. La meilleure segmentation pour chaque mot est identifiée via l'algorithme A*. Cette méthode consiste à sélectionner la segmentation dont le coût global est minimal. Le coût global pour une segmentation est la somme des coûts associés à chaque segment s_i. Nous avons expérimenté deux mesures de coût différentes, basées sur la fréquence des segments $f(s_i)$:

$$coût_1(s_i) = -log\frac{f(s_i)}{\sum_i f(s_i)}$$

$$coût_2(s_i) = -log\frac{f(s_i)}{\max_i[f(s_i)]}$$

Les fonctions de coûts sont complétées par les contraintes morphotactiques exposées dans la section précédente.

4.2.6 Obtention de familles morphologiques

Les résultats de la segmentation peuvent aussi être utilisés pour identifier les familles morphologiques. L'objectif premier du système est de découper les mots en segments et d'étiqueter les segments ainsi produits. Grâce à cet étiquetage, il est possible de grouper les mots en fonction des segments morphémiques partagés, et notamment les bases communes. Nous considérons

que les mots qui partagent la même base après segmentation appartiennent à la même *famille morphologique*.

Nous listons ci-dessous quelques exemples de familles obtenues pour le corpus volcano-fr, avec N=5, a=0.9 et b=0.1, la base commune est indiquée en gras :

océan	océane ; transocéanique ; sub-océaniques ; océan ; intra-océaniques ; océaniques ; océans ; océanique ; subocéanique ; intra-océanique ; proto-océanique ; océanie ; continent-océan ; non-océaniques
basalt	basaltique ; basalt ; paléobasaltes ; sous-basaltique ; basalte ; basaltes ; basaltes-hôtes ; basaltiques ; basaltique-gabbroïque ; sous-basaltiques ; basaltique-andésitique ; basalts
hydrat	hydraté ; hydrate ; hydrates ; déshydratés ; déshydratées ; déshydrate ; déshydratation ; hydratation ; hydratées ; hydratés ; hydratée
abouti	aboutissant ; abouti ; aboutir ; aboutissait ; aboutit ; aboutissez ; aboutissement ; aboutirent ; aboutissent

4.3 Évaluations

Le système a été soumis a plusieurs évaluations. La première a été effectuée lors du challenge de segmentation morphologique non supervisée organisé en 2005–2006 dans le cadre du réseau d'excellence européen PASCAL. Ce challenge a permis d'évaluer non seulement les segmentations produites mais aussi de comparer les résultats à ceux d'autres systèmes et d'estimer leur utilité dans une application réelle, à savoir la reconnaissance de la parole. Le système a également été évalué dans le cadre d'une application de synthèse de la parole. Enfin, la dernière expérimentation évalue la qualité des familles morphologiques obtenues.

4.3.1 Évaluation dans le cadre de Morpho Challenge 2005

Le challenge de segmentation non supervisée de mots en morphèmes, ou Morpho Challenge, a été organisé par Mikko Kurimo, Mathias Creutz et Krista Lagus de l'Université de Helsinki (Neural Networks Research Centre) dans le cadre du réseau d'excellence européen PASCAL (Pattern Analysis, Statistical Modelling and Computational Learning)[1]. Le challenge avait pour objectif l'élaboration d'un algorithme d'apprentissage capable de segmenter les mots en morphèmes. Les résultats de ce challenge, ouvert à tous, ont été présentés lors d'un atelier à Venise le 12 avril 2006, conjointement à d'autres challenges du réseau PASCAL.

Nous allons dans ce qui suit présenter les objectifs de ce challenge, ainsi que les compétitions proposées et les résultats obtenus par notre système [Bernhard, 2006c]. Cette présentation constitue un résumé de l'analyse plus détaillée de [Kurimo *et al.*, 2006].

L'intérêt pour la segmentation morphologique exprimé par ce challenge est lié à l'utilisation possible des segments morphémiques pour diverses tâches telles que la traduction automatique, la recherche d'information, la reconnaissance de la parole ou la modélisation statistique des langues [Kurimo *et al.*, 2006]. Les organisateurs ont fixé les objectifs suivants pour le challenge[2] :

– Apprendre quels phénomènes sont sous-jacents à la construction des mots dans les langues naturelles.

[1] Le site web du challenge, `http://www.cis.hut.fi/morphochallenge2005/`, fournit de nombreuses informations, dont les jeux de données, les résultats et les articles décrivant les divers algorithmes.

[2] Notre traduction.

– Découvrir des approches utilisables pour un grand nombre de langues.
– Faire avancer les méthodes d'apprentissage automatique.

Deux compétitions différentes ont été proposées dans le cadre du challenge, afin de comparer les différents systèmes :

– **Compétition 1** : comparaison des segmentations proposées à des segmentations de référence basées sur les résultats de FINTWOL pour le finnois, CELEX pour l'anglais et un analyseur morphologique développé à l'Université Bogaziçi pour le turc.
– **Compétition 2** : utilisation des segmentations pour découper les mots dans des corpus en finnois et en turc afin d'entraîner un modèle de langage n-gramme utilisé pour des expériences de reconnaissance de la parole.

Au total, le challenge a réuni 12 compétitrices et compétiteurs de 6 pays différents en Europe et en Amérique du Nord, totalisant 14 méthodes différentes, dont 10 ont été appliquées aux trois langues proposées. Près de la moitié des algorithmes ont été conçus par des étudiants de l'université de Leeds au Royaume-Uni, dans le cadre d'un projet en traitement automatique des langues.

Nous allons dans la suite décrire plus précisément les données fournies, les méthodes d'évaluation utilisées pour les deux compétitions ainsi que les résultats obtenus par notre système en comparaison avec les autres systèmes ayant pris part à la compétition, y compris les différentes versions du programme Morfessor développé par les organisateurs.

Données

Les données fournies par les organisateurs consistaient en des listes de mots associées à leur fréquence, dans trois langues différentes : l'anglais (167 377 mots différents), le finnois (1 636 336 mots différents) et le turc (582 923 mots différents). Nous donnons ci-dessous un extrait de chaque liste :

Anglais	**Finnois**	**Turc**
28 celebrities	1 ennustemallista	20 jazzcI
66 celebrity	2 ennustemallit	3 jazzcIdIr
5 celer	1 ennustemenetelmien	5 jazzcIlar
3 celeres	1 ennustemuutokset	1 jazzcIlardan
1 celeriac	1 ennustepalvelu	5 jazzcIlarI
3 celeries	12 ennustepäällikkö	3 jazzcInIn
2 celeris	1 ennustepäällikön	2 jazzcIsInIn
41 celerity	1 ennusteryhmä	2 jazzcIyI
99 celery	8 ennusteta	35 jazzda
76 celeste	525 ennustetaan	14 jazzdan

Ces listes de mots ont été extraites de diverses sources. La liste finnoise a été acquise à partir de journaux, de dépêches et de livres en version électronique. La liste anglaise a été produite à partir des publications et romans du Projet Gutenberg, une partie du corpus anglais Gigaword ainsi que le corpus Brown. Enfin, la liste turque a été extraite de publications collectées sur Internet, de journaux et de nouvelles sportives. Lors de l'extraction des mots, les organisateurs ont choisi de conserver les marques du possessif en anglais (*'s*) mais ont supprimé les tirets. De plus, compte tenu des sources utilisées, les listes de mots contiennent également des mots étrangers qui peuvent nuire à la qualité des analyses. On trouve par exemple dans la liste de mots anglais un certain nombre de mots allemands, comme *Augenkrankheiten* ou *Geschlechtsempfindungen*, qui font partie des mots les plus longs de la liste.

Les listes de mots à traiter pour le challenge étaient considérablement plus grandes que celles que nous avions eu à traiter jusqu'alors. Pour l'anglais, nous avons effectué l'apprentissage sur la liste complète de mots. Cependant, pour le finnois et le turc, nous n'avons utilisé que les 300 000 mots les plus fréquents, essentiellement pour des problèmes de consommation excessive de mémoire[1].

Des exemples des segmentations attendues pour quelques centaines de mots dans chaque langue ont également été fournis, ainsi que les programmes PERL permettant de calculer la précision, le rappel et la F-mesure par rapport à ces exemples. Nous allons détailler ces mesures dans la section suivante.

Compétition 1

Dans le cadre de la compétition 1, les segmentations proposées ont été comparées avec des segmentations de référence dans les trois langues. Cette évaluation a été effectuée sur un ensemble de mots tenu secret, comprenant 10% des mots des listes fournies pour chaque langue. Le programme d'évaluation, ainsi qu'un échantillon des segmentations attendues, étaient également téléchargeables sur le site Web du challenge. La Figure 4.6 présente un extrait de la trace d'exécution de ce programme d'évaluation, pour une de nos soumissions au challenge en anglais.

```
DES: about, SUG: about, #hits: 0, #ins: 0, #del: 0
DES: accelerate, SUG: accelerat e, #hits: 0, #ins: 1, #del: 0
DES: accurst, SUG: accurs t, #hits: 0, #ins: 1, #del: 0
DES: act ion 's, SUG: action 's, #hits: 1, #ins: 0, #del: 1
DES: adult 's, SUG: adul t 's, #hits: 1, #ins: 1, #del: 0
DES: aero plane s ', SUG: aero plane s ', #hits: 3, #ins: 0, #del: 0
DES: agree ab ly, SUG: agree ably, #hits: 1, #ins: 0, #del: 1
```

FIG. 4.6: Trace de l'exécution du programme d'évaluation de MorphoChallenge.

Les segmentations désirées, présentes dans la liste des segmentations standard, sont marquées par `DES`. Ainsi, la segmentation désirée pour le mot *adult's* est *adult 's* (les frontières morphémiques sont marquées par un espace). La segmentation proposée par notre système, marquée par `SUG`, est *adul t 's*. Pour l'évaluation, le nombre de frontières morphémiques correctement identifiées (`#hits`), insérées (`#ins`) et supprimées (`#del`) est comptabilisé. Dans le cas du mot *adult's*, le système a correctement identifié la frontière entre *adult* et *'s*, mais a inséré une frontière erronée entre *adul* et *t's*. Il y a donc une frontière morphémique correctement identifiée (`#hits: 1`) et une autre insérée (`#ins: 1`) pour ce mot.

A partir de ces décomptes, trois mesures d'évaluation sont calculées :

- La **précision** correspond au nombre de frontières correctement identifiées H divisé par le nombre total de frontières proposées (somme du nombre de frontières correctement identifiées H et insérées I) : $Précision = \frac{H}{H+I}$
- Le **rappel** est le nombre de frontières correctement identifiées H divisé par le nombre total de frontières attendues (somme du nombre de frontières correctement identifiées H et supprimées D) : $Rappel = \frac{H}{H+D}$

[1] Nous avons depuis modifié le programme pour utiliser des structures de données plus efficaces.

– La **F-mesure** est la moyenne harmonique de la précision et du rappel :
$F - mesure = \frac{2 \cdot H}{2 \cdot H + I + D}$

Le système remportant la compétition pour chaque langue est celui qui obtient la plus grande F-mesure.

Pour participer au challenge, nous avons sélectionné les meilleures valeurs pour les paramètres N, a et b en fonction des résultats obtenus pour les données d'évaluation fournies (évaluation partielle). Ces valeurs sont toutefois très proches pour les 3 langues. Le Tableau 4.12 détaille les valeurs de paramètres utilisées et les résultats obtenus. La segmentation finale a été obtenue par réutilisation de la liste de segments obtenus après apprentissage (voir Section 4.2.5, page 85). La méthode 1 correspond aux résultats obtenus en appliquant la première fonction de coût ($coût_1$) et la méthode 2 à ceux obtenus en appliquant la seconde ($coût_2$). La seconde mesure (évaluation finale) correspond aux résultats obtenus lors de la compétition.

				F-mesure			
				Évaluation partielle		Évaluation finale	
Langue	N	a	b	méthode 1	méthode 2	méthode 1	méthode 2
Anglais	5	0.85	0.1	64.29	61.05	66.6	62.4
Finnois	5	0.8	0.1	63.18	64.44	63.3	64.7
Turc	5	0.7	0.1	55.93	66.06	55.3	65.3

Tab. 4.12: Valeurs des paramètres et résultats obtenus pour la compétition 1 de MorphoChallenge.

La Figure 4.7 détaille la F-mesure des 10 systèmes ayant concouru pour l'ensemble des langues, ainsi que les résultats obtenus par les différentes versions du système Morfessor développé par les organisateurs [Creutz et Lagus, 2006]. Les résultats de notre système sont indiqués par Bernhard_1 pour la méthode 1 et Bernhard_2 pour la méthode 2.

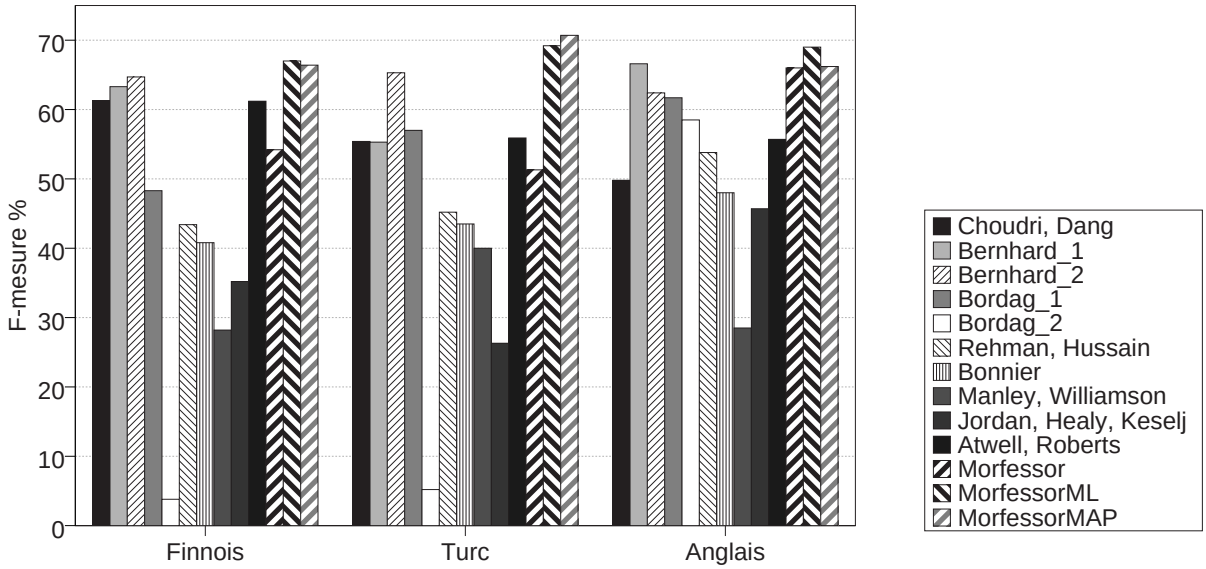


Fig. 4.7: F-mesures obtenues par les différents systèmes pour la compétition 1 de MorphoChallenge.

Notre système a remporté les compétitions à la fois pour le finnois et le turc. Ces résultats sont d'autant plus remarquables que le système n'avait jamais été testé sur d'autres langues que l'anglais ou le français au cours de sa conception. Il a toutefois été dépassé en finnois et en turc par les algorithmes Morfessor ML et MAP (hors compétition car ils ont été développés par les organisateurs). En anglais, la compétition a été remportée par le système[1] de S. Keshava et E. Pitler, de l'université de Yale [Keshava et Pitler, 2006]. Le système, dénommé RePortS, a obtenu une F-mesure de 76.8 %, dépassant ainsi également les systèmes Morfessor. Dans cette compétition, notre système est arrivé en deuxième position.

Cette évaluation a permis de constater une dissymétrie dans les fonctions de coût utilisées. La première obtient de meilleurs résultats pour l'anglais, dont la morphologie est relativement simple. A l'inverse, la seconde fonction de coût obtient de meilleurs résultats pour le finnois et le turc, avec une différence de 10 % de la F-mesure entre les deux fonctions en turc. La seconde fonction permet en réalité de sélectionner un plus grand nombre de segments, ce qui explique les meilleurs résultats obtenus en finnois et en turc, où le nombre des segments différents d'un même mot est bien plus important.

Compétition 2

Pour la compétition 2, les segmentations ont été utilisées pour entraîner un modèle de langage n-gramme pour des expériences en reconnaissance de la parole. Le système remportant la compétition pour chaque langue est celui qui obtient le taux d'erreur par lettre (LER) le plus bas en reconnaissance de la parole. Le taux LER correspond à la somme du nombre de lettres remplacées, insérées et supprimées divisé par le nombre de lettres dans la transcription correcte des données.

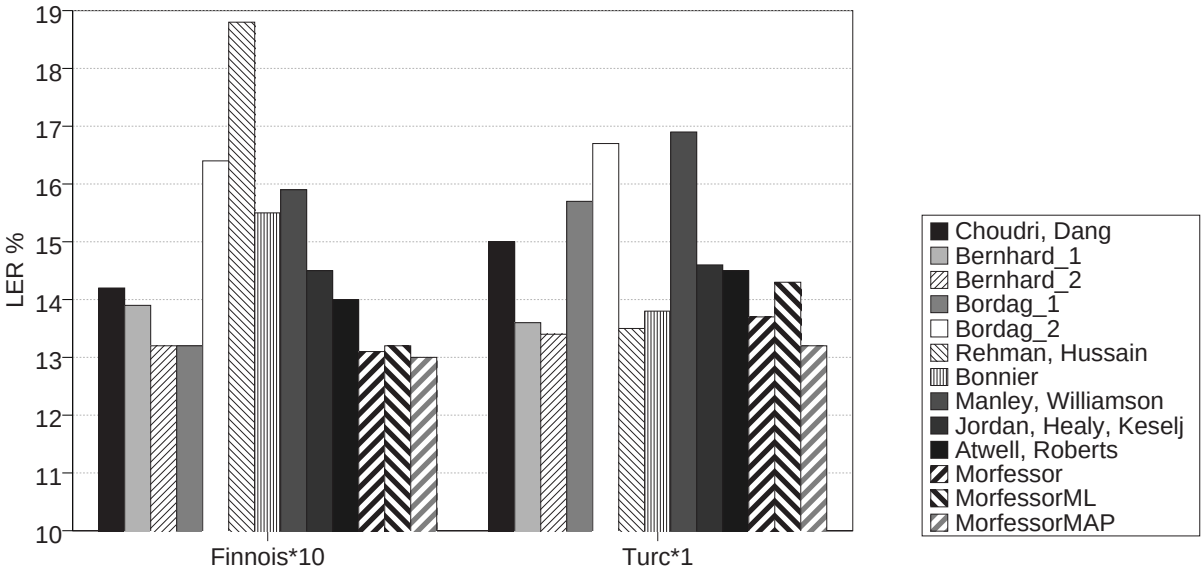


FIG. 4.8: LER des différents systèmes pour la compétition 2 de MorphoChallenge.

[1]Les résultats de ce système n'apparaissent pas sur la Figure 4.7 car il n'a été utilisé que pour segmenter les données en anglais.

La Figure 4.8 détaille les résultats obtenus par les 10 systèmes ayant concouru pour l'ensemble des langues, ainsi que les résultats obtenus par les différentes versions du système Morfessor développé par les organisateurs.

Notre système a également remporté la compétition 2 en turc, et en finnois, à égalité dans cette dernière langue avec l'algorithme de S. Bordag [Bordag, 2006].

Les résultats obtenus par notre système au cours de ce challenge sont très bons, compte tenu de la complexité des données à analyser. Le finnois et le turc sont des langues agglutinantes qui comprennent énormément de suffixes et nous n'avions jamais traité ces langues avant de participer au challenge. De plus, nous étions dans l'incapacité de juger les résultats obtenus et donc dans l'obligation de nous fier aux données d'évaluation. La complexité était également liée aux listes de mots fournies, qui étaient d'une taille importante (allant d'environ 170 000 mots pour l'anglais à plus de 1 600 000 mots en finnois) et qui contenaient de nombreux mots non attestés (mots étrangers, noms propres, mots mal orthographiés).

4.3.2 Évaluation dans le cadre d'une application de synthèse de la parole

Le système a également été évalué par V. Demberg [Demberg, 2006] dans le cadre d'une application de synthèse de la parole pour l'allemand, et plus particulièrement le problème de la conversion des lettres en phonèmes. La conversion de lettres en phonèmes peut être divisée en diverses sous-tâches : alignement des lettres et des phonèmes, syllabification, positionnement de l'accentuation et enfin conversion des graphèmes (chaînes d'une ou plusieurs lettres qui correspondent à un phonème) en phonèmes (g2p). Il a été démontré que l'analyse morphologique augmente les performances des systèmes de synthèse de parole. En effet, les informations morphologiques sont cruciales pour le positionnement de l'accentuation et d'autres phénomènes tels que le dévoisement de certaines syllabes.

Notre système, ainsi que ceux de S. Bordag [Bordag, 2006], S. Keshava et E. Pitler [Keshava et Pitler, 2006] et le système Morfessor de M. Creutz et K. Lagus [Creutz et Lagus, 2002] ont été évalués par V. Demberg pour connaître leurs effets sur les performances d'un système de conversion de graphèmes en phonèmes pour l'allemand. Les résultats sont assez décevants : aucun des 4 systèmes n'améliore les résultats de la conversion. Pire encore, les résultats sont encore plus mauvais qu'ils ne le sont sans analyse morphologique non supervisée. Il semblerait donc que les performances des systèmes de segmentation morphologique non supervisée ne soient pas encore suffisantes pour leur permettre d'être d'une utilité quelconque dans les systèmes de conversion de graphèmes en phonèmes. Selon V. Demberg, une F-mesure supérieure à 70% est indispensable.

4.3.3 Évaluation des familles morphologiques

L'évaluation directe des segmentations ne permet pas de mesurer la validité des liens morphologiques induits, et surtout la validité des liens sémantiques corrélés aux liens morphologiques. Une autre méthode d'évaluation consiste à mesurer la pertinence des familles morphologiques identifiées. Comme nous l'avons montré en Section 4.2.6, p. 85, il est possible d'obtenir des familles morphologiques à partir des segmentations obtenues en regroupant l'ensemble des mots partageant une même base. L'évaluation des familles morphologiques permet également d'estimer, indirectement du moins, la qualité de l'étiquetage des segments obtenus.

Pour procéder à cette évaluation, il est nécessaire de disposer de familles morphologiques de référence. Nous avons utilisé deux sources pour ces familles : nous avons d'une part élaboré

manuellement des listes de référence et, pour l'anglais, nous avons extrait des familles de référence à partir des segmentations proposées par CELEX.

Les listes de référence construites manuellement contiennent des familles de mots pour le domaine du cancer du sein en français et en anglais. Ces listes englobent 3 250 familles en anglais et 1 964 familles en français. Elles ont été élaborées à partir de corpus construits manuellement, et donc de taille réduite par rapport aux corpus utilisés dans nos évaluations (de l'ordre de 10 000 mots différents en français et en anglais).

Nous listons ci-dessous quelques unes des familles présentes dans la liste de familles morphologiques construites manuellement pour le français :

intense	intensifie ; intensité ; intensifier ; intensives ; intensive ; intensifications ; intensification ; intensifs ; intensifiées ; intense
kyste	microkystique ; scléro-kystique ; kystiques ; microkystes ; kystes ; kyste ; polykystiques ; intrakystiques ; fibrokystiques ; kystique ; fibrokystique
statistique	biostatistiques ; statisticiens ; statistique ; statistiques ; statistiquement

La base CELEX comprend quant à elle plusieurs types d'informations en trois langues (allemand, anglais et néerlandais) : orthographe, phonologie, syntaxe, morphologie et fréquences. Nous avons utilisé deux fichiers différents pour l'acquisition de familles morphologiques de référence en anglais à partir de CELEX. Le premier, `eml.cd`, contient notamment la segmentation des différents lemmes et permet ainsi, par récursivité, d'obtenir la base minimale. Le second, `emw.cd`, associe les mots avec leurs lemmes. Les familles morphologiques maximales sont déterminées à partir des lemmes et des relations morphologiques de dérivation, de composition ou de conversion qu'ils entretiennent.

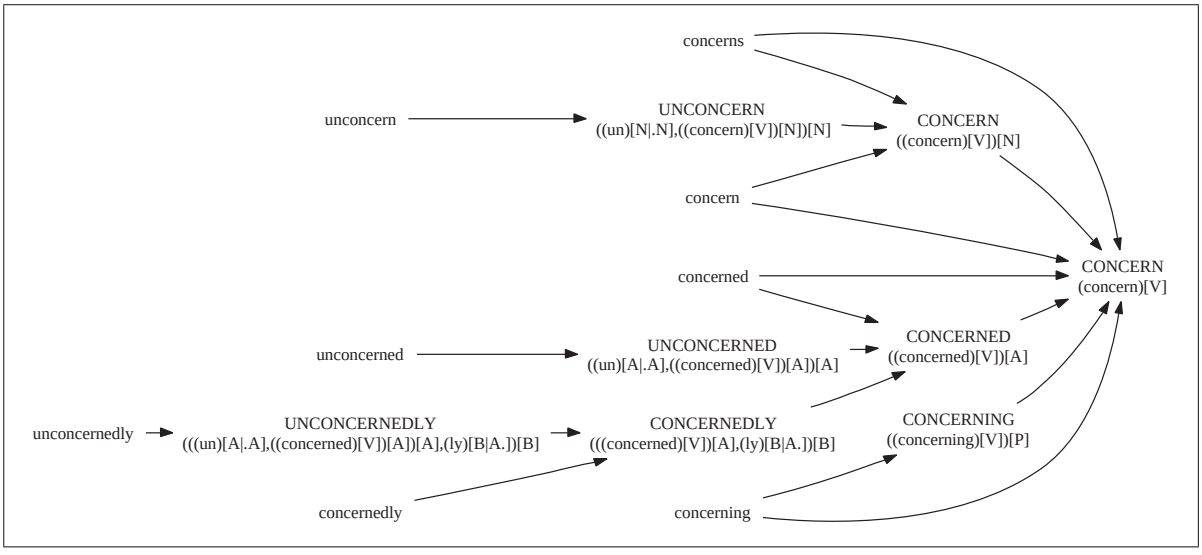


FIG. 4.9: Liens morphologiques dans CELEX. Les mots sont représentés en caractères minuscules tandis que les lemmes apparaissent en majuscule. Les décompositions proposées pour les lemmes sont également représentées.

Prenons l'exemple des mots associés à la base *concern* représentés sur la Figure 4.9. Tous les mots liés à un même lemme appartiennent à la même famille ; c'est le cas par exemple pour *concern*, *concerned*, *concerns* et *concerning* qui ont tous pour lemme le verbe CONCERN. Cette

famille peut être agrandie par l'ajout du mot *unconcerned* qui a pour lemme UNCONCERNED, formé à partir de l'adjectif CONCERNED par ajout du préfixe *un-*. CONCERNED est lui-même formé à partir du verbe CONCERN. En procédant par transivité pour les autres mots, on obtient ainsi la famille morphologique de référence suivante : [*concerned* ; *unconcerned* ; *concerns* ; *unconcernedly* ; *concerning* ; *concernedly* ; *unconcern* ; *concern*].

Après extraction des familles morphologiques de CELEX, nous obtenons 14 880 familles morphologiques pour l'anglais.

Afin de comparer ces familles morphologiques de référence aux familles obtenues par notre système, nous utilisons les mesures proposées par [Schone et Jurafsky, 2000, Schone et Jurafsky, 2001]. Celles-ci comparent les familles morphologiques de chaque mot w selon l'algorithme à évaluer d'une part et selon les données de référence d'autre part. La méthode consiste à faire la somme du nombre de mots corrects (C), insérés (I) et supprimés (D) dans les familles morphologiques de tous les mots de la liste d'évaluation. Si X_w est l'ensemble des mots appartenant à la famille morphologique de w selon le système à évaluer et Y_w est l'ensemble des mots appartenant à la famille morphologique de w selon CELEX ou toute autre base de référence, alors :

$$C = \sum_{\forall w} \frac{|X_w \cap Y_w|}{|Y_w|} \quad ; \quad D = \sum_{\forall w} \frac{|Y_w - (X_w \cap Y_w)|}{|Y_w|} \quad et \quad I = \sum_{\forall w} \frac{|X_w - (X_w \cap Y_w)|}{|Y_w|}$$

Lors du calcul de ces valeurs, seule l'intersection des mots de la base de référence et de la liste de mots analysés par le système est utilisée.

Par exemple, supposons que le système propose la famille suivante et que l'on cherche à évaluer la famille proposée pour le mot *concern* :

[*concern* ; *concerningly* ; *concerns* ; *concerted* ; *concert* ; *concerning* ; *concerned* ; *concern*].

Si *unconcerned*, *unconcernedly*, *concernedly* et *unconcern* n'appartiennent pas à la liste d'apprentissage et *concerningly* n'appartient pas à la base de référence, alors :

X_w = [*concern* ; *concerns* ; *concerted* ; *concert* ; *concerning* ; *concerned* ; *concern*]

Y_w = [*concerned* ; *concerns* ; *concerning* ; *concern*]

Donc, pour le mot *concern* : $C_w = \frac{4}{4} = 1.0$, $D_w = \frac{0}{4} = 0$ et $I_w = \frac{3}{4} = 0.75$. On procède de même pour l'ensemble des mots et on calcule la somme de l'ensemble des valeurs de C_w, D_w et I_w pour obtenir C, D et I.

À partir de ces valeurs, il est également possible de calculer la précision, le rappel (et par conséquent la F-mesure) du système. La précision est égale à $C/(C+I)$ et le rappel à $C/(C+D)$.

Il faut toutefois remarquer que les mesures proposées par Schone et Jurafsky tendent à pénaliser plus fortement les insertions que les suppressions. En effet, I peut être supérieur à 1, tandis que C et D sont toujours inférieurs à 1. Ces mesures favorisent donc les algorithmes qui prennent peu de risques.

Nous avons comparé les résultats obtenus par notre système à ceux d'une méthode *baseline* pour laquelle chaque mot constitue sa propre famille. Une telle méthode obtient une précision de 100% car il n'y a jamais d'insertion. Par contre, le rappel est bien plus faible, car il y a beaucoup de suppressions.

Les résultats pour différentes valeurs des paramètres N et a, les différents corpus et les différentes familles de référence (CELEX et familles construites manuellement) sont présentés dans les Figures 4.10 à 4.15.

Ces différents résultats présentent des tendances communes, indépendamment du domaine et de la langue : augmentation de la précision lorsque la valeur de a augmente, parallèlement à une diminution modérée du rappel. Ceci correspond aux effets attendus du paramètre : plus a

est grand, moins le système accepte les segmentations contenant des affixes non acquis lors de la première étape. La F-mesure augmente également lorsque a augmente. Elle est généralement meilleure que celle de la méthode *baseline*. Il faut toutefois remarquer que la précision (et par conséquent la F-mesure) obtenue pour le corpus volcano-en avec N=10 est assez faible. Cette chute de la précision est due au grand nombre d'affixes extraits pour N=10 (voir Table 4.7, p. 78), et donc vraisemblablement au plus grand nombre d'affixes non valides extraits. La méthode est effectivement sensible à la quantité et à la qualité des affixes extraits, les meilleurs résultats étant obtenus pour N=5.

Les résultats sont légèrement moins bons en français. Cette tendance est identique pour la méthode baseline : le rappel en anglais est supérieur à 30%, tandis que le rappel en français se situe au-delà de la barre des 20%. Ceci reflète la plus grande complexité morphologique du français, qui a pour conséquence des familles morphologiques de taille plus importante, ne serait-ce que pour les verbes. En effet, un verbe anglais peut prendre au plus 4 formes différentes : infinitif (*to dance*), prétérit (*danced*), gérondif (*dancing*) et présent, à la troisième personne du singulier (*dances*). En français, la conjugaison verbale est plus complexe.

On constate également que les résultats obtenus par comparaison avec les ressources manuelles sont généralement meilleurs, surtout en ce qui concerne la mesure de précision. Ces ressources contiennent un nombre plus faible de familles et de mots et les résultats ont donc été obtenus uniquement à partir de moins de 10 000 mots (voir Tableau 4.13). Lorsque le nombre de mots est moins important, il y a moins de cas d'insertion et nous avons vu que les mesures d'évaluation pénalisent plus fortement les cas d'insertion que de suppression. Il faut également noter que CELEX est loin de couvrir l'ensemble du vocabulaire de nos corpus. Il y a plusieurs raisons à cela. D'une part, les listes de mots utilisées contiennent nombre de mots propres et de mots mal orthographiés qui n'ont pas été filtrés. À cela s'ajoute la grande spécialisation des corpus, qui contiennent une grande proportion de mots techniques et de néologismes. Les problèmes de couverture de CELEX ont déjà été mis en évidence auparavant par [Schone et Jurafsky, 2000] notamment. Nous avons également constaté l'absence de certains liens morphologiques dans CELEX : par exemple, aucun découpage n'est proposé pour le mot *seismic*, qui n'est donc pas rattaché aux autres membres de sa famille morphologique comme *seism* ou *seismograph*. De plus, certains lemmes ne correspondent pas à leur décomposition : par exemple, le nom *crude* est en réalité une ellipse du composé *crude oil*, et sa décomposition, `(((crude)[A],(oil)[N])[N])[N]`, ne correspond donc pas à la forme de surface du lemme.

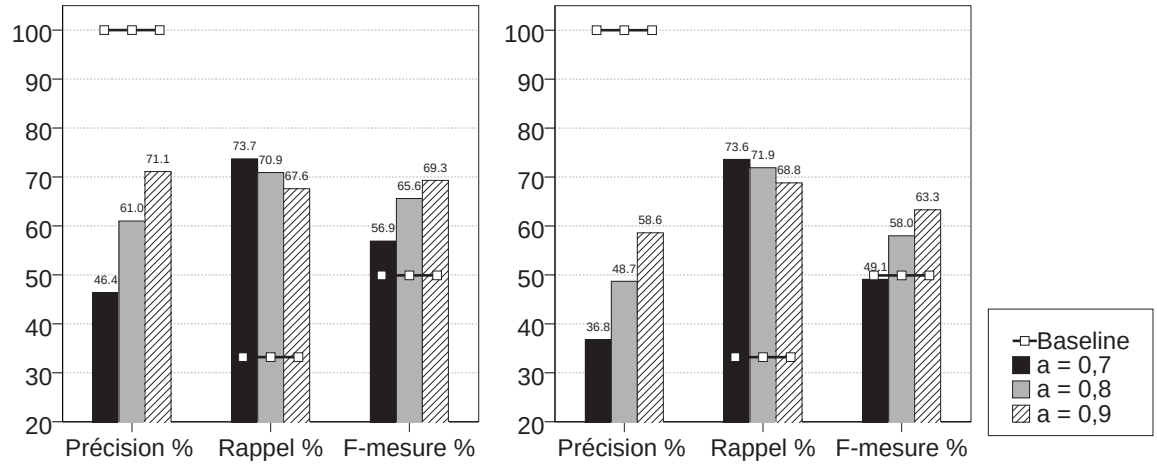
	CELEX	liste manuelle
cancer-en	22 174	9 064
volcano-en	18 643	7 065

(a) Corpus anglais

	liste manuelle
cancer-fr	8 544
volcano-fr	6 810

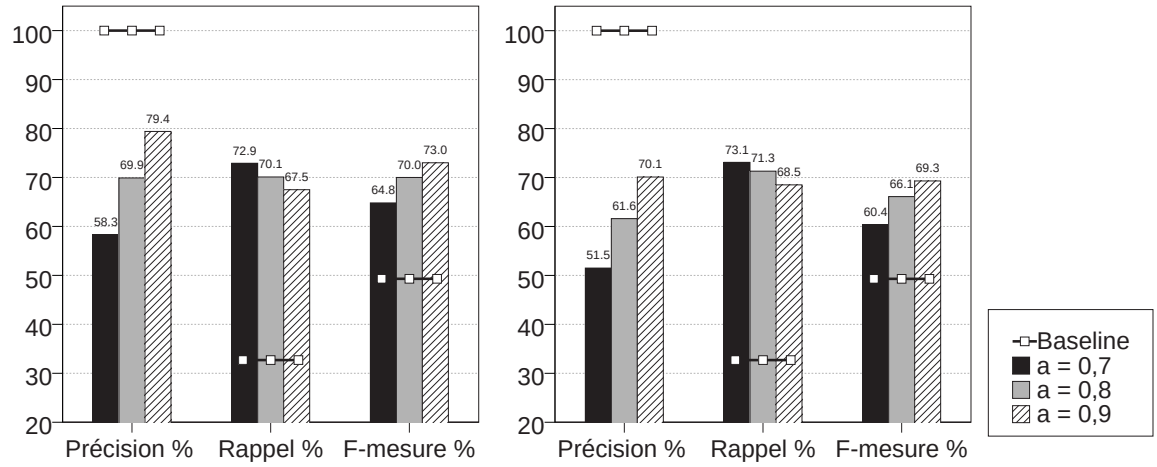
(b) Corpus français

Tab. 4.13: Nombre de mots évalués en fonction du corpus et des ressources de référence.



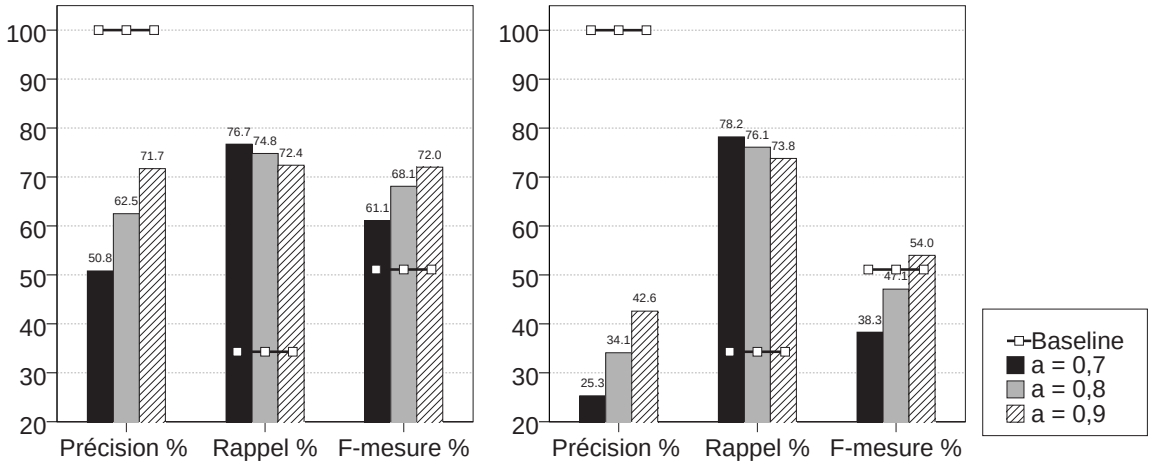
(a) N=5 (b) N=10

FIG. 4.10: Résultats de l'évaluation pour le corpus cancer-en, par comparaison avec CELEX et pour différentes valeurs de a et de N : (a) N = 5 ; (b) N = 10 (b constant à 0.1).



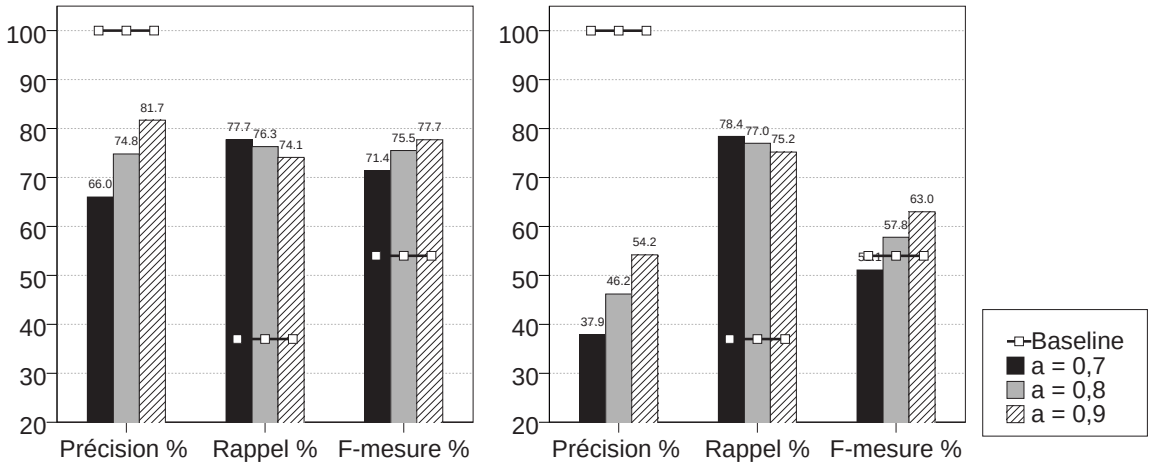
(a) N=5 (b) N=10

FIG. 4.11: Résultats de l'évaluation pour le corpus cancer-en, par comparaison avec la liste de référence manuelle et pour différentes valeurs de a et de N : (a) N = 5 ; (b) N = 10 (b constant à 0.1).



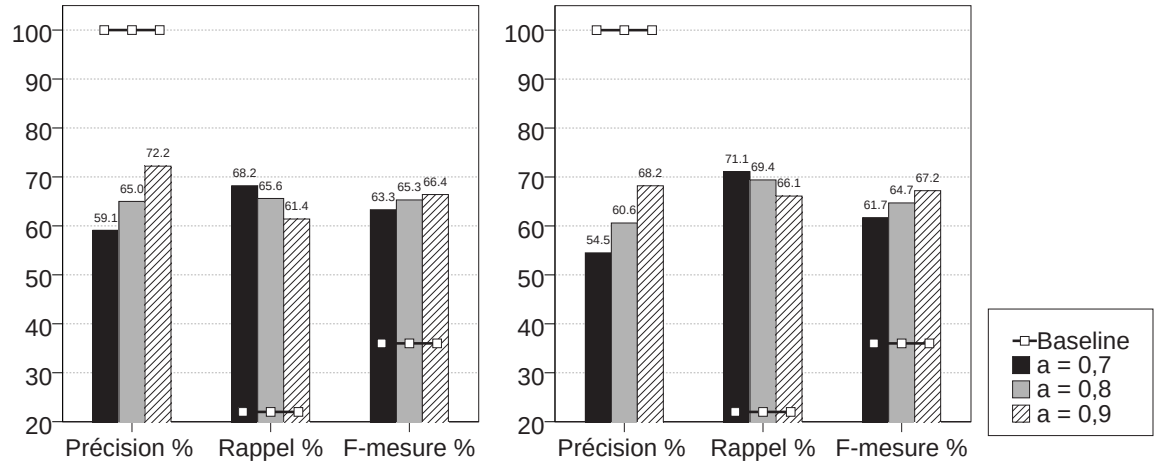
(a) N=5 (b) N=10

FIG. 4.12: Résultats de l'évaluation pour le corpus volcano-en, par comparaison avec CELEX et pour différentes valeurs de a et de N : (a) N = 5 ; (b) N = 10 (b constant à 0.1).



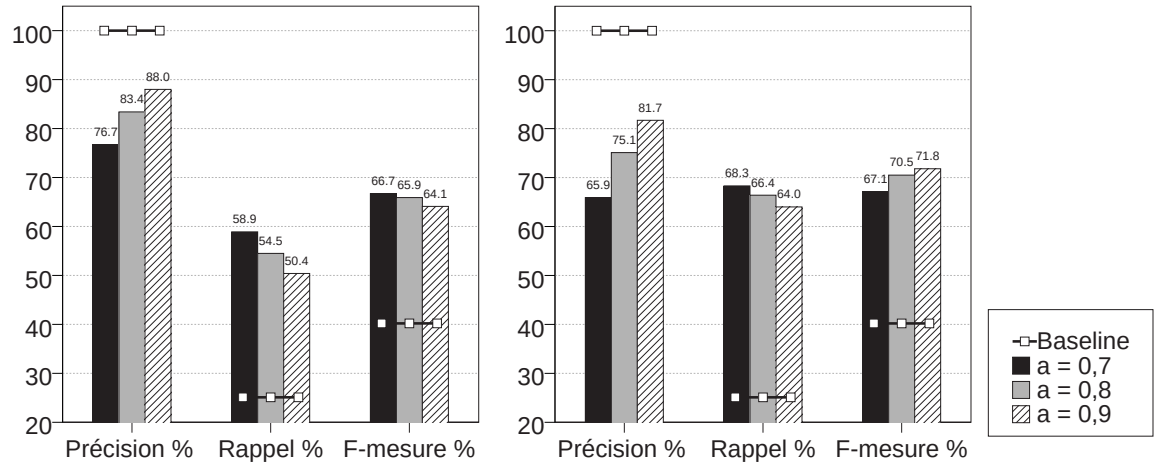
(a) N=5 (b) N=10

FIG. 4.13: Résultats de l'évaluation pour le corpus volcano-en, par comparaison avec la liste de référence manuelle et pour différentes valeurs de a et de N : (a) N = 5 ; (b) N = 10 (b constant à 0.1).



(a) N=5 (b) N=10

FIG. 4.14: Résultats de l'évaluation pour le corpus cancer-fr, par comparaison avec la liste de référence manuelle et pour différentes valeurs de a et de N : (a) N = 5 ; (b) N = 10 (b constant à 0.1).



(a) N=5 (b) N=10

FIG. 4.15: Résultats de l'évaluation pour le corpus volcano-fr, par comparaison avec la liste de référence manuelle et pour différentes valeurs de a et de N : (a) N = 5 ; (b) N = 10 (b constant à 0.1).

4.3.4 Analyse des résultats

L'analyse quantitative des résultats est cohérente avec les résultats obtenus au cours du challenge de segmentation morphologique. Le système obtient de bons résultats, avec des valeurs de la F-mesure supérieures à 60, voire 70 %. Nous allons maintenant analyser les résultats de manière qualitative, afin d'identifier les problèmes principaux de la méthode.

On distingue généralement deux défauts dans l'évaluation des algorithmes de segmentation : la sur-segmentation et la sous-segmentation. La sur-segmentation correspond à l'insertion de frontières morphémiques incorrectes, ce qui induit une baisse de la précision. La sous-segmentation correspond quant à elle à l'absence de frontières morphémiques désirées, ce qui induit une baisse du rappel. Le Tableau 4.14 donne quelques exemples de segmentations correctes, de sur- et de sous-segmentations.

Segmentations correctes	**Sur-segmentations**	**Sous-segmentations**
$\text{ultra}_\text{p} + -_\text{l} + \text{violet}_\text{b}$	$\text{mari}_\text{b} + \text{n}_\text{s} + \text{s}_\text{s}$	$\text{phréat}_\text{b} + \text{o}_\text{l} + \text{magmati}_\text{b} + \text{que}_\text{b}$
$\text{transport}_\text{b} + \text{é}_\text{s} + \text{e}_\text{s}$	$\text{épi}_\text{b} + \text{ques}_\text{s}$	touchant_b
$\text{dériv}_\text{b} + \text{ation}_\text{s}$	$\text{con}_\text{p} + \text{sul}_\text{b} + \text{s}_\text{s}$	$\text{proposait}_\text{b}$
$\text{périod}_\text{b} + \text{ique}_\text{s} + \text{s}_\text{s}$	$\text{cav}_\text{b} + \text{a}_\text{s} + \text{le}_\text{s}$	$\text{réintroduit}_\text{b}$
$\text{micro}_\text{p} + \text{fossil}_\text{b} + \text{e}_\text{s}$	$\text{ent}_\text{p} + \text{rai}_\text{b} + \text{n}_\text{s}$	$\text{éloquemment}_\text{b}$

TAB. 4.14: Exemples de segmentations correctes, de sur- et de sous-segmentations, tirés des résultats pour le corpus volcano-fr avec N=5, a=0,9 et b=0,1.

Les cas de sur-segmentation sont notamment liés à la détection d'affixes qui peuvent être valables dans d'autres contextes, comme *+ques* dans *épiques*, ou *con+* dans *consuls*. Ceci constitue des cas difficiles à traiter pour le système. En effet, la segmentation de *marins* en $\text{mari}_\text{b} + \text{n}_\text{s} + \text{s}_\text{s}$ va relier *marins* à *mari*, mais c'est également le seul segment commun avec des mots comme *maritime*, qui appartient à la même famille que *marins*. On atteint ici les limites des méthodes basées uniquement sur les graphies : la seule manière de résoudre ce type de problèmes serait d'utiliser des mesures de similarité sémantique.

Les cas de sous-segmentation sont quant à eux directement liés à la faible fréquence de certaines bases et affixes dans le corpus. Considérons le cas de *éloquemment*. On trouve dans le corpus les mots suivants, qui appartiennent à la même famille : *éloquente*, *éloquents*, *éloquentes* et *éloquence*. Pour pouvoir reformer cette famille à partir de la segmentation des mots, il faudrait que la base *éloque* soit reconnue dans chacun de ces mots. Or, les suffixes *+nte*, *+nts*, *+ntes* *+nce* et *+mment* ne sont pas découverts lors de la phase d'apprentissage des affixes. Le mot *éloquemment* n'est donc pas segmenté. On touche là au problème de l'acquisition d'une liste d'affixes qui est sufisamment précise, tout en ayant une couverture suffisante. La méthode d'acquisition de la liste d'affixes présente également un autre défaut, qui est plus gênant en français qu'en anglais : les terminaisons verbales sont moins bien reconnues que les autres. Ceci explique la non segmentation de *proposait*, ainsi que d'autres mots de la même famille. Les suffixes verbaux *+ait*, *+era*, *+ons*, *+eront* et *+erai* ne sont pas identifiés lors de l'étape 1 d'acquisition des affixes. Nous avons fait le choix de ne segmenter que les mots les plus longs lors de cette étape, d'une part parce que les méthodes basées sur la prévisibilité sont plus efficaces lorsque les mots sont longs et d'autre part parce que les mots longs sont également ceux qui sont les plus susceptibles d'être affixés. Or les mots longs sont généralement des noms composés, on trouve peu de verbes.

Comme nous avons déjà pu l'entrevoir par quelques exemples, certains problèmes sont liés à la méthode elle-même, à savoir l'analyse morphologique par segmentation. Il est bien sûr difficile

de prendre en compte les cas de variation des radicaux, notamment les différences d'accentuation ou les phénomènes de doublement des consonnes en fin de radical. Ceci conduit à la formation de familles distinctes, alors même que les mots sont morphologiquement liés, et donc produit une baisse du rappel. Nous présentons dans le Tableau 4.15 quelques exemples de familles disjointes. Ce phénomène se produit en français aussi bien qu'en anglais.

Base	**Famille**
dimensionn	tridimensionnel ; dimensionnement ; dimensionnées ; tridimensionnelle
dimension	dimension ; dimensions
cratère	cratère-dôme ; cratères ; cratères-trous ; intra-cratère ; demi-cratère ; cratère-puits ; intracratère ; cratère
cratéri	cratérisés ; cratérisées ; intracratérique ; cratérique ; intracratériques ; intra-cratérique
try	try ; trying
tri	trials ; tried ; tries ; trial

Tab. 4.15: Exemples de familles, tirés des résultats pour les corpus volcano-fr et volcano-en avec N=5, a=0,9 et b=0,1.

On peut distinguer trois types principaux d'alternances dans les bases conduisant à ce phénomène :

- Doublement des consonnes en fin de radical : *dimension**n*** – *dimension*
- Changement d'accentuation : *crat**è**re* – *crat**é**ri*.
- Remplacement d'une lettre par une autre : *try* – *tri*. Ce dernier exemple met également en évidence les problèmes éventuels liés au seuil de longueur minimale d'une base, que nous avons fixé à 3.

Ces problèmes pourraient être résolus a posteriori, après la segmentation, par regroupement des bases résultant d'une alternance. Ce regroupement pourrait prendre en compte à la fois des informations de similarité graphique des bases ainsi que des informations contextuelles permettant d'estimer la similarité sémantique.

4.4 Conclusion

Nous avons présenté une méthode d'analyse morphologique non supervisée qui découpe les mots en segments étiquetés. Le système obtient de très bons résultats, en comparaison à d'autres systèmes similaires. Les évaluations ont été réalisées pour des langues très différentes, comme l'anglais, le français, le finnois et le turc. À cela s'ajoutent des expériences conduites en allemand pour la conversion de graphèmes en phonèmes. Ces résultats ont été obtenus pour des listes de mots issues à la fois de la langue générale (Morpho Challenge) et de domaines de spécialité comme la médecine ou la volcanologie. Le système n'est donc ni dépendant de la langue traitée, ni du domaine dont est issu le vocabulaire.

Les résultats des diverses évaluations démontrent l'utilité des indices que nous avons utilisés : probabilités transitionnelles, comparaison de graphies, longueur et fréquence des segments morphémiques, contraintes morphotactiques. L'idée d'utiliser les probabilités transitionnelles a été directement inspirée de travaux en psychologie cognitive qui montrent que les êtres humains utilisent ce type d'informations pour segmenter la parole en mots.

Il faut toutefois remarquer que pour traiter l'ensemble des procédés de formation morphologique, et notamment la composition, il est nécessaire de combiner un certain nombre d'indices.

En effet, l'utilisation des probabilités transitionnelles seules ne permettrait pas d'obtenir des résultats de qualité suffisante. Ceci explique pourquoi notre système, ainsi que les autres systèmes qui ont des résultats équivalents, sont relativement complexes et combinent divers types d'informations. Nous n'avons toutefois pas encore exploité l'ensemble des données disponibles en corpus. En effet, nous n'avons pas pris en compte la fréquence des mots dans le corpus, ainsi que leur contexte d'occurrence. Tous ces éléments pourraient encore améliorer le système, s'ils sont utilisés à bon escient. Avant de les intégrer, il faudrait déterminer les étapes auxquelles ces informations pourraient être bénéfiques. Nous avons vu que les données contextuelles peuvent intervenir à différents moments dans l'analyse morphologique non supervisée : soit en tout début d'analyse, soit à la fin. Ces diverses pistes sont à explorer, même si leur intégration dans le système actuel demande réflexion.

Le système que nous venons de présenter dans ce chapitre produit une analyse morphologique des mots par segmentation et étiquetage des segments obtenus. Il est possible d'obtenir des familles morphologiques à partir de cette segmentation, avec des résultats qui dépendent bien sûr fortement de la qualité de la segmentation et de l'étiquetage. Afin de contourner certaines des limites que nous avons identifiées dans la section précédente, nous avons développé un autre système d'analyse morphologique non supervisée, dont l'objectif premier est non pas de segmenter les mots, mais de les regrouper pour former des familles morphologiques. Ce système est décrit dans le chapitre suivant.

Chapitre 5

Analyse morphologique par classification

5.1 Introduction

Nous avons vu que le système d'analyse par segmentation produit d'assez bon résultats, mais achoppe sur des cas de variation du radical (accentuation, doublement de consonnes). Face à ces problèmes, nous avons développé un autre système, en considérant l'analyse morphologique non plus comme un problème de découpage de mots, mais comme un problème de classification, visant à regrouper les mots sans forcément procéder à leur segmentation. Les classes de mots obtenues sont des familles morphologiques, comme par exemple [mesura ; mesurée ; mesure ; mesurant ; mesuré ; mesurait ; mesurés ; mesurer ; ...]. Ce système doit permettre l'acquisition de familles morphologiques, tout en utilisant des critères d'équivalence entre radicaux plus souples, rendant ainsi possible la résolution de certains des problèmes de l'analyse par segmentation.

La classification est avant tout une technique d'analyse de données. Nous allons tout d'abord décrire les méthodes de classification de ce point de vue, avant d'aborder les applications possibles de la classification en traitement automatique des langues et plus particulièrement pour l'acquisition de connaissances morphologiques.

5.1.1 La classification en analyse de données

Les algorithmes de classification classiques visent à organiser un ensemble d'éléments en groupes homogènes et contrastés, aussi appelés classes ou catégories (*clusters* en anglais)[1]. Ces méthodes sont notamment utilisées pour faciliter l'analyse des informations. Il existe deux types d'algorithmes de classification :

- **Classification non hiérarchique** : un ensemble de groupes initiaux sont améliorés de manière itérative, de sorte à obtenir les meilleurs groupes possibles. La méthode des k-means est un exemple d'algorithme de classification non hiérarchique.
- **Classification hiérarchique** : le résultat de la classification est une hiérarchie de groupes. Un groupe d'éléments correspond dans ce cas à l'union de deux groupes de niveau inférieur. On distingue généralement deux approches :
  - Les approches par **partitionnement** (ou **descendantes**) qui visent à obtenir des groupes de taille inférieure à partir d'un groupe contenant l'ensemble des données. À

[1] Voir [Manning et Schütze, 1999, chapitre 14] ou [Heyer *et al.*, 2006] pour une revue des algorithmes de classification

chaque étape, le groupe dont la cohérence interne est la plus faible est divisé en sous-groupes. Le processus se termine lorsqu'il ne reste plus que des groupes à un élément.
- Les approches par **agglomération** (ou **ascendantes**) qui cherchent à construire des classes de taille supérieure à partir d'un ensemble de groupes contenant chacun un élément. À chaque étape de la classification, les deux groupes d'éléments les plus similaires sont fusionnés afin de former un nouveau groupe. Le processus se termine lorsque tous les groupes ont été agglomérés dans un seul groupe, qui forme la racine de la hiérarchie. Les feuilles de la hiérarchie correspondent aux éléments initiaux analysés.

Les méthodes de classification sont généralement basées sur une mesure de la distance (ou de la similarité) entre les objets à classer. Elles visent à minimiser la distance à l'intérieur des groupes et à maximiser la distance entre les groupes.

Aux méthodes de classification que nous venons de décrire s'ajoutent les cartes auto-organisatrices de Kohonen (ou SOM pour Self-Organising Maps) et les algorithmes basés sur les graphes. Les cartes auto-organisatrices de Kohonen permettent de projeter des données sur un espace à deux dimensions subdivisé en sous-unités de sorte que les données proches sont groupées dans la même unité. Les algorithmes basés sur les graphes consistent à repérer soit les zones de forte densité, soit les cliques, qui sont des sous-graphes complets dont tous les sommets sont connectés deux à deux.

5.1.2 Classification et traitement automatique des langues

Les méthodes de classification décrites précédemment ont diverses applications pour l'analyse de données textuelles. Elles peuvent notamment être utilisées pour la classification de documents. Dans ce cas, les documents doivent d'abord être représentés sous forme de vecteurs contenant la mesure de fréquence normalisée des termes de l'index. Ces vecteurs permettent de calculer une mesure de similarité entre documents, utilisée par l'algorithme de classification. On peut procéder de manière similaire pour classer les mots en fonction de leur similarité contextuelle. Dans ce cas, les vecteurs contiennent la fréquence normalisée des mots co-occurrents, dans une fenêtre de taille donnée.

La classification pour l'analyse de données morphologiques nécessite une approche différente, car il ne s'agit pas de mesurer la similarité sémantique des mots, mais la similarité structurelle des mots. L'intérêt pour la classification des mots appartenant à la même famille morphologique est ancien, du fait des besoins suscités par la recherche d'informations. En effet, les regroupements effectués à partir de la similarité morphologique constituent des classes d'équivalence, utilisables pour l'indexation ou pour l'extension de requête. Ces classes d'équivalence sont susceptibles d'améliorer le rappel, c'est-à-dire le nombre de documents pertinents retrouvés. Les algorithmes de racinisation, présentés en Section 2.3.1, sont une des méthodes communément utilisées. La racinisation est effectuée par l'application de règles permettant de réduire les mots à des radicaux partagés avec d'autres mots. Une classe d'équivalence regroupe alors l'ensemble des mots partageant le même radical. Il ne s'agit toutefois pas d'une méthode de classification, dans le sens donné par le domaine de l'analyse des données, à cause de l'utilisation de règles explicites. D'autres méthodes visent à obtenir des résultats similaires, de manière non supervisée ou peu supervisée. Elles ont pour objectif de grouper les mots en familles de sorte que tous les mots de la famille partagent un même radical. Les familles de mots ainsi formées sont des familles morphologiques.

Il existe deux types de critères pour la classification morphologique non supervisée. Les méthodes proches des algorithmes de classification classiques utilisent des mesures de similarité entre mots basées notamment sur la similarité graphique. La classification peut également être

effectuée indépendamment de toute mesure de similarité ou de distance, par des critères de regroupement explicites.

Les mesures de similarité morphologique proposées sont parfois très simples. Ainsi, [Adamson et Boreham, 1974] proposent de calculer un coefficient de similarité des mots (coefficient de Dice) basé sur les digrammes, c'est-à-dire les suites de deux lettres, partagés par les mots. Les mots sont alors regroupés par un algorithme de classification ascendante hiérarchique utilisant ces coefficients. Cette méthode de regroupement est également applicable au regroupement de chaînes plus longues, à savoir les titres de documents. [Gaussier, 1999] utilise également un algorithme de classification ascendante hiérarchique mais définit une mesure de similarité différente, basée non pas sur la similarité orthographique mais sur la productivité des suffixes contenus dans chaque mot. [Jacquemin, 1997] définit quant à lui une mesure de distance entre suffixes permettant de calculer une distance entre classes.

Contrairement aux méthodes décrites précédemment, l'algorithme proposé par [Hathout, 2005] n'utilise aucune mesure de distance morphologique, mais repère les cliques dans un graphe morphologique, construit à partir de schémas d'affixation. Ces cliques correspondent à des familles morphologiques car dans les cliques, tout comme dans les familles morphologiques, tous les éléments sont liés les uns aux autres et donc apparentés. Ce type de méthode est à rapprocher des méthodes d'identification des sens des mots à partir des zones de forte densité dans des graphes de co-occurrence [Véronis, 2004].

L'approche que nous proposons est assez différente des exemples cités précédemment. Celle-ci repose sur divers indices, dont la similarité graphique. Elle est à rapprocher des méthodes de classification ascendante hiérarchique car la construction des familles morphologiques se fait par fusion de familles de taille inférieure. Mais la similarité avec les algorithmes de classification s'arrête là : nous n'utilisons pas de mesure de similarité et le processus ne se termine pas par la formation d'une famille morphologique unique. Nous allons maintenant décrire notre méthode.

5.2 Description de la méthode

La Figure 5.1 présente l'architecture globale du système, qui comprend 3 étapes principales. Ces diverses étapes seront détaillées dans les sections suivantes.

5.2.1 Données

Le système repose sur trois types de données :
- Une liste des mots d'un corpus L.
- Une liste de préfixes P
- Une liste de signatures S, c'est-à-dire des paires de suffixes.

Les deux dernières listes sont obtenues à l'aide du module d'apprentissage d'affixes décrit en Section 4.2.1, p. 73. Nous avons adapté ce dernier module pour qu'il produise non seulement une liste de préfixes et de suffixes mais également une liste de signatures. Les signatures sont des paires de suffixes apparaissant avec la même base et qui sont donc mutuellement substituables sur l'axe paradigmatique. Par exemple, les suffixes de la paire (*s*,*ique*) peuvent se combiner à la base *climat* pour former les mots *climat**s*** et *climat**ique***. La même signature se retrouve dans les paires de mots *volcan**s*** – *volcan**ique*** et *océan**s*** – *océan**ique***.

La notion de signature est présente dans de nombreux travaux en acquisition automatique de connaissances morphologiques, parfois sous des dénominations différentes. [Gaussier, 1999] utilise tout simplement la dénomination de *paires de suffixes*. [Grabar et Zweigenbaum, 1999] parlent de *règles morphologiques* et [Hathout, 2005] de *schémas de suffixation*. [Freitag, 2005] utilise

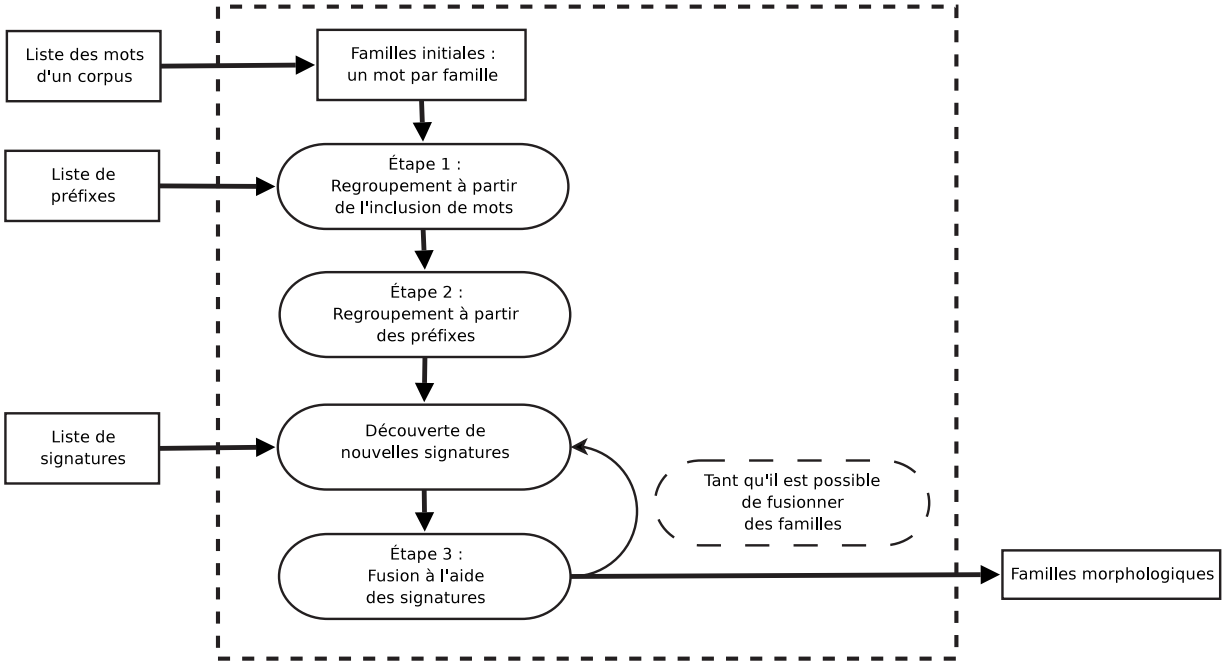


FIG. 5.1: Architecture globale du système d'analyse morphologique par classification.

quant à lui le terme de *transformations* (*transforms*). [Goldsmith, 2001] définit la signature d'un radical comme l'ensemble des suffixes avec lesquels il apparaît. Cet ensemble de suffixes forme un paradigme, comme par exemple `NULL.ing.ingly`, associé entre autres à *despair* (*despair, despairing, despairingly*), *pity* (*pity, pitying, pityingly*) et *insult* (*insult, insulting, insultingly*). La notion de signature peut également être étendue à la description des termes composés de deux mots [Jacquemin, 1997]. Dans ce cas, la signature correspond en réalité à un quadruplet de suffixes. En effet, la signature du couple de termes (*continuous measure-ment*, *continuous-ly measure-d*) est $\{(\epsilon,ment),(ly,d)\}$.

Les listes de préfixes et de signatures nécessaires au système peuvent bien sûr être acquises de diverses manières. Dans une version initiale de l'outil, présentée dans [Bernhard, 2006b], nous proposions d'acquérir automatiquement les préfixes à partir de l'expression régulière suivante : `([aio]-)?(\w{3,}[aio])-`. Celle-ci permet de reconnaître les chaînes de caractères de longueur supérieure ou égale à 4, se terminant par *a*, *i* ou *o* et immédiatement suivies par un tiret. La première partie de l'expression régulière prend en compte les mots comportant plusieurs préfixes les uns à la suite des autres (comme par exemple dans le mot *hépato-gastro-entérologues*). Cette expression régulière permet l'identification d'un certain nombre de préfixes et d'éléments de formation de composés savants. Elle a de plus l'avantage d'être applicable à plusieurs langues, notamment le français (*chimio-radiothérapie*), l'anglais (*chemo-radiotherapy*) et l'allemand (*Chemo-radiotherapie*). Nous n'utilisons pas cette méthode d'identification des préfixes dans les expériences décrites par la suite, mais elle constitue une des variantes possibles du système.

Nous allons maintenant détailler l'ensemble des étapes menant à la construction des familles morphologiques.

5.2.2 Familles initiales

Avant apprentissage, il y a autant de familles que de mots dans la liste donnée en entrée : chaque mot constitue sa propre famille. L'objectif est de regrouper les mots pour arriver à une partition optimale des données.

Les familles formées au cours du processus d'apprentissage sont définies de la manière suivante :

> Chaque famille F est caractérisée par les informations suivantes :
> - un radical R, permettant de représenter la famille ;
> - deux sous-familles, c'est-à-dire des familles englobées dans F ;
>
> lorsque la famille correspond à une feuille dans la hiérarchie, la famille contient un mot m unique et n'a pas de sous-familles.

5.2.3 Étape 1 : regroupement de familles à partir de l'inclusion de mots

Le premier critère de regroupement des familles est l'inclusion de mots : il s'agit de repérer les mots formés par préfixation à partir d'un autre mot de la liste, selon une procédure détaillée ci-dessous :

> Soient :
> – $F_1, F_2, \ldots, F_i$ des familles telles que $F_1 = [m_1], F_2 = [m_2], \ldots, F_i = [m_i]$;
> – F_j une famille telle que $F_j = [m_j]$;
> – $m_1, m_2, \ldots, m_i$ et m_j des mots de longueur minimale égale à 4.
>
> Les familles $F_1, F_2, \ldots, F_i$ et F_j sont regroupées pour former une nouvelle famille F_k si $m_1 = E + m_j, m_2 = E + m_j, \ldots, m_i = E + m_j$
> où E représente la suite maximale d'un ou plusieurs préfixes de la liste P, éventuellement séparés par des tirets, tels que chaque préfixe ait une longueur minimale de 3.
> Le radical de la nouvelle famille F_k est m_j.

Par exemple, si F_1 = [subvolcaniques], F_2 = [sub-volcaniques], F_3 = [post-volcaniques] et F_4 = [volcaniques], alors il est possible de former une nouvelle famille F_5 telle que $F_5 = F_1 \cup F_2 \cup F_3 \cup F_4$ = [subvolcaniques, sub-volcaniques, post-volcaniques, volcaniques]. En effet, les mots *subvolcaniques*, *sub-volcaniques* et *post-volcaniques* contiennent tous le mot *volcaniques*. De plus, ils débutent par les préfixes *sub+*, *sub-+* et *post-+*. Le radical de la nouvelle famille est le mot inclus dans l'ensemble des autres mots de la famille, à savoir *volcaniques*.

5.2.4 Étape 2 : regroupement de familles à partir des préfixes

Après avoir procédé à un premier regroupement des mots en fonction des mots inclus, nous utilisons d'autres critères de regroupement, basés sur la double constatation suivante :

1. Les méthodes basées sur la comparaison de graphies, et notamment la méthode du plus long préfixe commun, ne sont pas suffisamment précises si elles sont utilisées de manière indépendante. Il est donc nécessaire de les combiner à d'autres données, comme des informations contextuelles ou des ressources externes, permettant de restreindre l'ensemble des possibilités [Jacquemin, 1997].

2. Les informations internes au mot peuvent s'avérer suffisantes pour utiliser la méthode du plus long préfixe commun. En effet, lorsque deux mots partagent un même préfixe et que leurs bases sont graphiquement similaires, alors il y a de fortes chances pour qu'ils soient également morphologiquement liés. Prenons l'exemple des mots suivants : *neuro-oncologist* et *neuro-oncology*. Ces deux mots débutent tous deux par le préfixe *neuro-* suivi d'une même chaîne de caractères de longueur 7 : *oncolog*. La combinaison de deux indices, à savoir le partage d'un préfixe, suivi d'une chaîne commune, est un indice suffisant dans la plupart des cas pour conclure que les mots sont morphologiquement liés.

Nous appliquons ces remarques de la manière suivante :

Soient :
- F_1 et F_2 deux familles ;
- R_1 le radical représentant F_1 ;
- R_2 le radical représentant F_2.

Les deux familles F_1 et F_2 sont regroupées dans une nouvelle famille F_3 ssi :

1. $R_1 = \alpha + s_1$ et $R_2 = \alpha + s_2$, où α est une chaîne de caractères de longueur minimale égale à 4 et s_1 et s_2 sont des chaînes de caractères différant au moins par leur premier caractère.
2. Il existe au moins un mot $m_1 \in F_1$ et un mot $m_2 \in F_2$ tels que m_1 et m_2 incluent le même préfixe.

Le radical R_3 de la nouvelle famille F_3 est le mot le plus court parmi R_1 et R_2.

Par exemple, si :
- F_1 = [paléo-volcan, volcan] avec R_1 = *volcan* ;
- F_2 = [subvolcanique, post-volcanique, néovolcanique, paléovolcanique, volcanique] avec R_2 = *volcanique*

alors il est possible de former une nouvelle famille :
$F_3 = F_1 \cup F_2$ = [paléo-volcan, volcan, subvolcanique, post-volcanique, néovolcanique, paléo-volcanique, volcanique].

En effet, R_1 et R_2 partagent une chaîne initiale commune de longueur 6, *volcan*, et les mots *paléo-volcan* de F_1 et *paléovolcanique* de F_2 ont en commun le préfixe *paléo*. Le radical de F_3 est le radical le plus court, à savoir *volcan*.

Le dendrogramme de la Figure 5.2 présente un exemple de résultats obtenus à l'issue des deux premières étapes. La première étape de classification par le critère d'inclusion de mots permet la formation des familles suivantes :
- F_1 = [paléo-volcan, volcan]
- F_2 = [volcanique, paléovolcanique, néovolcanique, post-volcanique, subvolcanique]
- F_3 = [sub-volcaniques, volcaniques, subvolcaniques, post-volcaniques, postvolcaniques]
- F_4 = [paléovolcanisme, volcanisme]

Puis, au cours de l'étape 2 de fusion à partir des préfixes, la familles F_1 est tout d'abord fusionnée avec F_2. La famille résultante est ensuite groupée avec F_3. Enfin, F_4 est réunie avec la classe ainsi obtenue.

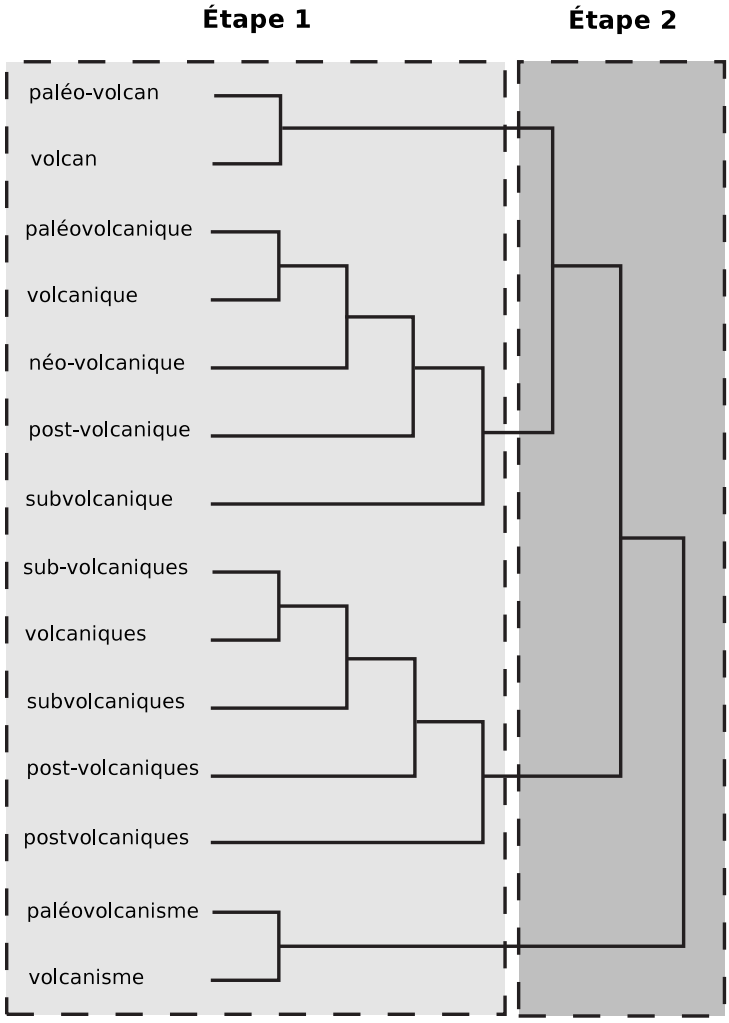


FIG. 5.2: Familles obtenues par classification à l'issue des deux premières étapes.

5.2.5 Étape 3 : regroupement de familles à partir des signatures

La dernière étape de la classification procède par **bootstrapping** [Heyer *et al.*, 2006]. Le bootstrapping consiste à découvrir de nouveaux éléments à partir d'un petit ensemble d'éléments initiaux. Cette technique est utilisée par [Enguehard, 1992] pour découvrir de nouveaux termes à partir d'une liste initiale de termes et par [Baroni et Bernardini, 2004] pour construire automatiquement des corpus de textes de spécialité à partir d'Internet. Dans notre cas, les données initiales correspondent à un ensemble de signatures fournies en entrée au système, identifiées par le module d'apprentissage d'affixes décrit en Section 4.2.1, p. 73. De nouvelles signatures sont découvertes à partir des regroupements opérés lors des étapes précédentes et permettent à leur tour d'effectuer de nouveaux regroupements. Le processus se termine lorsqu'il n'est plus possible de découvrir de nouvelles signatures.

Découverte de nouvelles signatures

La découverte de nouvelles signatures se fait à partir des familles déjà constituées. Les mots non préfixés de chaque famille sont comparés deux à deux afin d'obtenir une liste de signatures, selon la méthode suivante :

Soient :
m_1 et m_2 deux mots non préfixés appartenant à la famille F
tels que
$m_1 = \alpha + s_1$ et $m_2 = \alpha + s_2$
avec
$|\alpha| \geq 4$ et s_1 et s_2 des chaînes de caractères différant au moins par leur premier caractère.

Nous appellerons signature la paire (s_1,s_2) et $sig(F, F)$ l'ensemble des signatures formées à partir d'une famille F, c'est-à-dire par comparaison bijective des mots non préfixés de F. Toutes ces signatures sont ajoutées à la liste des signatures S.

Prenons l'exemple de la famille suivante, formée lors des étapes 1 et 2 :
[trachyandésite, andésite, trachy-andésite, andésites, trachy-andésites, trachyandésites, andésitique, trachy-andésitique, trachyandésitique, trachy-andésitiques, trachyandésitiques, andésitiques].

La comparaison des graphies des mots non préfixés de cette famille conduit à l'identification des signatures suivantes : (ϵ, s), $(e, ique)$, $(e, iques)$, $(es, ique)$ et $(es, iques)$ (voir Figure 5.3).

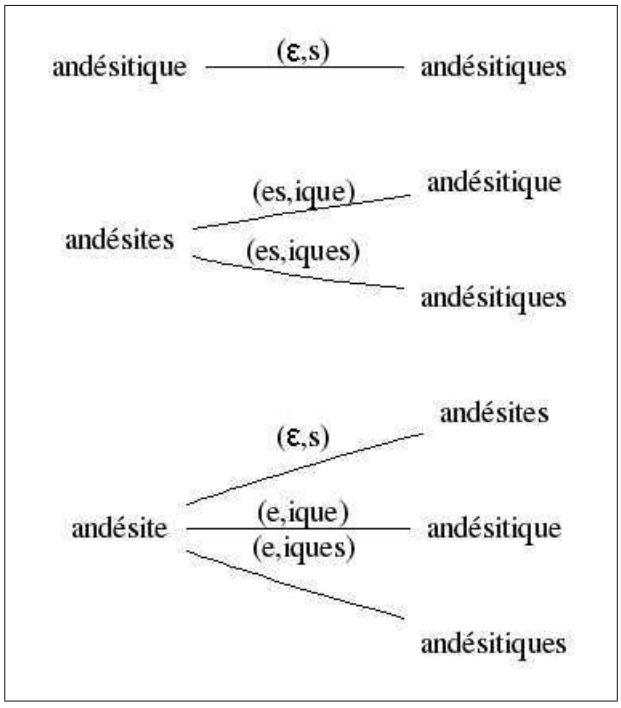


FIG. 5.3: Identification de signatures

Fusion de familles à l'aide des signatures

Les signatures ainsi acquises sont utilisées pour fusionner des familles. Le critère d'agglomération repose sur un indice p qui mesure la proportion de signatures valides partagées entre deux familles que l'on cherche à fusionner :

Soient :
- F_1 et F_2 deux familles ;
- l_1 le nombre de mots non préfixés de F_1 ;
- l_2 le nombre de mots non préfixés de F_2 ;
- S la liste de signatures fournies en entrée et découvertes à partir des familles déjà constituées.

L'indice p se calcule de la manière suivante :

$$p = \frac{|sig(F_1, F_2) \cap S|}{l_1 \cdot l_2}$$

Dans les expériences relatées dans la suite de ce chapitre, nous avons fusionné deux familles lorsque $p \geq 0.5$.

Prenons l'exemple des familles suivantes F_1 et F_2 et de la liste de signatures S :
- F_1 = [mesurent ; mesura ; mesuraient ; mesurait ; mesurant ; mesure]
- F_2 = [mesurer]
- S = [(ant, er) ; (ϵ, r) ; (nt, r) ; ...]

Les familles F_1 et F_2 sont fusionnées car :
- $sig(F_1, F_2)$ = [(nt, r) ; (a, er) ; (aient,er) ; (ait, er) ; (ant, er) ; (ϵ, r)],
  donc
  $sig(F_1, F_2) \cap S$ = [(ant, er) ; (ϵ, r) ; (nt, r)] et $|sig(F_1, F_2) \cap S| = 3$
- $l_1 = 6$ et $l_2 = 1$
  donc $l_1 \cdot l_2 = 6$
- $p = \frac{3}{6} = 0.5$.

Le dendrogramme de la Figure 5.4 illustre l'intérêt des regroupements à partir des signatures. La seule famille formée à l'issue des deux premières étapes est [océanique, subocéanique, subocéaniques, océaniques, non-océaniques]. L'étape de fusion de famille à partir des signatures partagées permet le regroupement de mots comme [océan, océans] ou [océanien, océanienne, océaniennes, océaniens].

Ce processus d'agglomération à partir des signatures partagées est répété tant que de nouvelles signatures sont acquises et tant que ces signatures permettent de regrouper des classes. La Figure 5.5 présente l'évolution du nombre de signatures en fonction du nombre d'itérations et du paramètre N utilisé pour l'acquisition des préfixes et paires de suffixes par le module présenté en Section 4.2.1, p. 73. Le nombre total de signatures en fonction du nombre d'itération suit une courbe de croissance logistique : le nombre de signatures différentes augmente fortement au cours des premières itérations, puis se stabilise. Le nombre de nouvelles signatures identifiées augmente jusqu'à la troisième itération environ, puis décroît. Le processus d'acquisition de nouvelles signatures s'achève au bout de 10 à 15 itérations. La courbe 5.5d présente un profil particulier en raison de la forte différence entre le nombre de signatures découvertes pour N=5 et N=10.

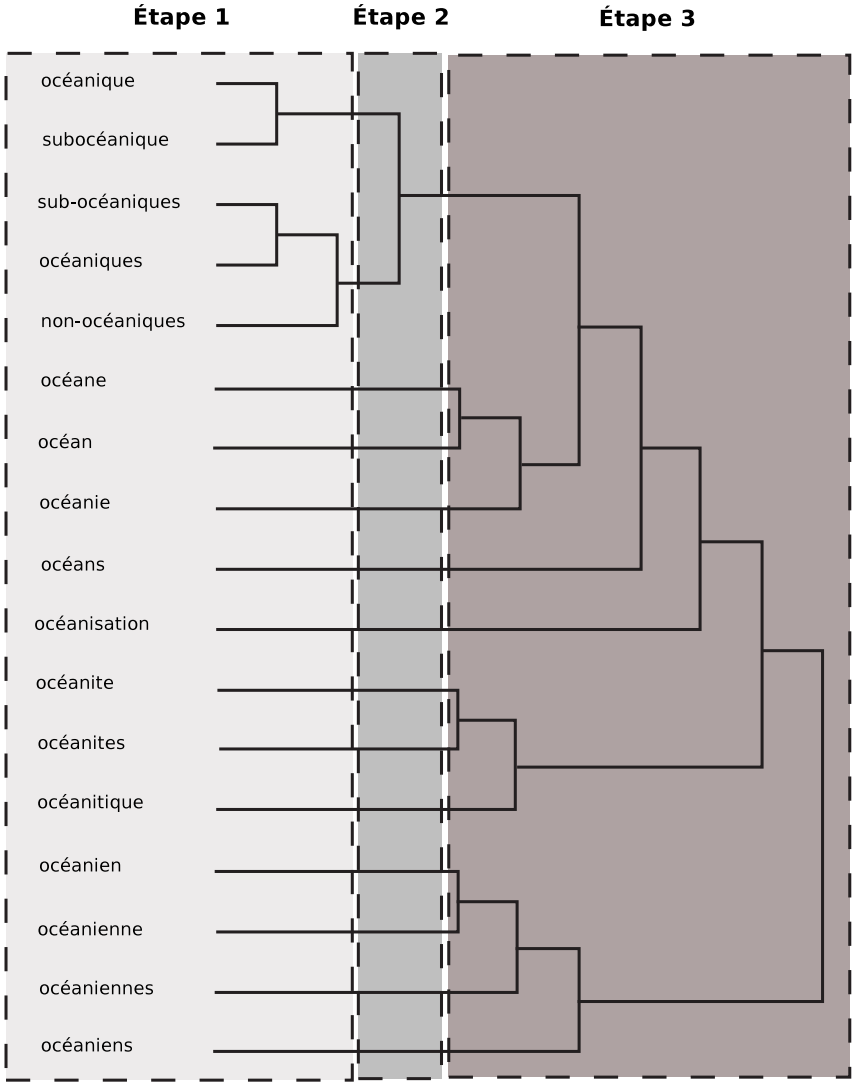
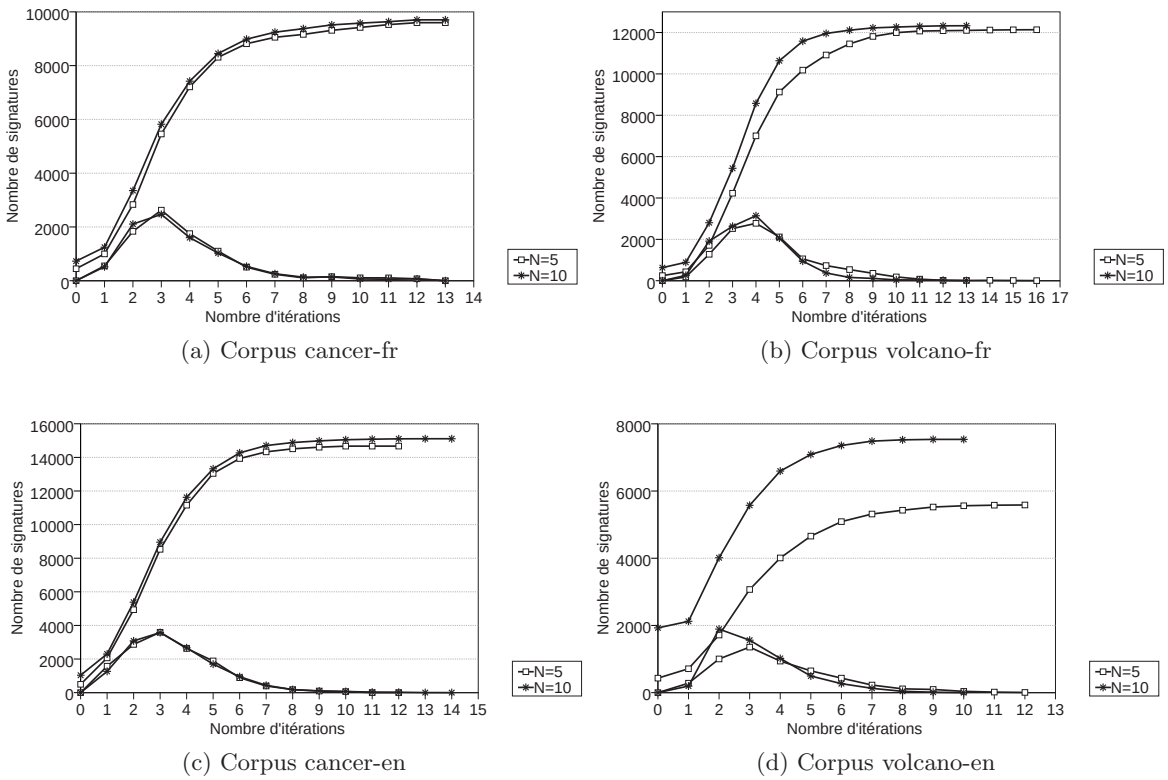


FIG. 5.4: Familles obtenues par classification à l'issue des trois étapes.



(a) Corpus cancer-fr

(b) Corpus volcano-fr

(c) Corpus cancer-en

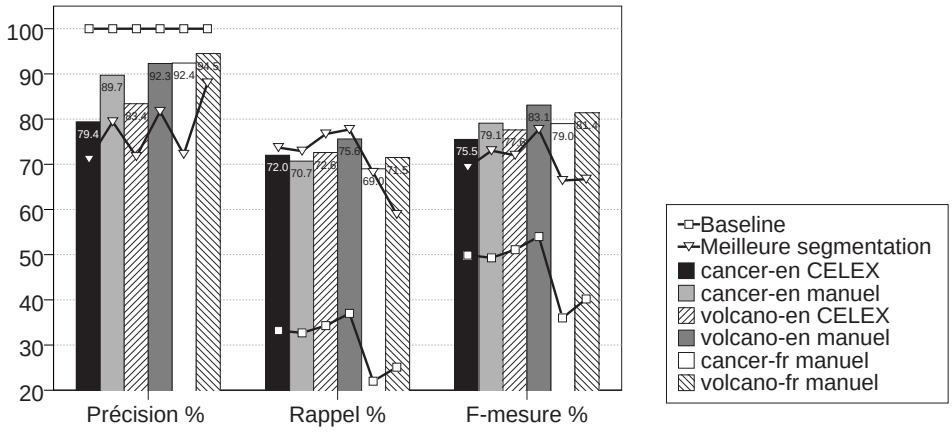
(d) Corpus volcano-en

FIG. 5.5: Évolution du nombre de signatures identifiées au cours du processus d'apprentissage pour les corpus : (a) cancer-fr ; (b) volcano-fr ; (c) cancer-en ; (d) volcano-en. La courbe du haut présente le nombre total de signatures tandis que la courbe du bas présente le nombre de nouvelles signatures.

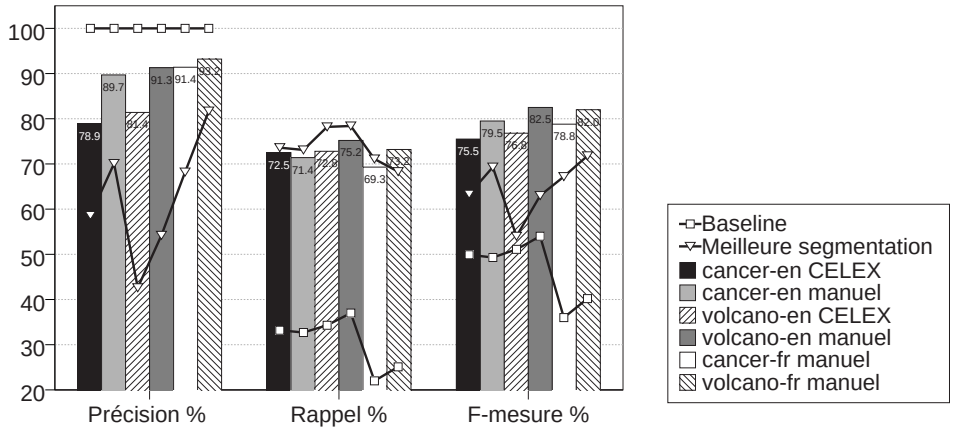
5.3 Évaluation

5.3.1 Résultats

Nous avons évalué les familles induites en utilisant la méthode décrite dans la Section 4.3.3, p. 91. Les résultats sont présentés sur la Figure 5.6. Les diagrammes présentent également les résultats de la méthode *baseline* (familles formées à partir de tous les mots et contenant un mot unique) ainsi que les meilleurs résultats obtenus par l'algorithme d'analyse morphologique par segmentation.



(a) N=5



(b) N=10

FIG. 5.6: Résultats obtenus par le système d'analyse morphologique par classification, pour différentes valeurs de N : (a) N = 5 ; (b) N = 10

Même si les résultats de la segmentation présentent parfois un rappel supérieur, la précision ainsi que la F-mesure sont toujours en faveur de l'analyse morphologique par classification. Celle-ci semble également moins sensible à la qualité des données initiales (préfixes et suffixes). En effet, les résultats de la segmentation sont fortement dégradés pour le corpus volcano-en avec N=10 en raison du très grand nombre d'affixes initiaux acquis. L'utilisation de ces mêmes listes pour la formation des familles morphologiques n'a aucun effet notable, que ce soit sur la précision ou le rappel. De plus, malgré les contraintes imposées sur la longueur minimale des

préfixes et des bases, le rappel est assez élevé et se situe autour de la barre des 70%. Ce résultat est d'autant plus remarquable car la méthode ne traite pas le cas des mots composés, si ce n'est de manière indirecte, si leurs composants sont identifiés comme étant des préfixes ou des suffixes.

Les résultats peuvent s'expliquer de différentes manières :

- La méthode d'évaluation mesure directement la qualité des familles morphologiques. Or l'objectif premier du système d'analyse par segmentation est le découpage des mots en sous-unités. Les familles ne sont qu'un dérivé de la segmentation. À l'inverse, le système d'analyse par classification forme directement des familles morphologiques, sans procéder auparavant à la segmentation des mots.
- Le système d'analyse par classification utilise la notion de signature, qui est absente de l'autre système. Les signatures permettent de conditionner l'identification d'un suffixe à sa co-occurrence avec un autre suffixe, dans une même série paradigmatique et donc d'augmenter la précision des résultats.
- Les objectifs du système d'analyse par classification sont bien plus modestes : ils ne visent pas à fournir une segmentation détaillée, ni un étiquetage des segments produits. Ceci explique également la plus grande simplicité de la méthode.

5.3.2 Analyse des résultats

Différents types de variantes sont groupés par l'algorithme :

- variantes graphiques et orthographiques comme *tumor* (variante américaine) et *tumour* (variante britannique).
- variantes flexionnelles comme *traitement* et *traitements*.
- variantes dérivationnelles suffixées comme *traiter* et *traitement* et préfixées comme *auto-examen* et *examen*.
- composés savants comme *hormonothérapie* et *immunothérapie*.

Il faut toutefois noter que les composés ne sont regroupés que si l'un de leurs éléments est identifié dans la liste des préfixes ou dans la liste des suffixes. De plus, la hiérarchie produite par la classification ne permet pas la parenté multiple : un composé est donc toujours associé avec un seul de ses composants.

Nous avions identifié divers problèmes de l'analyse morphologique par segmentation, dont l'absence de regroupement pour des mots dont les bases présentent des variations comme le doublement de consonne ou le changement d'accentuation. Ces problèmes sont résolus par l'algorithme de classification dans certains cas, si les conditions de regroupement le permettent :

- Doublement des consonnes en fin de radical : les signatures découvertes permettent de former la famille suivante : [dimensionnement ; dimension ; dimensionnées ; dimensions].
- Changement d'accentuation : les familles [pseudocratère ; intra-cratère ; cratère ; intracratère] et [intracratérique ; intra-cratérique ; cratérique] sont fusionnées grâce au partage du préfixe *intra+* et de radicaux suffisamment proches.

Malgré sa bonne précision, l'algorithme d'analyse par classification commet deux erreurs principales : sur-regroupement, c'est à dire le groupement de mots qui n'appartiennent pas tous à la même famille morphologique et sous-regroupement, c'est à dire l'absence de regroupement pour des mots appartenant à la même famille. La première erreur a pour conséquence de faire baisser la précision du système, tandis que la seconde conduit à une baisse du rappel.

Ces erreurs peuvent survenir à toutes les étapes de la classification :

- À l'étape 1, malgré les contraintes imposées sur la longueur des préfixes et des mots, des mots peuvent être injustement considérés comme étant formés par combinaison d'un préfixe et d'un suffixe. Ainsi, le mot *missing* est analysé comme étant la forme préfixée par *mis* du mot *sing*.
- À l'étape 2, il arrive que des familles soient fusionnées alors même qu'elles sont morphologiquement disjointes. Par exemple, la famille [médiane, paramédiane] est fusionnée avec la famille [socio-médical, paramédical, médical] car les mots *médiane* et *médical* commencent par la même chaîne de caractères de longueur 4 et apparaissent tous deux sous forme préfixée avec *para*.
- À l'étape 3 enfin, les contraintes imposées sur la longueur minimale de la base, égale à 4, peuvent empêcher le regroupement de certaines familles comme [île] et [îles], et donc induire une baisse du rappel. L'inverse est également vrai toutefois, comme en témoigne le regroupement des mots *rocket* et *rockets* avec *rock* et *rocks*.

5.4 Conclusion

Nous avons présenté une méthode d'analyse morphologique non supervisée procédant par classification. Malgré la simplicité de la méthode, les résultats obtenus sont très bons, notamment en ce qui concerne la précision. Même si les cas de composition ne sont pas explicitement traités, le rappel n'est pas trop dégradé par rapport à l'analyse par segmentation.

Des évaluations complémentaires sont toutefois nécessaires. Nous n'avons pour l'heure testé le système que sur du vocabulaire issu de corpus de spécialité en français et en anglais. Il serait intéressant d'évaluer les performances pour d'autres langues plus complexes, comme l'allemand, et pour du vocabulaire non technique. L'utilisation des préfixes aux deux premières étapes de la classification suppose que la langue traitée utilise ce procédé de formation. Or ce procédé n'est pas présent dans toutes les langues : le turc par exemple n'emploie pas de préfixes, ce qui rend les deux premières étapes du traitement inutiles. Reste alors à déterminer si le système est capable de produire des regroupements pertinents en utilisant uniquement les signatures. On peut se poser une question similaire pour le vocabulaire moins technique, où le procédé de préfixation est utilisé moins fréquemment. Des expérimentations complémentaires pourront nous permettre de répondre à ces questions.

Les améliorations possibles à apporter au système sont diverses. À l'heure actuelle, nous n'utilisons aucune autre information que la liste de mots. Les informations contextuelles, disponibles dans les corpus, pourraient permettre d'améliorer encore les résultats, notamment en terme de précision. Ces informations pourraient être intégrées à toutes les étapes de la classification, afin de valider le fusionnement de deux familles.

De plus, le système procède à une classification hiérarchique ascendante stricte, sans parenté multiple et donc sans possibilité pour un mot d'appartenir à deux voire à plusieurs familles différentes. Ceci est souhaitable pour les mots composés, qui font partie de plusieurs familles morphologiques. Il faudrait donc recourir à une forme de classification « floue ».

Le système ne permet pas non plus la découverte de nouveaux préfixes, en complément de ceux injectés dans le système lors de la phase d'initialisation. On pourrait envisager d'appliquer une phase de bootstrapping similaire à celle qui permet la découverte de nouvelles signatures.

Il reste également à résoudre le problème du traitement des mots absents du corpus utilisé pour l'apprentissage des familles morphologiques. La manière la plus simple de procéder semble

de ré-appliquer la dernière étape du système, basée sur les signatures, aux mots inconnus pour leur assigner une famille morphologique ou créer de nouvelles familles.

Enfin, tout comme on peut déduire des familles morphologiques à partir des résultats de la segmentation, il serait souhaitable de pouvoir décomposer les mots en sous-unités à partir de leur famille morphologique. En effet, les familles morphologiques sont utiles à des applications de recherche d'informations ou de classification de documents, mais pour d'autres tâches, comme la reconnaissance ou la synthèse de la parole, il est nécessaire de disposer d'un découpage des mots en segments.

Conclusion

Pour conclure cette partie consacrée à l'acquisition de connaissances morphologiques à partir de textes, nous allons ré-examiner les deux méthodes proposées en fonction des objectifs que nous nous sommes fixés :

Travail sur corpus Les seules données d'apprentissage utilisées par les deux méthodes sont des listes de mots issues de corpus construits automatiquement à partir d'Internet. Nous n'avons employé aucune ressource externe.

Langue de spécialité Le vocabulaire des domaines spécialisés, comme la médecine, présente souvent une structure morphologique complexe, faisant appel au procédé de composition. Ceci impose de disposer de systèmes d'analyse morphologique capables de traiter aussi bien de la flexion, que de la dérivation ou de la composition. Le système d'analyse morphologique par segmentation est capable de détecter plusieurs bases dans chaque mot et remplit ainsi cette mission. Le second système répond partiellement à ce besoin, mais devrait encore être amélioré pour permettre l'appartenance d'un mot à plusieurs familles morphologiques.

Apprentissage et approche statistique Les deux systèmes procèdent par apprentissage non supervisé. Il est toutefois difficile de se passer totalement de certains paramètres ou valeurs seuils qui garantissent un certain niveau de précision et de rappel aux résultats.

Indépendance aux langues Le système d'analyse morphologique par segmentation a été testé sur 5 langues : le français, l'anglais, le finnois, le turc et l'allemand. Ces langues sont assez différentes et présentent des degrés de complexité morphologique variés. Le système obtient toutefois d'assez bons résultats. Le second système n'a pour l'heure été testé que sur l'anglais et le français. Des évaluations pour d'autres langues sont envisageables, notamment l'allemand et le néerlandais qui font également partie de la base CELEX.

Les deux systèmes remplissent donc les objectifs fixés, avec de bons résultats, notamment en comparaison aux autres systèmes de même type existant à l'heure actuelle, comme le système Morfessor. Il reste à les intégrer dans un processus global d'acquisition de ressources lexico-sémantiques à partir de textes. C'est ce que nous allons présenter plus en détail dans la partie suivante.

Troisième partie

Exploitation des résultats

Introduction

Nous venons de décrire deux systèmes d'analyse morphologique. Le but de notre travail n'était pas uniquement de concevoir des algorithmes capables de segmenter les mots ou de les regrouper en fonction de leur similarité morphologique. Nous nous étions fixé pour objectif d'utiliser la morphologie pour des tâches d'acquisition automatique de ressources lexicales. Dans le premier chapitre, nous avons distingué deux tâches principales : l'identification des termes et mots clés du domaine et la structuration de ces termes par la découverte des relations sémantiques qu'ils entretiennent.

Nous allons montrer dans cette partie que la morphologie peut être employée pour la résolution de ces problèmes. Le Chapitre 6 décrit une application de pondération et de visualisation de mots clés basée sur les familles morphologiques identifiées par le système décrit dans le Chapitre 5. Puis, nous présentons une expérience d'acquisition automatique de relations sémantiques à partir des résultats de la segmentation morphologique produite par le système décrit dans le Chapitre 4.

Chapitre 6

Pondération et visualisation de mots clés

6.1 Introduction

Nous allons traiter dans ce chapitre de deux problèmes : (a) la pondération et l'acquisition de mots clés à partir de leur famille morphologique et (b) la visualisation des résultats sous forme de liste pondérée au format HTML.

6.1.1 Méthodes de pondération de mots clés

Les mots clés sont les mots qui décrivent le mieux le contenu d'un document ou d'un corpus. Cette propriété est généralement corrélée à un nombre d'occurrences important du mot dans le corpus ou le document considéré. Les méthodes classiques d'acquisition de mots clés reposent sur la comparaison des fréquences des mots dans le corpus d'analyse et dans un corpus de référence. Les fréquences comparées correspondent aux *fréquences de surface* des mots (pour reprendre le vocabulaire de la psychologie cognitive). Ces fréquences constituent un indice statistique pour l'identification des mots clés.

Les mots ne sont toutefois pas les seules unités linguistiques susceptibles de décrire le contenu d'un document ou d'un corpus. En effet, les mots ne sont pas des unités atomiques mais peuvent être décomposés en morphèmes. Selon [Witschel et Biemann, 2006], les constituants de mots sont de bons indicateurs du contenu d'un corpus et peuvent donc être utilisés pour la classification de documents. Nous avons vu également en Section 1.3.1, p. 18 que les affixes spécifiques à un domaine permettent l'identification de termes. La morphologie correspond donc à un indice structurel pour l'identification des mots clés.

Les systèmes que nous avons décrits dans les sections précédentes identifient les familles morphologiques d'un corpus. Nous proposons d'utiliser ces familles morphologiques pour extraire les mots clés d'un domaine, en identifiant les familles morphologiques spécifiques au domaine. Pour déterminer les familles spécifiques, nous utilisons la mesure de *fréquence cumulée* qui correspond à la somme de la fréquence de surface des mots de la famille [Meunier, 2003]. De nombreuses expériences ont démontré que la fréquence cumulée joue un rôle lors de la reconnaissance des mots, et cet effet est d'autant plus perceptible que la fréquence cumulée est importante. Les fréquences cumulées des familles dans le corpus d'analyse sont comparées aux fréquences cumulées de ces mêmes familles dans un corpus de référence. La méthode d'identification de mots clés que nous

proposons combine donc indices structurels, par l'appartenance à une famille morphologique, et indices statistiques, par la pondération des familles en fonction de leur fréquence cumulée.

6.1.2 Visualisation de données

En complément de l'identification des mots clés via leur famille morphologique, nous présentons un mode de visualisation des résultats, utilisable pour cartographier les mots clés et donc les thématiques propres à un corpus ou à tout ensemble de documents.

Les méthodes de visualisation font à l'heure actuelle partie intégrante de tout système traitant d'un grand ensemble de données, ceci afin de faciliter leur analyse et de ne pas noyer l'utilisateur sous le flot des informations. On peut distinguer deux types de représentations pour l'analyse des données textuelles : les graphes et les cartes.

Les graphes sont adaptés à la visualisation de données structurées par des relations. Ces relations peuvent être de divers types :

- Liens contextuels, représentés sous forme de graphes de mots co-occurrents [Heyer *et al.*, 2001, Véronis, 2003] ;
- Liens sémantiques, représentés sous forme de graphes de concepts [Le Priol, 2001] ;
- Liens thématiques, représentés sous forme de graphes des thèmes spécifiques au domaine [Ibekwe-Sanjuan et Sanjuan, 2004].

Les cartes permettent quant à elles de représenter des mesures de distance entre données : la proximité sur la carte est proportionnelle à la similarité des éléments. Les cartes présentent généralement le résultat d'une classification ou d'une projection des données sur un espace à deux dimensions, par les méthodes suivantes :

- Cartes auto-organisatrices de Kohonen appliquées aux mots [Honkela *et al.*, 1995, Lagus *et al.*, 2002] ou aux documents [Kohonen *et al.*, 2000] ;
- Analyse en composantes principales [Ploux et Victorri, 1998].

Le mode de visualisation que nous utilisons se rapproche des cartes, car les données sont représentées dans un espace à deux dimensions. Toutefois, la distance n'y est pas significative, on parle donc plutôt de *liste pondérée*, car la taille et la couleur d'un élément dépendent de son poids.

Nous allons dans un premier temps définir les mesures utilisées pour la pondération des familles morphologiques en fonction de la fréquence cumulée. Puis, nous présenterons un mode de visualisation des données, consistant à présenter les familles sous forme de liste pondérée au format HTML. Enfin, nous procéderons à une évaluation des résultats obtenus en fonction de glossaires décrivant les termes des domaines considérés.

6.2 Pondération des familles morphologiques

Une famille morphologique est constituée par un ensemble de mots partageant un même radical. Chaque mot de la famille peut être caractérisé par son nombre d'occurrences dans le corpus d'analyse. Nous allons calculer deux mesures différentes à partir de ce nombre d'occurrences : la fréquence cumulée de la famille (CFF = Cumulative Family Frequency) et le log du rapport de vraisemblance (LLR = Log Likelihood Ratio). Grâce à ces mesures, il sera possible de classer les familles morphologiques par leur importance et donc de pondérer les mots qu'elles contiennent : les mots-clés du domaine appartiennent aux familles morphologiques les plus importantes.

6.2.1 Fréquence cumulée

La première mesure correspond à la fréquence cumulée (CFF) des mots de la famille morphologique. Elle se calcule de la manière suivante :

Soient :
- F une famille morphologique composée des mots $m_1, m_2, \ldots, m_n$;
- $f(m_i)$ le nombre d'occurrences du mot m_i dans le corpus analysé.

$$CFF = \sum_{i=1}^{n} f(m_i)$$

6.2.2 Log du rapport de vraisemblance

Le log du rapport de vraisemblance (LLR) est notamment utilisé pour comparer le nombre d'occurrences de mots dans un corpus de spécialité par rapport à un corpus de référence. En effet, les mots clés spécifiques au domaine spécialisé apparaissent de manière significativement plus fréquente dans le corpus de spécialité que dans le corpus de langue générale. Le log du rapport de vraisemblance permet d'estimer la significativité statistique de la différence des occurrences des mots dans le corpus de spécialité et dans le corpus de référence. La formule du LLR est donnée page 23. Nous avons adapté cette mesure à la comparaison des fréquences cumulées d'une famille morphologique dans le corpus analysé et dans un corpus de référence aux fréquences cumulées attendues selon l'hypothèse nulle. Le calcul de cette mesure repose sur la table de contingence 6.1.

	Corpus 1	Corpus 2	Total
Fréquence cumulée de la famille	a	b	a+b
Fréquence cumulée des autres familles	c-a	d-b	c+d-a-b
Total	c	d	c+d

TAB. 6.1: Table de contingence pour la comparaison des fréquences cumulées des familles morphologiques entre corpus.

La formule du log du rapport vraisemblance pour la famille F, selon la formule donnée par [Rayson et Garside, 2000] est la suivante :

$$LLR = 2\left(a\ \ln\left(\frac{a}{E_1}\right) + b\ \ln\left(\frac{b}{E_2}\right)\right)$$

La mesure du LLR permet de comparer les fréquences cumulées observées de la famille morphologique dans chacun des corpus : $O_1 = a$ (Corpus 1) et $O_2 = b$ (Corpus 2) aux effectifs attendus selon l'hypothèse d'indépendance : $E_1 = c \cdot \frac{a+b}{c+d}$ (Corpus 1) et $E_2 = d \cdot \frac{a+b}{c+d}$ (Corpus 2) où c est le nombre d'occurrences de mots dans le Corpus 1 et d le nombre d'occurrences de mots dans le Corpus 2[1].

[1]Les mots du corpus de référence qui n'apparaissent pas dans le corpus utilisé pour l'acquisition des familles constituent des familles dont ils sont les uniques membres.

6.3 Visualisation des mots clés

Une fois les familles pondérées par la fréquence cumulée CFF ou le rapport du log de vraisemblance LLR, se pose la question de leur présentation. Les résultats de la pondération des mots clés sont généralement présentés sous formes de liste ordonnées, par ordre de mesure décroissante. Ce mode de visualisation rend l'analyse des données assez fastidieuse, car il faut parcourir l'ensemble de la liste, jusqu'à un seuil limite de la mesure en-deçà duquel les mots clés ne sont plus considérés comme pertinents. Nous proposons donc une présentation différente, sous forme de liste pondérée au format HTML. Dans une telle liste, la taille de la police et la couleur utilisées sont dépendantes du poids de l'élément représenté. Plus un élément est important, plus la taille de la police est grande et plus la couleur est foncée. À l'inverse, moins l'élément est important, plus la taille de la police est petite et plus la couleur est claire.

Ce mode de visualisation, basé sur le principe des « cartes de chaleur » (*heatmap*), est utilisé par de nombreux sites Web pour présenter les mots clés ou étiquettes (*tags*) associés à divers types de ressources. Dans ce cas, les listes pondérées sont appelées « nuages d'étiquettes » (*tag clouds*). Les *tags* sont des mots ou des expressions permettant de décrire une ressource (photo, page web, flux RSS, etc.) sur Internet et qui constituent ainsi des méta-données pour la ressource décrite [Wikipedia, 2006b]. Les *tags* peuvent notamment être utilisés pour classer les ressources décrites ou effectuer des recherches [1].

Le principe des « nuages d'étiquettes » est utilisable pour la visualisation de toute liste de mots, à partir du moment où l'on dispose d'une mesure de pondération pour ces mots. Par exemple, [Eiken *et al.*, 2006] présentent les mots clés extraits quotidiennement de sites d'actualités sous forme de nuages de mots clés. De manière similaire, J. Véronis, sur son blog « Technologies du Langage » (`http://aixtal.blogspot.com/`), utilise les nuages de mots pour représenter les mots clés de la presse ou les résultats de requêtes effectuées sur un moteur de recherche. Son outil, le Nébuloscope [Véronis, 2006], génère des nuages de mots à partir des résultats de requêtes sur le moteur Dir.com. La Figure 6.1 présente un tel nuage de mots généré à partir de la requête « volcan ».

L'avantage majeur de ce type de représentations est qu'elle permet d'identifier rapidement les mots clés les plus importants. De plus, elle diffère des représentations des listes classiques triées par ordre de fréquence décroissante car les mots clés sont distribués sur une carte à deux dimensions.

6.3.1 Listes pondérées de familles de mots

Compte tenu des avantages des listes pondérées, nous avons choisi ce mode de représentation pour les familles de mots. La visualisation des familles plutôt que celle des mots clés conduit à une réduction de la taille de la liste et évite les redondances. En effet, la Figure 6.1, obtenue directement à partir des mots sans autre traitement, présente un certain nombre de répétitions qui pourraient être évitées grâce à une regroupement préalable des mots en familles, comme par exemple [eruption, eruptions], [hotel, hotels], [ile, iles] ou [volcan, volcanique, volcans]. On retrouve là l'idée selon laquelle la morphologie permet une *compression* des données textuelles, idée qui est exploitée par les algorithmes d'apprentissage de connaissances morphologiques basés sur le principe de la longueur minimale de description (voir Section 2.3.3, p. 48). Par conséquent, au lieu de représenter directement les mots clés sur la liste, nous y projetons les familles morphologiques, représentées par un mot clé typique permettant l'accès à l'ensemble des éléments de la famille. Ce mot clé typique peut être sélectionné en fonction de sa forme : il pourra alors s'agir

[1] Voir par exemple les sites Del.icio.us `http://del.icio.us/` ou Flickr `http://www.flickr.com/`

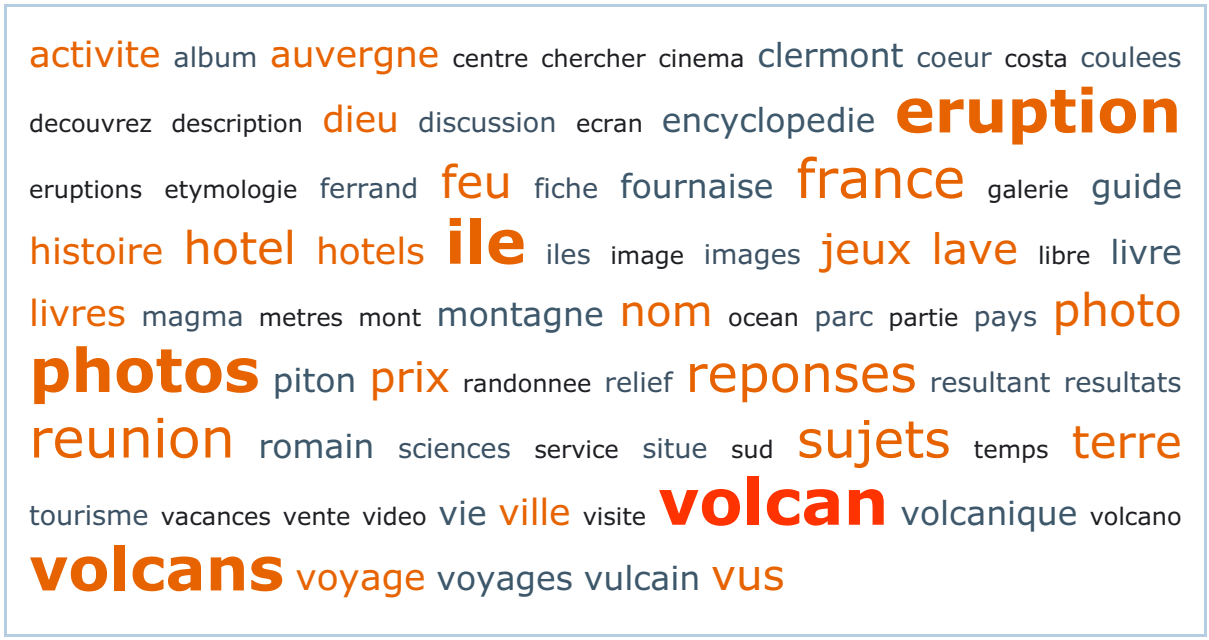


FIG. 6.1: Exemple de nuage de mots produit par le Nébuloscope de J. Véronis à partir de la requête « volcan ».

du radical de la famille. Nos expériences ont toutefois montré que les formes les plus fréquentes, et donc les plus saillantes dans le corpus, ne sont pas toujours les radicaux de la famille. Nous avons donc opté pour une autre solution : le mot clé prototypique est le mot le plus fréquent de la famille.

Nous allons maintenant expliciter le principe de construction des listes pondérées au format HTML.

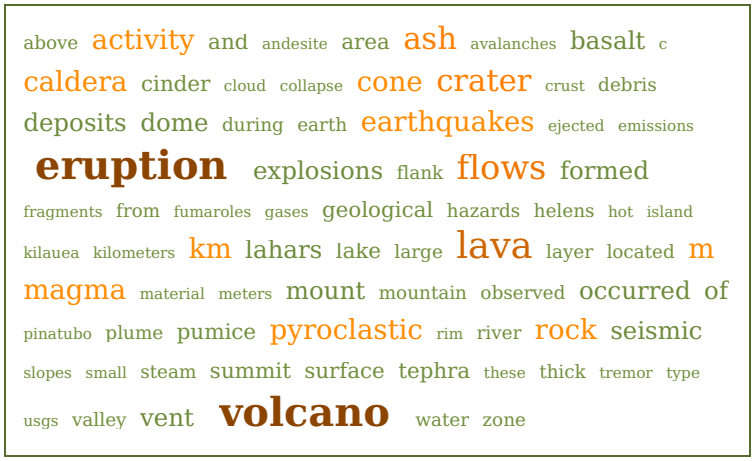
6.3.2 Construction d'une liste pondérée de familles de mots

Le principe des listes pondérées au format HTML est très simple à mettre en œuvre : il consiste à représenter les données par une police de taille et de couleur différente en fonction d'un poids qui est attribué à chaque élément. Les tailles de police et les couleurs sont attribuées par niveau dans la liste : les niveaux les plus bas, de poids faible, correspondent à une couleur claire et à une taille de police petite tandis que les niveaux les plus élevés, de poids élevé, ont une couleur foncée et une police de grande taille. Le nombre de niveaux de la liste est déterminé à l'avance : dans nos expériences, nous produisons des cartes à 9 niveaux.

Le processus de construction d'une liste pondérée de familles de mots de L niveaux et M éléments nécessite la donnée d'un ensemble de familles associées à un poids (CFF ou LLR). Seules les M familles qui ont le poids le plus élevé sont conservées. Parmi ces M familles, on recherche le poids maximal max et le poids minimal min. Ces deux valeurs sont utilisées pour calculer les intervalles de valeurs de poids associés à chaque niveau. L'ensemble des valeurs de poids situées entre min et max sont découpées en L intervalles où un intervalle correspond à $(\log(max) - \log(min))/L$. Le niveau d'une famille morphologique est alors calculé en divisant le logarithme de son poids moins le logarithme du poids minimal par la valeur de l'intervalle.

La Figure 6.2 présente un exemple de résultat, pour L=9 (9 niveaux) et M=80 (80 familles sur la carte). La liste ne présente que les représentants de chaque famille. L'accès au détail de

la famille se fait en cliquant sur le mot clé représentant la famille sur la carte : tous les éléments de la famille sont alors affichés à droite de la liste. Ici, les détails sont donnés pour la famille de mots représentée par « andesite ».



andesictic	1
andesine	5
andesit	1
andesite	1336
andesite-based	1
andesite-dacite	15
andesites	132
andesitic	581
andesiting	1
basalt-andesite	4
basalt-andesite-dacite	3
dacitic-andesitic	4
	2084

FIG. 6.2: Exemple de nuage de familles de mots pour le corpus volcano-en.

6.4 Évaluation

Il reste maintenant à déterminer une méthodologie pour évaluer les résultats de l'identification et la visualisation des mots clés via leur famille morphologique.

Nous avons présenté dans les chapitres précédents deux systèmes d'apprentissage de données morphologiques non supervisés. Les deux systèmes permettent l'acquisition de familles morphologiques. Les résultats du système d'analyse morphologique par classification sont toutefois meilleurs et nous utiliserons donc ses résultats pour les expérimentations présentées dans ce chapitre (résultats obtenus pour N=5).

Nous avons évalué et comparé les listes pondérées produites à partir de quatre méthodes différentes :

- Baseline-CFF : mots pondérés par leur nombre d'occurrences (fréquence de surface).
- Baseline-LLR : mots pondérés par le log du rapport de vraisemblance.
- Familles-CFF : familles morphologiques obtenues par classification, pondérées par la fréquence cumuluée.
- Familles-LLR : familles morphologiques obtenues par classification, pondérées par le log du rapport de vraisemblance.

Les méthodes *baseline* correspondent au cas où aucun regroupement morphologique n'a été opéré : chaque mot forme alors sa propre famille morphologique et la mesure du CFF est égale au nombre d'occurrences du mot (fréquence de surface).

Les mesures du log du rapport de vraisemblance ont été obtenues par comparaison aux corpus de référence issus de la collection de corpus de l'Université de Leipzig décrits dans l'Annexe A.2, page 151.

Afin d'évaluer les listes pondérées ainsi produites ainsi que la pertinence des mots clés représentés via leur famille, nous avons comparé les éléments représentés aux vocables définis dans divers glossaires disponibles sur Internet, que nous appellerons glossaires de référence.

6.4.1 Glossaires de référence

Pour chaque corpus (cancer-fr, cancer-en, volcano-fr, volcano-en) nous avons manuellement sélectionné dix glossaires différents sur Internet. Les sites d'origine de ces glossaires sont listés en Annexe C, p. 159. Nous avons extrait les items définis dans chaque glossaire afin de constituer des listes de référence. De plus, nous avons découpé les items polylexicaux de ces glossaires en vocables en prenant soin d'éliminer les mots outils.

6.4.2 Mesures d'évaluation

Nous avons défini deux mesures permettant d'évaluer les listes pondérées produites : densité et couverture. La mesure de densité rend compte du taux d'éléments intéressants présentés dans la liste. Un élément de la liste est considéré comme étant d'intérêt si la famille qu'il représente contient au moins un vocable défini dans les glossaires de référence. Plus le nombre d'éléments d'intérêt est important, plus la liste est dense en informations pertinentes. La mesure de couverture est quant à elle proche des mesures de rappel classiques. Elle permet de mesurer le nombre de vocables issus des glossaires de référence apparaissant directement ou indirectement, via leur famille, dans la liste pondérée[1].

Les mesures de densité et de couverture se calculent de la manière suivante :

Soient :
- $F_1, F_2, \ldots, F_M$ les familles présentes dans la liste pondérée évaluée de taille M ;
- $|F_i|$ le nombre de mots dans la famille F_i.
- $V = v_1, v_2, \ldots, v_n$ les vocables issus des glossaires utilisés pour l'évaluation.

$$\text{couverture} = \frac{1}{n} \sum_{i=1}^{M} |F_i \cap V|$$

$$\text{densité} = \frac{1}{M} \sum_{i=1}^{M} g(i) \quad \text{où} \quad g(i) = \begin{cases} 1 & \text{si } F_i \cap V \neq \emptyset \\ 0 & \text{autrement} \end{cases}$$

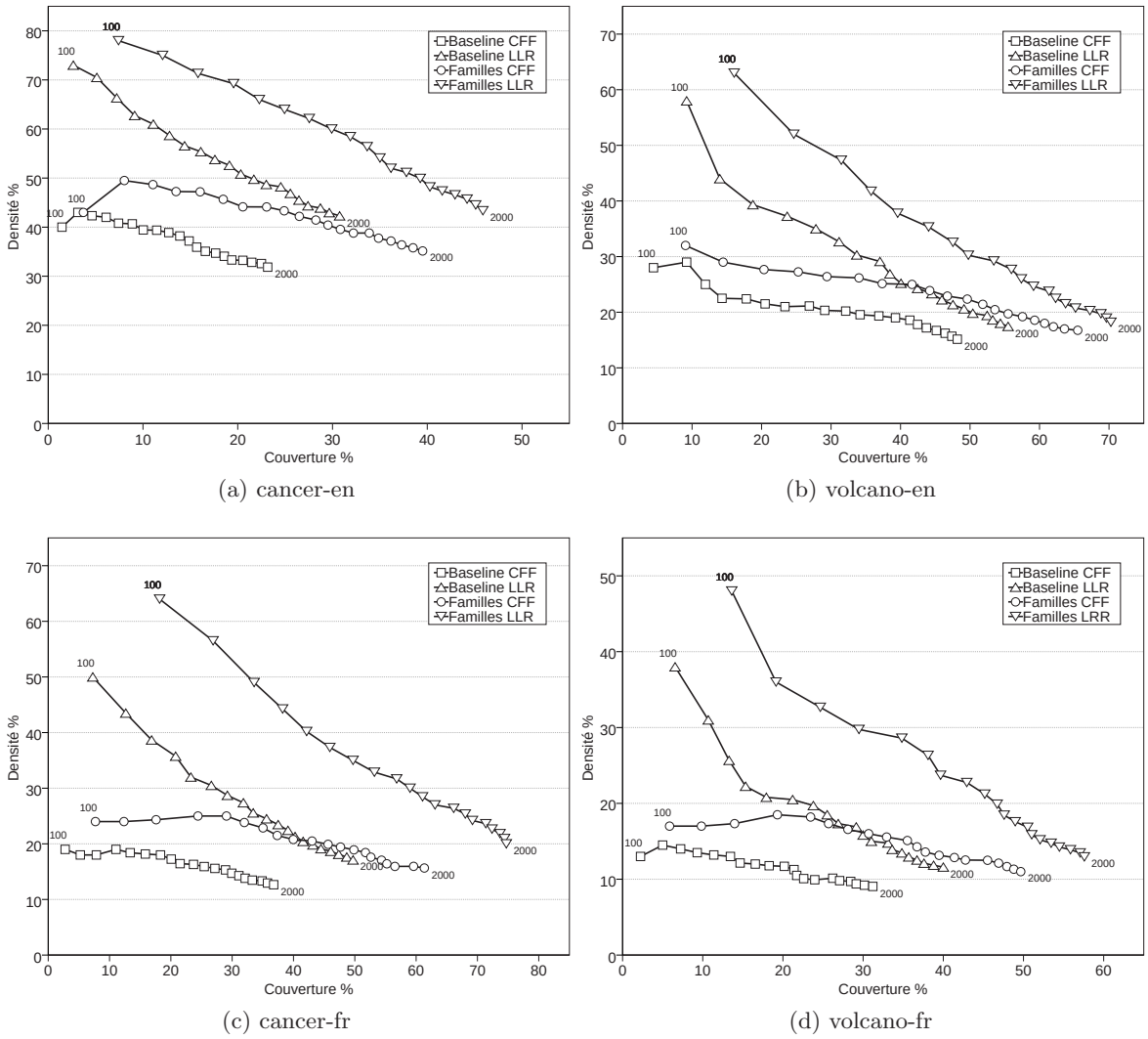
6.4.3 Résultats

Nous allons dans un premier temps analyser les résultats des mesures de densité et de couverture avant de comparer les représentations obtenues par les diverses méthodes.

Densité et couverture

Nous avons mesuré la densité et la couverture pour nos 4 corpus et des listes pondérées comprenant de 100 à 2 000 éléments. Il faut noter qu'une couverture de 100 % est impossible à atteindre car tous les vocables définis dans les glossaires de référence ne se trouvent pas forcément dans les corpus.

[1] Des mesures similaires sont définies par [Ninova *et al.*, 2005] pour évaluer l'adéquation d'une ressource lexicale à un corpus. La couverture mesure alors la proportion des mots du corpus apparaissant également dans la ressource et la densité compare la fréquence moyenne des items de la ressource à celle des mots du corpus.



(a) cancer-en (b) volcano-en

(c) cancer-fr (d) volcano-fr

FIG. 6.3: Résultats de l'évaluation des listes pondérées de familles de mots pour les différents corpus et différentes tailles M (variant de 100 à 2 000, par intervalle de 100 éléments).

Les couvertures maximales (en pourcentage) pour les différents corpus sont les suivantes :
- cancer-en : 84.6 % (2 327 / 2 751)
- volcano-en : 96.7 % (608 / 629)
- cancer-fr : 90.8 % (625 / 688)
- volcano-fr : 86.7 % (503 / 580)

La Figure 6.3 présente les résultats obtenus.

Quel que soit le corpus et le nombre d'éléments de la liste, les meilleurs résultats sont obtenus par la méthode Familles-LLR, c'est-à-dire pour les familles morphologiques pondérées par la mesure du log du rapport de vraisemblance. De plus, l'évolution de la densité en fonction de la couverture dépend de la méthode de pondération utilisée. La pondération par le LLR permet d'obtenir une densité élevée lorsque le nombre d'éléments évalués est faible. Cependant, cette densité chute fortement à mesure que la couverture augmente. Ainsi, alors que la courbe Baseline-

LLR se situe au-dessus de la courbe Familles-CFF, elle passe en-dessous pour des valeurs de couverture comprises entre 30 et 40 %.

Listes pondérées obtenues

Les Figures 6.4 à 6.7 présentent les listes pondérées obtenues pour les différentes méthodes à partir du corpus volcano-en, avec $M = 80$. L'analyse visuelle de ces listes est conforme aux résultats donnés par les mesures de précision et de couverture.

La pondération par le LLR est bien supérieure à la pondération par le CFF, ce qui est le résultat logiquement attendu. Les listes Baseline-CFF et Familles-CFF présentent surtout des mots outils, qui sont les mots les plus fréquents du corpus (voir Figures 6.4 et 6.6). On trouve également certains termes du domaine comme *crater*, *eruption*, *flows*, *lava* ou *volcano*. Ces mots sont toutefois peu visibles car ils appartiennent aux niveaux moyens de la liste, et sont donc représentés par une police de taille réduite. L'utilisation de familles morphologiques améliore légèrement le rendu global car certains termes apparaissent à des niveaux plus élevés comme par exemple *volcano* ou *eruption*.

Les meilleures représentations des données sont obtenues par la pondération du LLR. La liste Familles-LLR présente toutefois une meilleure couverture des différents termes que la liste Baseline-LLR. En effet, le regroupement des mots en familles morphologiques évite les informations redondantes et permet donc de présenter une plus grande diversité d'informations pertinentes dans une liste de même taille. Cette méthode rend donc possible une compression des données. On trouve par exemple trois éléments distincts appartenant à la même famille que *volcano* dans la liste Baseline-LLR (Figure 6.5), à savoir *volcanic*, *volcano* et *volcanoes*. Ces trois éléments sont regroupés dans la même famille représentée par le mot *volcano* dans la liste Familles-LLR (Figure 6.7).

Nous avons utilisé la mesure du LLR pour comparer les fréquences observées dans nos corpus spécialisés aux fréquences dans les corpus de références issus de la collection de corpus de l'Université de Leipzig. Nous avons constaté que la comparaison des corpus en français conduisait à l'identification de nombreux mots outils anglais comme *and*, *of*, *the* ou *to*. Ces mots apparaissent souvent dans des références bibliographiques en anglais qui sont fréquentes dans les corpus techniques. Même si nous avons utilisé un module permettant de filtrer les mots apparaissant dans des paragraphes en langue étrangère, certains mots anglais sont toutefois considérés comme valides car ils apparaissent au sein de paragraphes majoritairement écrits en français. Il reste donc de nombreux mots anglais dans les listes de mots que nous utilisons. Par contre le corpus de référence français ne contient que des phrases en français. La comparaison des fréquences peut donc également mettre en évidence certains éléments liés au mode de construction des corpus comparés, s'ils n'ont pas été construits de la même manière ou par les mêmes outils. La sélection du corpus de référence joue donc un rôle essentiel sur les résultats obtenus.

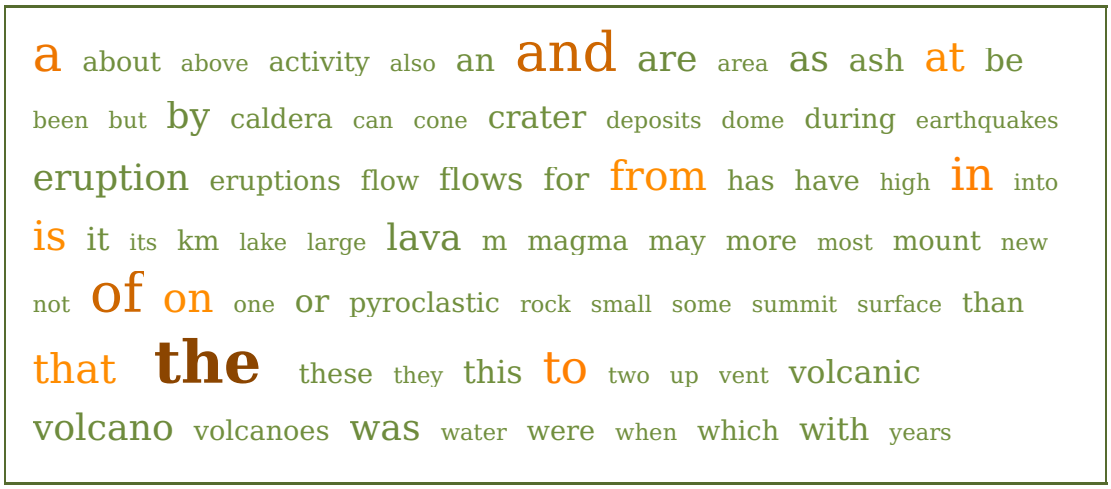


FIG. 6.4: Nuage Baseline-CFF à 80 éléments pour le corpus volcano-en

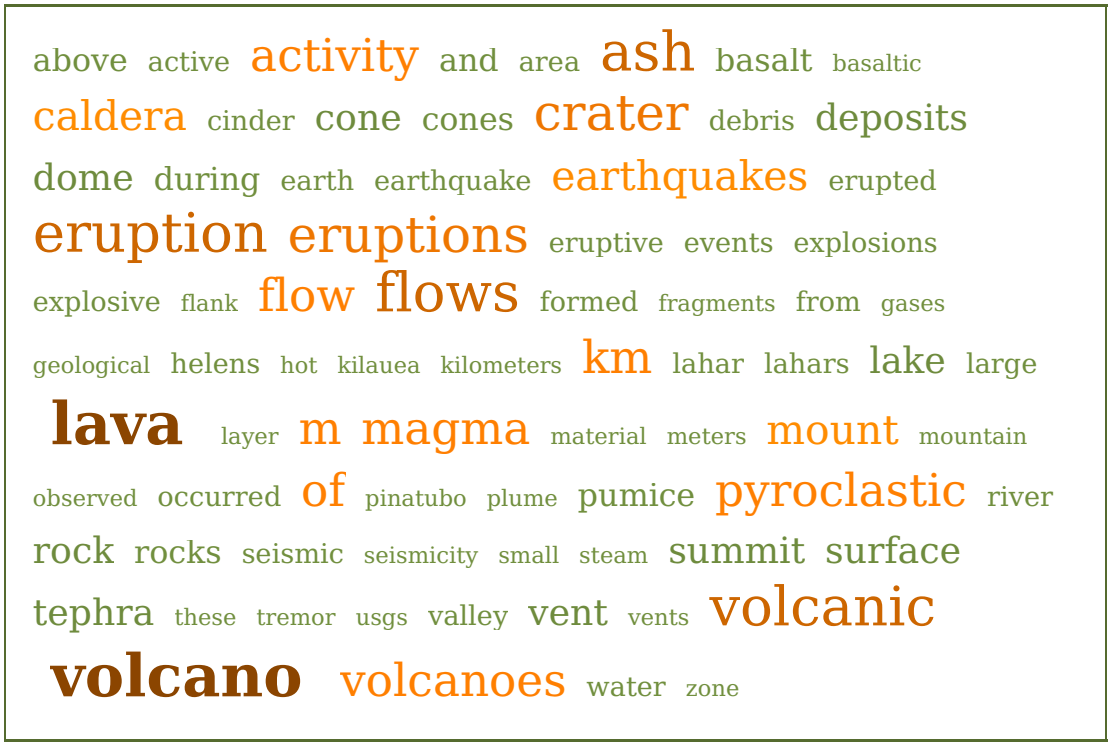


FIG. 6.5: Nuage Baseline-LLR à 80 éléments pour le corpus volcano-en

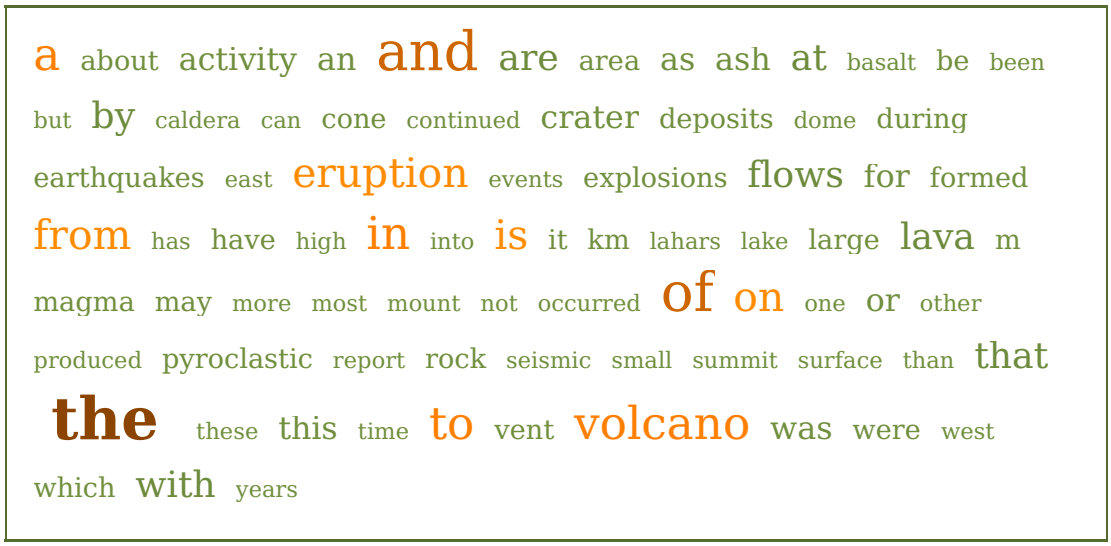


FIG. 6.6: Nuage Familles-CFF à 80 éléments pour le corpus volcano-en

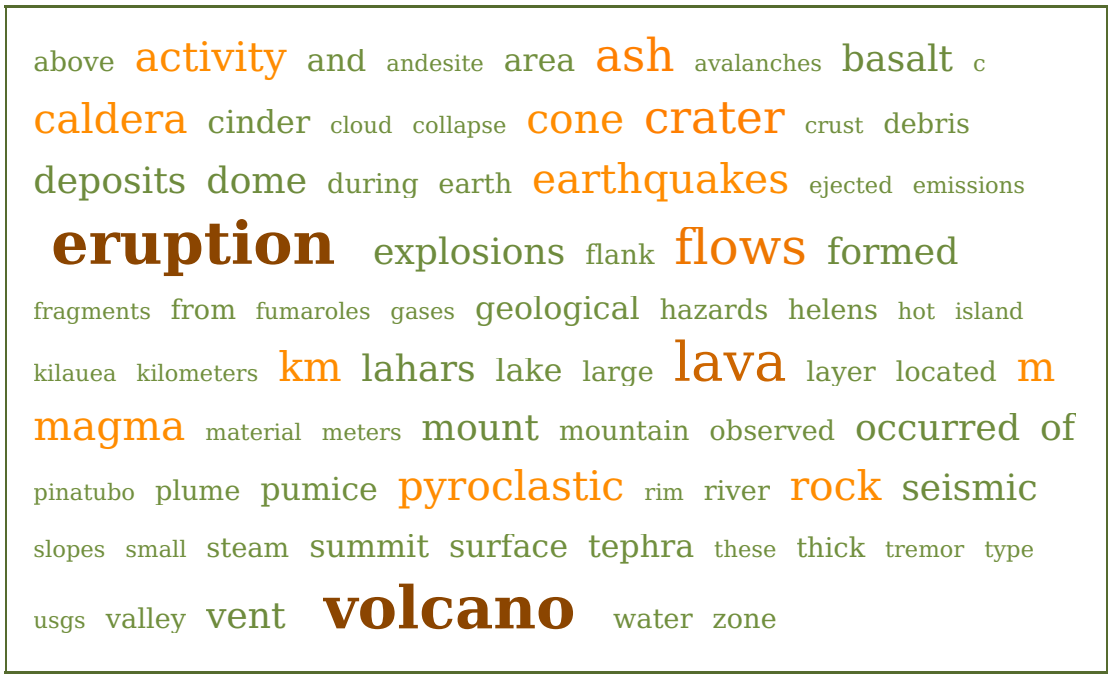


FIG. 6.7: Nuage Familles-LLR à 80 éléments pour le corpus volcano-en

6.5 Conclusion

Nous avons présenté une méthode de pondération de mots clés basée sur leur famille morphologique d'appartenance. Cette méthode est supérieure aux méthodes classiques d'acquisition de mots clés car elle combine l'utilisation d'indices structurels à des indices statistiques. De plus, elle facilite la visualisation des données en permettant leur représentation sans redondance sous forme de liste pondérée au format HTML. Ce mode de visualisation est largement utilisé sur Internet à l'heure actuelle car il met en évidence les thématiques majeures des ressources décrites.

La méthode n'est pas limitée en principe aux corpus de textes de spécialité et pourrait donc être utilisée pour représenter tout type de listes de mots pondérés issues d'autres corpus. Des expériences supplémentaires, notamment pour des corpus moins techniques, sont toutefois nécessaires pour évaluer cette possibilité.

De plus, nous n'avons utilisé qu'une seule mesure permettant de quantifier la spécificité d'une famille morphologique, celle du log du rapport de vraisemblance. Il pourrait être utile de comparer les résultats obtenus avec ceux d'autres mesures, comme par exemple le calcul des spécificités par approximation normale du calcul hypergéométrique [Lebart et Salem, 1994].

Dans le chapitre suivant, nous allons décrire une autre application des connaissances morphologiques pour l'acquisition de ressources lexicales, à savoir l'identification de relations sémantiques.

Chapitre 7

Acquisition de relations sémantiques

7.1 Introduction

Les familles morphologiques, acquises par segmentation ou par classification, groupent des mots qui sont non seulement liés par leur structure morphologique mais également par leur sens. Ces relations de sens sont diverses : les variantes flexionnelles ont un sens très proche tandis que les variantes dérivationnelles et les mots composés ont un sens plus éloigné. Afin d'analyser plus précisément ces liens sémantiques et de les quantifier, nous avons défini des schémas spécifiques reposant sur la décomposition morphologique des mots et l'étiquetage des segments qu'ils contiennent [Bernhard, 2006a]. Ces schémas sont décrits dans la Section 7.2. Après identification de ces schémas à partir des segmentations produites pour le corpus cancer-en, nous avons comptabilisé les relations sémantiques reliant effectivement les mots morphologiquement liés dans deux ressources différentes. La première, WordNet, est une ressource généraliste tandis que la seconde, le thésaurus du National Cancer Institute (NCIT), décrit plus spécifiquement la terminologie médicale dans le domaine du cancer.

7.2 Définition de schémas à partir des résultats de la segmentation morphologique

Grâce à l'étiquetage des segments proposés par le système d'analyse morphologique par segmentation, il est possible d'identifier dans le mot une base qui sera considérée comme la tête du mot : il s'agit de la base qui se situe en fin de mot, à droite. Les préfixes et bases qui précèdent éventuellement la tête sont des modifieurs. Cet ordre est valable en anglais, aussi bien pour les composés populaires comme *windmill* (moulin à vent) que pour les composés savants comme *phototherapy* [Bauer, 1998]. En français même si la tête des composés savants se situe à droite, la tête des composés populaires se trouve généralement à gauche. En effet, un *oiseau-mouche* est bien un oiseau et non une mouche. La méthode que nous allons décrire pour l'anglais n'est donc pas directement transposable à l'ensemble des mots composés du français, exception faite des composés savants.

Nous avons défini deux schémas à partir des segments morphologiques étiquetés : inclusion et substitution. Le premier devrait permettre l'identification de relations de spécialisation, tandis que le second devrait correspondre à des relations de co-hyponymie. L'objectif est donc non seulement d'identifier des relations sémantiques, mais également d'être capable de les étiqueter avec précision. Les schémas morphologiques sont inspirés des relations structurelles définies pour les termes polylexicaux, telles que l'inclusion lexicale (expansion gauche, expansion droite et

inclusion) et la substitution (voir Section 1.4.4, p. 30). Les études ont montré que l'insertion et l'expansion gauche mettent en évidence des relations de spécialisation tandis que la substitution permet d'identifier des termes co-hyponymes [Ibekwe-SanJuan, 1998, Ibekwe-SanJuan, 2005].

Nous avons adapté la définition des relations structurelles d'inclusion et de substitution aux résultats de la segmentation morphologique :

1. L'**inclusion** correspond à deux constructions morphologiques :
   (a) **Expansion gauche** : le terme T_2 est obtenu par expansion gauche du terme T_1 si $T_2 = \alpha + T_1$ où α est une séquence de segments morphémiques incluant un seul préfixe ou une seule base et éventuellement d'autres types de segments. Par exemple, *lymphedema* est une expansion gauche de *edema* et *outpatient* est une expansion gauche de *patient*.
   (b) **Insertion** : le terme T_2 est obtenu par insertion d'un modifieur dans T_1 si T_2 est formé par l'insertion d'un préfixe ou d'une base dans T_1, avant la dernière base. Par exemple, *hepatosplenomegaly* est obtenu par l'insertion de *spleno* dans *hepatomegaly*.

   Les relations d'inclusion ne sont pas symétriques.
2. La **substitution** correspond à la construction suivante : le terme T_1 est obtenu par substitution à partir de T_2 si $T_1 = \beta + \alpha$ et $T_2 = \gamma + \alpha$ où β et γ sont des séquences de segments incluant au plus un préfixe ou une base. La relation de substitution est symétrique : si T_1 est obtenu par substitution à partir de T_2 alors T_2 est obtenu par substitution à partir de T_1. Par exemple, le couple de mots (*magnetotherapy*, *curietherapy*) correspond à un cas de substitution.

Les schémas ainsi définis décrivent aussi bien les mots composés, contenant plusieurs bases, que les mots dérivés par préfixation. Nous ne faisons aucune distinction entre ces deux cas. Ils devraient permettre d'induire des relations sémantiques précises entre termes. Nous faisons l'hypothèse que le schéma d'inclusion correspond à des relations de spécialisation tandis que le schéma de substitution correspond à des relations de co-hyponymie.

Nous allons maintenant décrire les données et ressources utilisées pour la vérification de nos hypothèses et l'analyse quantitative des relations sémantiques induites par ces schémas.

7.3 Données et ressources utilisées pour l'évaluation

7.3.1 Données

Nous avons vu en Section 1.4.5 p. 33 qu'il existe diverses manières d'évaluer les relations sémantiques acquises de manière automatique. Une des procédures d'évaluation les plus utilisées est la comparaison à une ressource de référence. Il est nécessaire dans ce cas de disposer de ressources pour la langue et le domaine cible. Ces ressources sont nombreuses en médecine et en anglais.

Nous avons donc évalué les relations sémantiques obtenues à partir des schémas morphologiques pour le corpus cancer-en. Les mots du corpus ont été découpés en segments morphémiques étiquetés par le système décrit dans le Chapitre 4, avec les valeurs de paramètres suivantes : N=5, a=0.9 et b=0.1.

Pour l'évaluation, nous avons utilisé deux ressources de référence : la première, WordNet, est une ressource de langue générale, tandis que la seconde, le thésaurus du NCI (NCIT) est un vocabulaire contrôlé spécifique au domaine du cancer. Nous allons tout d'abord présenter ces deux ressources avant de décrire les relations sémantiques qu'elles contiennent.

7.3.2 WordNet

WordNet est une base de données lexicales pour l'anglais [Miller, 1995]. Les mots sont groupés dans des classes de synonymes appelées *synsets*, qui représentent des concepts. Les synsets de noms sont organisés hiérarchiquement par la relation de spécialisation ou relation EST-UN. On trouve également des liens de méronymie (partie-de) et d'antonymie. Pour cette évalutation, nous avons utilisé la hiérarchie de noms de WordNet 2.0. Celle-ci comprend 114 648 noms différents et 79 689 synsets.

7.3.3 Le thésaurus du NCI

Le thésaurus du NCI (NCIT) est une ressource publiée par le National Cancer Institute et disponible sous une licence Open Source [National Cancer Institute, Office of Communications and Center for Bioinformatics, 2006]. Pour cette évaluation, nous avons utilisé la version 06.07d, rendue publique en août 2006, sous son format texte (*flat file*). Le format texte contient tous les termes associés aux concepts du NCIT (termes représentant le concept et termes synonymes) ainsi que les relations de spécialisation entre concepts. Le thésaurus contient environ 137 000 termes et 57 000 concepts. Environ 85% de ces concepts sont représentés par des termes polylexicaux et ne seront donc pas pris en compte dans cette évaluation qui ne concerne que les termes simples morphologiquement complexes.

7.3.4 Relations sémantiques présentes dans les ressources

La Figure 7.1 présente des sous-arbres issus de WordNet et du NCIT pour des concepts similaires dans les deux ressources.

Ces ressources contiennent les relations sémantiques suivantes :

- **Synonymie**. Dans WordNet, les synonymes sont regroupés dans le même synset. Dans le NCIT, les synonymes représentent le même concept. Par exemple, dans WordNet, *neurolysin* et *neurotoxin* appartiennent au même synset et dans le NCIT, *endotoxin* et *bacterial pyrogen* représentent le même concept. Les synonymes dans le NCIT incluent également des variantes flexionnelles et orthographiques comme *metal – metals* et *tumor – tumour*.
- **Spécialisation**. Les relations hiérarchiques de spécialisation relient des synsets dans WordNet et les concepts dans le NCIT. Si l'on reprend les exemples de la Figure 7.1, dans WordNet, le synset contenant *cytotoxin* est un hyponyme du synset contenant *toxin* et dans le NCIT le concept représenté par *enterotoxin* est un sous-concept de celui représenté par *toxin*.
- **Co-hyponymie**. Cette relation est définie entre synsets dans WordNet : le synset contenant *cytotoxin* et celui contenant *neurotoxin* et *neurolysin* sont co-hyponymes du synset contenant *toxin*. Dans le NCIT, la relation de co-hyponymie est définie entre concepts : le concept représenté par *enterotoxin* et celui représenté par *neurotoxin* sont co-hyponymes du concept représenté par *toxin*.

WordNet contient les relations supplémentaires suivantes :

- **Antonymie**. Cette relation symétrique est définie entre formes : par exemple, *victory* et *defeat* sont antonymes.
- **Méronymie**. Cette relation est définie entre synsets. Par exemple *brim* est un méronyme de *hat* : *hat* est donc le holonyme de *brim*.

Certaines de ces relations ne sont définies qu'entre synsets dans WordNet et entre concepts dans le NCIT. Pour pouvoir évaluer les relations induites par les schémas morphologiques, qui sont définis pour des mots, nous avons donc considéré les mots appartenant aux synsets de

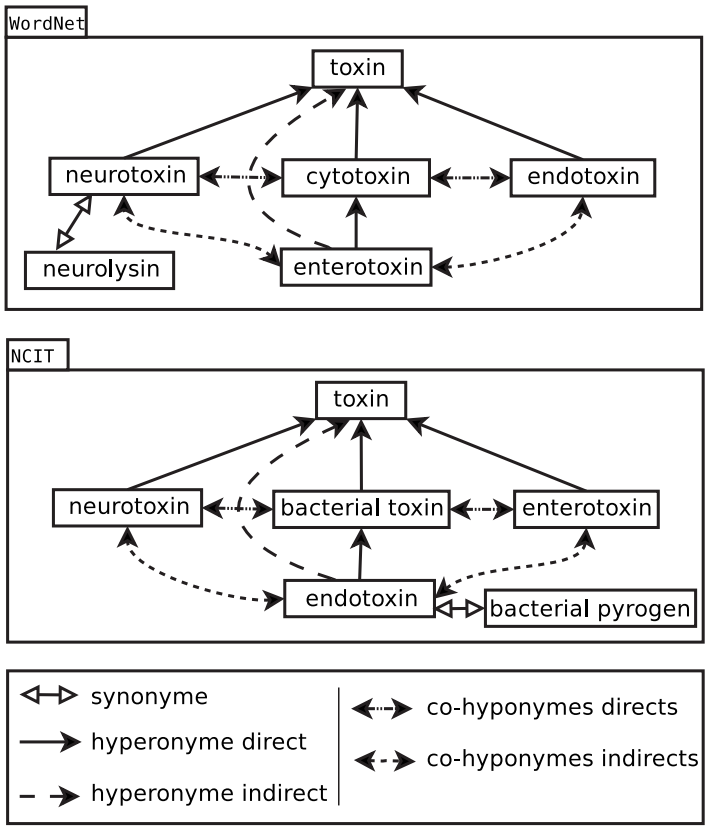


FIG. 7.1: Relations sémantiques dans WordNet et dans le NCIT.

WordNet et les termes représentant les concepts dans le NCIT. Les couples de mots identifiés par les schémas d'inclusion et de substitution dont les deux membres n'appartiennent pas aux ressources de référence ne sont pas pris en compte.

Nous avons également comptabilisé les relations sémantiques de spécialisation et de co-hyponymie indirecte :

- Un synset (resp. concept) A est un **hyperonyme indirect** d'un synset (resp. concept) B si A est un ancêtre de B dans la hiérarchie, obtenu par transitivité de la relation d'hyperonymie à partir de B.
- Un synset (resp. concept) B et un synset (resp. concept) C sont des **co-hyponymes indirects** d'un autre synset (resp. concept) A si A est un hyperonyme direct ou indirect de B et de C. Nous avons fixé la distance maximale de A à B et C dans la hiérarchie à 3, c'est-à-dire que A est au plus un arrière-grand-parent de B et C.

Nous allons maintenant détailler les résultats obtenus.

7.4 Résultats

Les Figures 7.2 à 7.5 présentent les résultats de la comparaison des relations induites par les schémas d'inclusion et de substitution aux relations sémantiques décrites dans les ressources. Les diagrammes 7.2 et 7.3 détaillent le nombre de relations de chaque type correspondant aux couples de mots identifiés par les schémas d'inclusion et de substitution. Les diagrammes 7.4 et 7.5 détaillent quant à eux la proportion de relations directes, indirectes et absentes des ressources.

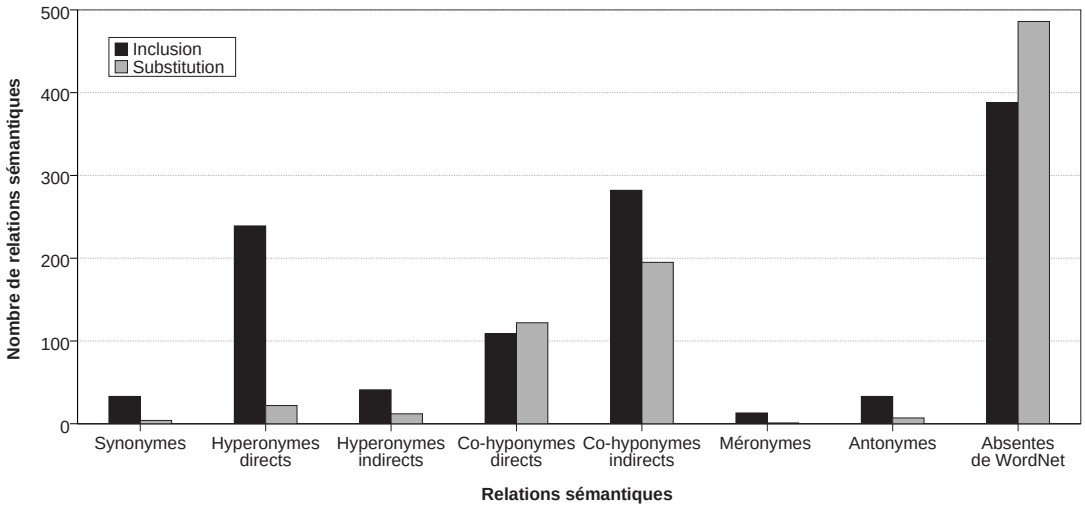


FIG. 7.2: Nombre de relations sémantiques de WordNet identifiées par les schémas d'inclusion et de substitution.

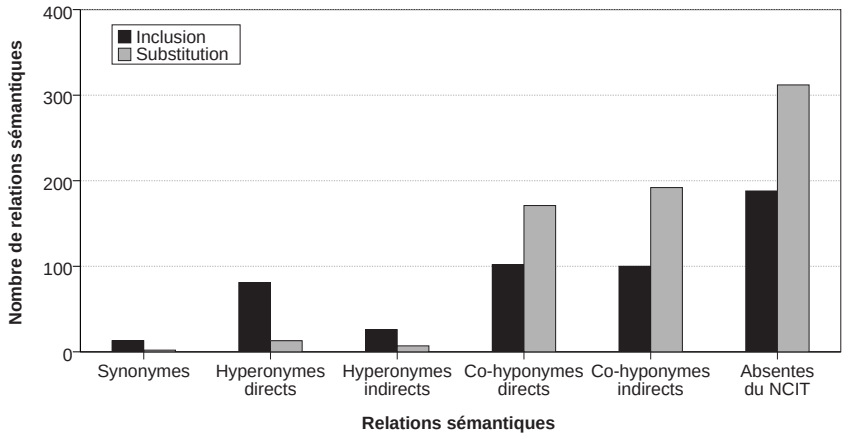


FIG. 7.3: Nombre de relations sémantiques du NCIT identifiées par les schémas d'inclusion et de substitution.

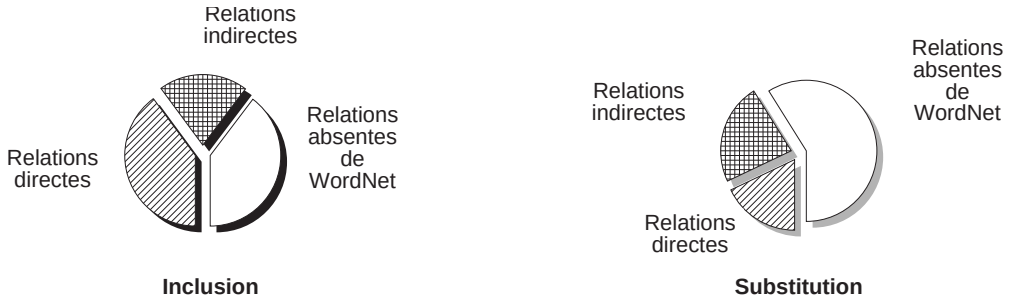


FIG. 7.4: Proportions de relations directes, indirectes et absentes dans WordNet.

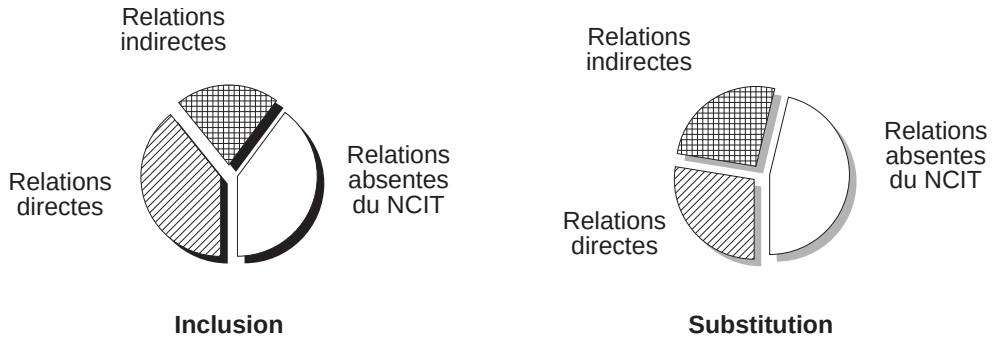


FIG. 7.5: Proportions de relations directes, indirectes et absentes dans le NCIT.

Nous allons maintenant analyser les résultats en fonction des différentes relations sémantiques.

7.4.1 Spécialisation

Dans la Section 7.2, nous avons fait l'hypothèse que le schéma d'inclusion devrait permettre d'identifier des relations de spécialisation tandis que le schéma de substitution devrait correspondre à la co-hyponymie. Les résultats montrent que les relations de spécialisation sont prédites de manière préférentielle par le schéma d'inclusion. Seul un faible nombre de relations de spécialisation directe ou indirecte sont prédites par le schéma de substitution. Cette remarque est valable pour WordNet aussi bien que pour le NCIT.

Nous présentons ci-dessous quelques exemples de relations hiérarchiques directes et indirectes découvertes par les schémas d'inclusion et de substitution. Les différentes relations sont indiquées par les sigles suivants : `i>` (hyperonymie directe – inclusion), `i>>` (hyperonymie indirecte – inclusion), `s>` (hyperonymie directe – substitution), `s>>` (hyperonymie indirecte – substitution).

WordNet	NCIT
conductor `i>` semiconductor	dermatitis `i>` radiodermatitis
fish `i>>` swordfish	filament `i>>` neurofilament
weekday `s>` tuesday	radiography `s>` esophagography
radiogram `s>>` myelogram	endoscopy `s>>` sigmoidoscopy

7.4.2 Co-hyponymie

Dans ce cas, la supériorité d'un schéma sur l'autre est moins flagrante. Le schéma de substitution ne prédit pas beaucoup plus de relations de co-hyponymie que le schéma d'insertion, parfois même moins, dans le cas des relations de co-hyponymie indirecte pour WordNet. Il faut toutefois remarquer que le schéma de substitution prédit avant tout des relations de co-hyponymie et très peu d'autres relations.

Nous présentons ci-dessous quelques exemples de relations de co-hyponymie directe et indirecte découvertes par les schémas d'inclusion et de substitution. Les différentes relations sont indiquées par les sigles suivants : `>i<` (co-hyponymie directe – inclusion), `>>i<<` (co-hyponymie indirecte – inclusion), `>s<` (co-hyponymie directe – substitution), `>>s<<` (co-hyponymie indirecte – substitution).

WordNet	NCIT
deoxyadenosine `>i<` adenosine	kilogram `>i<` gram
oxytetracycline `>>i<<` tetracycline	immunodiagnosis `>>i<<` diagnosis
chemotherapy `>s<` radiotherapy	cytopathology `>s<` neuropathology
microscope `>>s<<` laryngoscope	biotherapy `>>s<<` chemotherapy

Le schéma d'inclusion est relativement ambigu. En effet, l'inclusion ne correspond pas toujours à la spécialisation. Prenons l'exemple du couple de mots suivants : (*hypothalamus*, *thalamus*) : le terme *thalamus* est inclus dans le terme *hypothalamus*. Mais l'*hypothalamus* n'est pas un type de *thalamus*. La relation sémantique qui lie ces deux mots est spatiale : l'*hypothalamus* se trouve sous le *thalamus*. Dans WordNet, *thalamus* et *hypothalamus* sont co-hyponymes de *neural structure* et co-méronymes de *diencephalon* ; dans le NCIT, les deux termes sont co-hyponymes de *Brain_Part*.

On retrouve également parmi les relations de co-hyponymie marquées par l'inclusion les cas d'antonymie (*nonsmoker* – *smoker*) et le cas particulier des unités de mesure comme *volt* – *kilovolt*.

7.4.3 Synonymie

Les synonymes sont généralement des composés populaires comme (*paper, newspaper*) ou (*tan, suntan*). Dans ce cas, la première base, comme *sun* dans *suntan* est optionnelle lorsque le mot est utilisé dans des contextes non ambigus. On trouve également des cas de supplétion, c'est-à-dire des segments morphologiques qui ont le même sens mais dont la forme est différente, comme par exemple *mis-/dis-* (*mistrust, distrust*) ou *di-/bi-* (*biphosphonate, diphosphonate*).

7.4.4 Antonymie

Les couples d'antonymes s'opposent par leurs préfixes. On trouve les oppositions suivantes :
- *an+/ε+* : anovulation – ovulation ;
- *de+/ε+* : decontamination – contamination ;
- *dis+/ε+* : disservice – service, disagreement – agreement ;
- *in+/ε+* : inactivation – activation, inactivity – activity ;
- *mis+/ε+* : misconception – conception ;
- *non+/ε+* : nonparticipation – participation, nonsmoker –smoker ;
- *un+/ε+* : unwillingness – willingness, unfamiliarity – familiarity ;
- *hyper+/hypo+* : hypertension – hypotension ;
- *in+/out+* : inflow – outflow.

La majorité des préfixes s'opposent à l'absence de préfixe, et la structure morphologique correspond donc au schéma d'inclusion. La substitution de préfixes opposés est plus rare.

7.4.5 Méronymie

Similairement aux antonymes, les méronymes sont généralement marqués par l'adjonction d'un préfixe ou d'un mot à leur holonyme :

- *half+/ε+* : half-hour – hour, half-century – century ;
- *mid+/ε+* : midwinter – winter, midnight – night ;
- *quarter+/ε+* : quarter-century – century ;
- *south+/ε+* : soutwest – west ;
- *sub+/ε+* : subfamily – family.

Les relations de méronymie marquées par l'inclusion correspondent avant tout à des relations temporelles et spatiales. On constate que les éléments joints à l'holonyme sont soit des préfixes comme *sub+*, soit des morphèmes libres comme *mid+*, *half+*, *quarter+* ou *south+*.

7.4.6 Relations manquantes

À peu près la moitié des couples de mots identifiés par les schémas d'insertion et de substitution ne sont liés par aucune relation sémantique dans le NCIT ou WordNet, que ce soit par des relations directes ou indirectes (voir les Figures 7.5 et 7.4). Le schéma d'inclusion semble un peu plus fiable de ce point de vue car il permet l'identification d'un plus grand nombre de relations sémantiques valides. Les relations manquantes peuvent l'être du fait (a) d'une lacune, motivée ou fortuite, dans la ressource et (b) d'un manque de précision de notre méthode d'identification des relations sémantiques. La première cause est difficile à juger : elle dépend des choix effectués par les concepteurs de la ressource ou encore de manques que les schémas d'inclusion et de substitution peuvent permettre de combler. Concernant la seconde, nous avons identifié les problèmes suivants :

Segmentations morphologiques incorrectes

Les résultats de la segmentation morphologique ne sont pas parfaits et les schémas d'inclusion et de substitution peuvent selon les cas être appliqués à des mots non morphologiquement liés :

- lobule = $\text{lobul}_\text{b} + \text{e}_\text{s}$
  globule = $\text{g}_\text{p} + \text{lobul}_\text{b} + \text{e}_\text{s}$
- copy = $\text{cop}_\text{b} + \text{y}_\text{s}$
  sigmoidoscopy =$\text{sigmoid}_\text{b} + \text{o}_\text{l} + \text{s}_\text{l} + \text{cop}_\text{b} + \text{y}_\text{s}$

Segments ambigus

Tout comme les mots, certains segments de mots peuvent avoir plus d'un sens. Considérons la liste de mots suivante : *gram*, *microgram*, *kilogram*, *roentgenogram*, *mammogram*. Tous ces mots partagent le segment final *+gram*. Mais dans ces exemples, *gram* a deux sens différents : il peut faire référence à une unité de mesure, comme dans *microgram* et *kilogram* ou à une image comme dans *roentgenogram* et *mammogram*. Par conséquent, des couples non liés comme (*kilogram*,*mammogram*) sont identifiés en même temps que des couples réellement liés comme (*kilogram*,*microgram*). Des cas d'ambiguïtés similaires ont également été relevés par [Grabar et Zweigenbaum, 2002b], pour les termes poly-lexicaux.

Lien sémantique trop général

Certaines bases identifiées au cours de la segmentation ont une contribution sémantique réduite. Prenons l'exemple du segment *+logy* qui se trouve dans des mots aussi divers que :

technology, *pathology* ou *pharmacology*. Certains mots qui se terminent par *+logy* ont parfois un sens très proche. C'est le cas par exemple pour *nephrology*, *hematology* et *rheumatology* qui sont tous trois co-hyponymes de *Internal_Medicine* dans le NCIT. Dans d'autres cas cependant, la contribution sémantique de *+logy* est trop faible pour considérer les mots comme sémantiquement reliés, comme par exemple *psychology*, *opthalmology* et *technology*.

7.5 Conclusion

Les résultats que nous obtenons à partir des schémas d'inclusion et de substitution montrent que les relations hiérarchiques sont avant tout marquées par l'inclusion. Les cas de co-hyponymie sont identifiés aussi bien par le schéma de substitution que par celui d'inclusion. Les relations d'antonymie et de méronymie correspondent quant à elles à l'utilisation de préfixes spécifiques.

Nous avons donc montré que les résultats de la segmentation morphologique pouvaient être utilisés pour détecter des relations sémantiques, par la recherche de certains schémas de construction spécifiques. Cependant, ces schémas sont trop ambigus pour permettre un étiquetage suffisamment précis des relations induites. Il reste donc à déterminer des méthodes pour lever ces ambiguités. Deux manières de procéder sont proposées dans la littérature.

La première consiste à utiliser la méthodologie du bootstrap en fournissant au système une liste de mots liés par un lien sémantique précis pour déterminer les structures morphologiques correspondantes [Schwab *et al.*, 2005]. Ces structures morphologiques peuvent, après validation, permettre de découvrir de nouveaux mots liés par la relation cible. Il s'agit d'une méthode semi-automatique, qui s'apparente à la méthode utilisée pour découvrir des relations sémantiques à partir de patrons lexico-syntaxiques [Hearst, 1992, Morin, 1998].

La seconde consiste à décrire dans un premier temps les opérations morphologiques et leur impact sémantique pour ensuite les identifier dans les données dont l'analyse morphologique a déjà été effectuée [Grabar et Hamon, 2006]. Cette méthode nécessite donc une description linguistique préalable.

Il semble donc qu'il est difficile de se passer d'une validation ou d'une interprétation humaine lorsqu'il s'agit de typer les relations sémantiques induites. Cette intervention peut se faire soit par le typage explicite de certains segments morphémiques, soit par la donnée de mots liés par une relation précise. Ceci constitue la limite des approches non supervisées.

La méthode que nous proposons a également d'autres limites. Les schémas définis se prêtent bien aux langues germaniques, mais ne sont pas directement utilisables pour le français, du moins en ce qui concerne la composition populaire. Les résultats qu'il est possible d'obtenir par cette méthode sont également fortement dépendants de la complexité morphologique du vocabulaire traité. D'autres expériences devraient permettre de mieux déterminer les domaines et langues d'application de la méthodologie proposée.

De plus, l'utilisation de la morphologie ne constitue pas une solution complète pour la détection de relations sémantiques. Elle ne permet pas de remplacer les méthodes basées sur les informations contextuelles [Light, 1996] car les indices morphologiques ne sont pas disponibles pour tous les mots. Nous n'avons pas étudié les possibilités de combinaison des deux indices, indices morphologiques et indices contextuels, pour l'acquisition de relations sémantiques. Une des pistes possibles consiste en l'utilisation des segments morphémiques comme unités contextuelles à la place des mots. Ceci pourrait permettre de résoudre un des problèmes des méthodes utilisant les similarités contextuelles, à savoir le manque de données disponibles en corpus pour les mots les plus rares.

Conclusion

Contributions

Rappelons l'objectif général que nous avions énoncé en introduction de ce mémoire : il s'agissait de mettre de « l'ordre dans le désordre », c'est-à-dire de structurer les données textuelles disponibles en abondance à l'heure actuelle. Cette structuration vise d'une part à découvrir les concepts ou unités de connaissance d'intérêt dans les textes et d'autre part à les organiser. Nous avions remarqué que le vocabulaire spécifique aux domaines techniques et notamment la médecine présente une caractéristique intéressante, celle d'inclure de nombreux termes morphologiquement complexes, formés par dérivation et composition savante. Nous avons donc proposé d'utiliser cette spécificité pour acquérir des connaissances à partir de textes de spécialité en intégrant la morphologie dans un processus global allant des textes aux ressources lexico-sémantiques.

La première étape d'un tel processus consiste en l'acquisition des données textuelles à analyser. Nous avons présenté une méthodologie d'acquisition automatique de corpus à partir d'Internet, grâce à une liste de mots amorce constituée de termes du domaine. Cette méthode nous a permis de construire quatre corpus couvrant deux thématiques différentes, le cancer et la volcanologie, en deux langues, le français et l'anglais.

Ces données textuelles ont ensuite été utilisées pour l'apprentissage de connaissances morphologiques. Nous avons développé deux systèmes qui diffèrent par les résultats obtenus. Le premier segmente les mots en sous-unités étiquetées tandis que le second forme des familles morphologiques. Ces deux systèmes procèdent par apprentissage non supervisé, c'est-à-dire qu'ils utilisent pour seules données les corpus d'apprentissage, sans aucune règle ou lexique spécifiques à la langue ou au domaine traités. Le premier système obtient de très bons résultats en français, en anglais, en finnois et en turc, qui sont pourtant des langues très différentes, et ce aussi bien pour des mots de la langue générale que des vocables spécifiques à des domaines techniques. Les performances du second système pour l'acquisition de familles morphologiques en anglais et en français sont encore meilleures. Le système d'analyse morphologique par segmentation a également été comparé à d'autres systèmes similaires au cours d'un challenge international de segmentation morphologique organisé dans le cadre du réseau d'excellence européen PASCAL. Il y a obtenu de très bons résultats, proches du meilleur système actuel (Morfessor) et a surpassé les autres systèmes participants en finnois et en turc. Il faut toutefois noter que des évaluations effectuées en allemand pour un système de conversion de graphèmes en phonèmes se sont révélées assez décevantes et ont montré que les performances des systèmes de segmentation morphologique non supervisée ne sont pas encore suffisantes pour une utilisation dans des applications réelles.

Dans la dernière partie du mémoire, nous avons décrit deux applications possibles des connaissances morphologiques acquises. Les résultats de l'analyse morphologique par classifi-

cation ont été utilisés pour identifier les familles morphologiques spécifiques au corpus étudié et produire des représentations graphiques au format HTML, appelées listes pondérées. Les listes pondérées facilitent l'analyse des données par la mise en valeur des thématiques les plus saillantes dans le corpus.

La deuxième application proposée est l'acquisition de relations sémantiques. Nous avons défini deux schémas reposant sur les résultats de la segmentation morphologique afin d'identifier les relations d'hyperonymie et de co-hyponymie. Les résultats obtenus montrent que la structure morphologique permet la découverte de relations sémantiques spécifiques mais aussi que ces relations sont marquées de diverses manières qui peuvent être ambiguës.

Retour sur les choix méthodologiques

L'ensemble des systèmes et méthodes décrits dans ce mémoire ont été élaborés dans un cadre méthodologique que nous avons défini dans l'introduction et qui imposait le travail sur des corpus de textes de spécialité, pour diverses langues, ainsi que l'utilisation de méthodes d'apprentissage non supervisées.

Les seules données nécessaires à l'apprentissage des connaissances morphologiques ont été extraites de corpus de textes construits automatiquement à partir d'Internet. Il nous a été ainsi possible de produire facilement des corpus de textes de spécialité en plusieurs langues. De plus, ces corpus constituent des données réalistes, caractéristiques des données textuelles que l'on trouve sur Internet. Bien sûr, le fait de disposer de corpus constitués rapidement à partir d'Internet présente aussi un certain nombre de désavantages. En effet, le contenu n'a pas été corrigé manuellement et les traitements que nous avons appliqués sont relativement sommaires. Il reste donc de nombreux mots étrangers ou mal orthographiés. Nous n'avons pas non plus traité le cas des noms propres et autres entités nommées. Tous ces éléments sont autant de facteurs qui compliquent encore l'apprentissage car il est effectué à partir de données bruitées. Il aurait bien sûr été possible de travailler à partir de listes de mots contrôlés, et peut-être d'ailleurs que nous obtiendrions des résultats encore meilleurs. Nous pensons toutefois qu'il est nécessaire de prendre en compte les données non standard lors de l'élaboration d'un système car elles garantissent sa souplesse d'utilisation dans des situations réalistes.

Ces mêmes critères de souplesse expliquent également pourquoi nous avons travaillé sur des corpus en différentes langues et décrivant des domaines distincts. Cette stratégie s'est d'ailleurs avérée payante car le système d'analyse morphologique par segmentation a obtenu de bons résultats en finnois et en turc, pour des données d'apprentissage non filtrées et ce alors même que nous n'avons aucune connaissance de ces deux langues.

Nous n'avons toutefois pas épuisé l'ensemble des indices disponibles en corpus car nous n'avons pour l'heure pas utilisé les informations liées aux occurrences des vocables dans le texte. On rejoint là les questions de sémantique qui, nous l'avons vu au cours des divers chapitres, peuvent difficilement être éludées. En effet, c'est dans cette direction qu'il faut aller pour améliorer les systèmes d'analyse morphologique non supervisés car le sens fait partie intégrante du niveau morphologique. Reste encore à déterminer le type d'informations sémantiques disponibles et applicables. L'utilisation de ressources externes au corpus, sous forme de thésaurus ou de terminologies, est une des possibilités qui présente un certain nombre d'avantages car les relations sémantiques sont explicitement encodées et validées. Cette approche suppose toutefois la disponibilité des ressources pour la langue et le domaine traités, ce qui n'est pas toujours le cas. Une

autre méthode consiste à estimer la similarité sémantique par des mesures de similarité contextuelle. Les proximités sémantiques obtenues sont moins précises et il peut être difficile d'obtenir des mesures pour les mots les plus rares (problème connu sous le nom de *data sparseness* en anglais).

Perspectives

Les perspectives de notre travail sont multiples et concernent aussi bien l'amélioration des méthodes proposées que leur utilisation pour d'autres applications.

Nous avons vu que nos systèmes d'analyse morphologique présentent certaines limites. Le système d'analyse morphologique par classification est le plus récent et par conséquent le plus perfectible. Nous envisageons de l'améliorer sur les points suivants :

- apprentissage de nouveaux préfixes et gestion de la composition par la classification multiple ;
- possibilité d'analyser les mots absents du corpus d'apprentissage, soit en les classant dans une famille existante, soit en formant de nouvelles familles ;
- production de segmentations à partir des résultats de la classification ;
- couplage des informations sur la graphie des mots et des informations contextuelles.

Cette dernière amélioration semble nécessaire pour les deux systèmes. En effet, nous utilisons uniquement une liste des mots du corpus pour l'apprentissage des connaissances morphologiques, sans référence aucune au contexte d'occurrence des mots traités. Cependant, il n'est pas évident de déterminer l'étape à laquelle les informations sémantiques ou contextuelles doivent être injectées dans le système d'analyse morphologique. Dans la Section 2.3.3, p. 52, nous avions distingué deux types d'approches : celles qui intègrent les informations sémantiques en début de processus et celles qui les intègrent en fin de processus. En début de processus, les informations sémantiques permettent d'initialiser l'apprentissage de manière aussi précise que possible. En fin de processus, elles permettent de filtrer les résultats non pertinents. Dans les deux cas, elles ne sont pas intégrées à toutes les étapes du traitement, mais interviennent uniquement à un moment précis de l'apprentissage. L'intégration des informations contextuelles à un système d'analyse morphologique demande donc réflexion.

Les améliorations des systèmes devraient également s'accompagner d'évaluations complémentaires pour d'autres langues que l'anglais et le français et d'autres types de corpus, notamment des corpus non techniques.

Nous souhaitons également agrandir l'éventail des utilisations possibles de nos systèmes. Dans ce mémoire, nous avons proposé deux applications pour l'acquisition de ressources lexicales. Le système d'analyse par segmentation a également été évalué pour deux autres applications liées au traitement de la parole : reconnaissance de la parole en finnois et en turc et conversion de graphèmes en phonèmes pour la synthèse de la parole en allemand.

Les autres applications que nous prévoyons d'évaluer dans un avenir proche sont la recherche d'information et la classification de documents. En recherche d'information, les ressources morphologiques peuvent être utilisées pour l'extension de requêtes [Moreau et Claveau, 2006] ou l'indexation. Pour la catégorisation de documents, les familles morphologiques ou les segments morphémiques peuvent servir de descripteurs aux documents à classer [Witschel et Biemann, 2006].

Au cours de cette thèse, nous avons tenté de traiter le problème de l'acquisition de connaissances lexicales et sémantiques en nous focalisant sur une unité linguistique, le morphème, qui n'est que marginalement prise en compte dans la littérature, pourtant abondante à ce sujet. Nous espérons avoir su montrer que la morphologie présente un intérêt certain pour cette problématique, car, ne l'oublions pas, le morphème est la plus petite unité linguistique porteuse de sens. Et n'est-ce pas finalement la quête du sens qui est au centre des stratégies visant à mettre de « l'ordre dans le désordre » apparent des données textuelles ? En mettant le morphème au cœur de cette problématique, nous ne faisons que lui rendre la place qui lui est due.

Annexe A

Caractéristiques des corpus

A.1 Corpus construits automatiquement à partir d'internet

Tous ces corpus ont été construits automatiquement à partir d'internet à l'aide de la méthode décrite dans le chapitre 3. Les mots ont été extraits après tokenisation et suppression des paragraphes écrits dans une langue différente de la langue cible. Les mots contenant des caractères spéciaux comme © ont également été exclus du décompte.

A.1.1 Corpus anglais

Amorces	breast ; cancer ; treatment ; therapy ; chemotherapy ; tumor ; mastectomy ; biopsy ; mammogram ; mammography ; carcinoma ; lump ; lymphedema ; radiotherapy ; lobular ; metastasis ; osteoporosis ; oncologist ; fibrocystic ; hyperplasia.
Nombre de documents	4 549
Taille	46,1 Mo
Taille moyenne d'un document	10,4 Ko
Nombre de formes différentes	86 149
Nombre d'occurrences	7 060 019

Fig. A.1: Caractéristiques du corpus anglais sur le cancer du sein (cancer-en).

Amorces	lava ; volcano ; eruption ; magma ; flows ; ash ; crater ; pyroclastic ; vent ; cones ; caldera ; basalt ; earthquakes ; tephra ; pumice ; cinder ; scoria ; lahar ; pahoehoe ; fumarole.
Nombre de documents	2 537
Taille	18,7 Mo
Taille moyenne d'un document	7,5 Ko
Nombre de formes différentes	47 789
Nombre d'occurrences	2 935 238

Fig. A.2: Caractéristiques du corpus anglais sur la volcanologie (volcano-en).

A.1.2 Corpus français

Amorces	sein ; cancer ; traitement ; radiothérapie ; récidive ; chimiothérapie ; axillaire ; mammaire ; tumeur ; ganglions ; métastases ; mastectomie ; biopsie ; mammographie ; exérèse ; carcinome ; microcalcifications ; curage ; tumorectomie ; lymphoedème.
Nombre de documents	1 233
Taille	10,3 Mo
Taille moyenne d'un document	8,5 Ko
Nombre de formes différentes	46 898
Nombre d'occurrences	1 455 747

FIG. A.3: Caractéristiques du corpus français sur le cancer du sein (cancer-fr).

Amorces	éruption ; volcan ; lave ; magma ; coulées ; cratère ; cendres ; cône ; caldeira ; basalte ; scories ; dôme ; ponces ; pyroclastiques ; lapilli ; lahars ; séismes ; fumerolles ; pahoehoe ; maar.
Nombre de documents	1 243
Taille	11,8 Mo
Taille moyenne d'un document	9,7 Ko
Nombre de formes différentes	59 768
Nombre d'occurrences	1 784 582

FIG. A.4: Caractéristiques du corpus français sur la volcanologie (volcano-fr).

A.2 Corpus de référence

En complément de ces corpus, couvrant des domaines spécifiques, nous avons également utilisé des corpus généralistes pour chaque langue, provenant de la collection de corpus de l'Université de Leipzig [Quasthoff *et al.*, 2006] téléchargeables à partir de l'url suivante : `http://corpora.uni-leipzig.de/`. Ces corpus contiennent des phrases sélectionnées au hasard à partir de journaux ou du web. Les phrases incomplètes ainsi que les phrases de langue étrangère ont été suprimées. Nous appelons leipzig-en le corpus anglais et leipzig-fr le corpus français. Les caractéristiques de ces deux corpus sont résumées dans le Tableau A.1.

	leipzig-en	**leipzig-fr**
Taille	12,9 Mo	13,5 Mo
Nombre de phrases	100 000	100 000
Nombre de formes différentes	68 484	84 274
Nombre d'occurrences	2 048 751	2 109 644
Sources	AP Associated Press Financial Times OTS news ticker Wall Street Journal	L'Humanité Le Monde Le Parisien Le Télégramme Zénit

Tab. A.1: Caractéristiques des corpus de référence.

Annexe B

Outils et programmes

B.1 Outils

L'ensemble des programmes ont été écrits dans le langage *Python* (`http://www.python.org/`). Python est un langage de script orienté objet portable, libre et gratuit. Il est très performant pour l'analyse de données textuelles et présente une syntaxe simple qui le rend très lisible. Ceci explique pourquoi Python est de plus en plus utilisé, que ce soit pour l'enseignement ou le développement d'applications professionnelles.

Il existe de nombreux projets de TAL (Traitement Automatique des Langues) qui utilisent Python. Nous n'en citons ici que quelques uns :

- Natural Language Toolkit (NLTK : `http://nltk.sourceforge.net/`) : Ensemble de librairies et programmes Python pour le traitement automatique des langues.
- TreeTaggerWrapper (`http://www.limsi.fr/Individu/pointal/python/treetaggerwrapper-doc/`) : Module permettant d'utiliser le TreeTagger via Python.
- PyWordNet (`http://osteele.com/projects/pywordnet/`) : Interface Python pour WordNet. Nous avons utilisé ce module pour l'évaluation des relations sémantiques décrite dans le Chapitre 7.
- Extensions de l'outil d'indexation plein texte TextIndexNG (`http://opensource.zopyx.com/software/textindexng3`) : Normalisation (retrait des accents), découpage du texte en mots, racinisation (basée sur Snowball), distance de Levenshtein.

B.2 Programmes pour la construction et le pré-traitement des corpus

Dans cette section nous détaillons les programmes annexes de construction et de pré-traitement des corpus que nous avons développés au cours de la thèse.

B.2.1 Construction automatique de corpus à partir d'Internet : Web Corpus Builder

Ce module permet la construction automatique de corpus à partir d'Internet. Nous l'avons initialement programmé pour nos besoins propres, mais il a également été utilisé par Eric Atwell de l'université de Leeds au Royaume-Uni pour les besoins d'un cours (DB32 : Technologies for Knowledge Management).

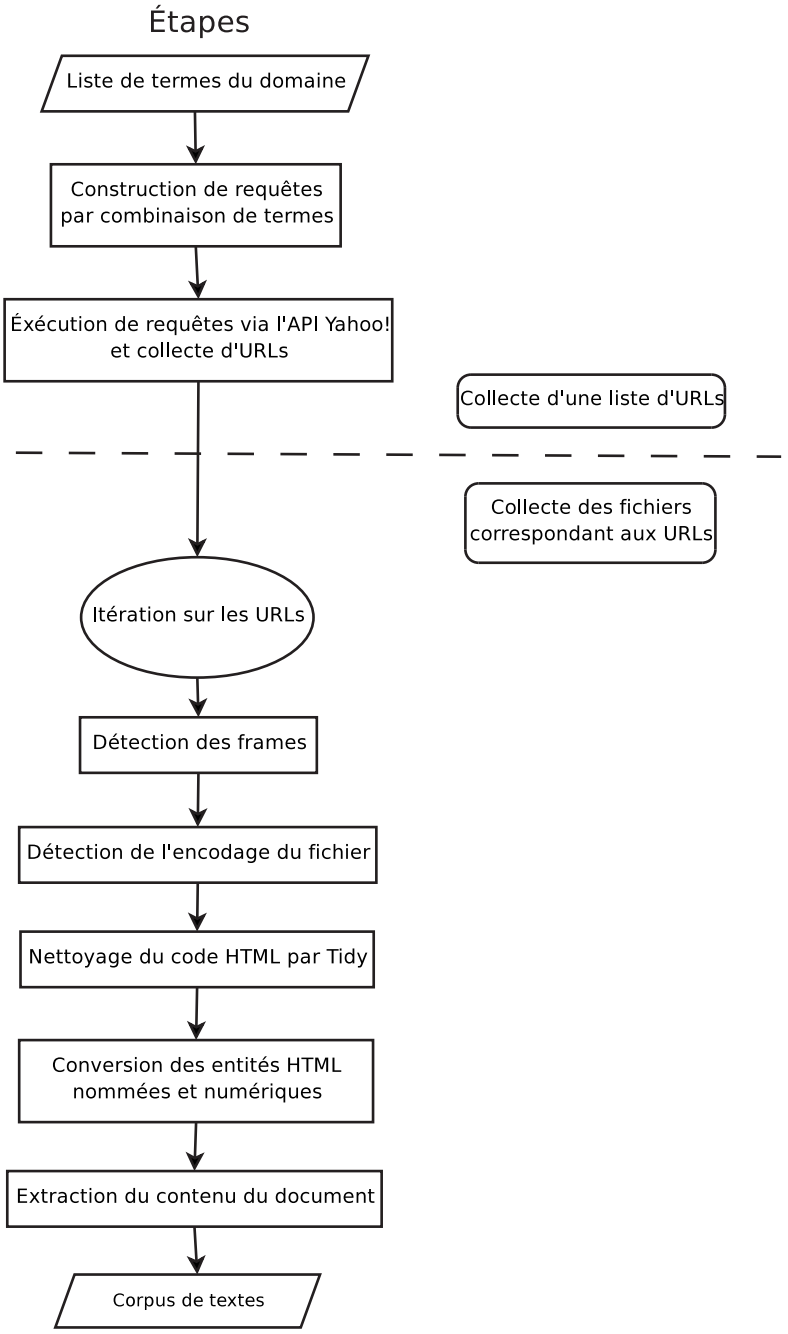


FIG. B.1: Étapes de construction d'un corpus à partir d'Internet.

Il s'inspire des programmes suivants :
- BootCaT [Baroni et Bernardini, 2004] (`http://sslmit.unibo.it/~baroni/bootcat.html`)
- BTE : Body Text Extraction [Finn *et al.*, 2001] (`http://www.aidanf.net/software/bte-body-text-extraction`)

Il utilise les modules externes suivants :
- μTidylib : Adaptateur Python pour TidyLib (`http://utidylib.berlios.de/`)
- Librairie Yahoo pYsearch-1.3 (`http://developer.yahoo.net/download/`)

B.2.2 Segmentation de textes en mots, phrases et paragraphes : PyTokeniser

Ce module permet le découpage de textes en mots, phrases et paragraphes. Il est composé de deux tokeniseurs différents :
- **SimpleTokeniser** à base d'expressions régulières simples :
  - Reconnaissance des limites de phrases par `([.!?...]+)\s+`.
  - Reconnaissances des mots par `(\w+)`.
  - Reconnaissance des autres types d'unités par `([^\w\s]+)`.
- **RegExpTokeniser** qui utilise des expressions régulières plus élaborées, dépendantes de la langue du texte pour certaines et inspirées de [Grefenstette, 1998, Pointal, 2004]. Elles permettent de distinguer divers types d'unités :
  - Les caractères indiquant une frontière de segment de manière non ambigue où qu'ils soient placés tels que `?`, `!`, `(`, `)`, etc.
  - Les caractères et chaînes de caractères qui doivent être séparés de la suite lorsqu'ils sont situés après un espace : `c'`, `l'`, `d'`, `s'`, `qu'` etc. Cette règle permet de considérer les pronoms avant un mot comme des unités indépendantes. Il en va de même pour les articles ou les conjonctions élidés.
  - Les caractères et chaînes de caractères qui doivent être séparés de ce qui précède lorsqu'ils sont situés avant un espace : `-elle`, `-elles`, `-en`, `-il`, `-ils`, `-je`, `-nous`, `-on`, `-t`, `-vous`. Cette règle permet notamment de considérer les pronoms situés à droite d'un verbe comme des unités lexicales indépendantes.
  - Les nombres.
  - Les abbréviations.
  - Les URLs.
  - Les adresses e-mail.

B.2.3 Identification de la langue d'un texte : Language Identifier

Ce module permet l'identification de la langue d'un texte. Il utilise la méthode de [McNamee, 2005] qui de décompose en deux étapes :

1. Apprentissage d'un profil des mots les plus fréquents pour chaque langue à partir d'un corpus d'apprentissage (voir Tableau B.1). Les mots sont pondérés en fonction de leur fréquence d'occurrence dans le texte : plus le mot est fréquent, plus son poids est important (voir Tableau B.2).
2. Utilisation des profils de fréquence pour identifier la langue d'un texte. La langue d'un texte est déterminée en calculant un score pour chaque langue. Ce score correspond à la somme des poids des mots du texte. Prenons l'exemple de la séquence « in die ». Les scores pour les différentes langues sont les suivants :
   - allemand : 1,61 + 2,46 = 4,07

– anglais : 2,59 + 0 = 2,59
– français : 0 + 0 = 0
– italien : 1,09 + 0 = 1,09
– néerlandais : 2,07 + 1,23 = 3,30

La langue de la séquence "in die" est donc l'allemand. Si aucun score ne dépasse 0, alors la langue du texte est inconnue.

Langue	**Auteur**	**Titre**	**Fichier**	**Taille**
Allemand	John Henry Mackay	Der Schwimmer	15068-8.txt	436,4 Ko
Anglais	F. B. Tarbell	A History Of Greek Art	hgrkr10.txt	318,0 Ko
Français	Jules Verne	Le Docteur Ox	11589-8.txt	434,8 Ko
Italien	Giuseppe Garibaldi	Clelia	8clel10.txt	485,7 Ko
Néerlandais	Jozef Muls	De Val van Antwerpen	11500-8.txt	200,7 Ko

Tab. B.1: Textes utilisés pour l'apprentissage par le programme d'identification de la langue d'un texte.

Allemand		**Anglais**		**Français**		**Italien**		**Néerlandais**	
und	3,82	the	8,66	de	3,82	di	3,03	de	5,48
er	3,57	of	5,73	le	2,48	e	3,03	en	3,04
die	2,46	in	2,59	et	2,34	la	2,00	van	2,78
der	2,16	and	2,46	la	2,33	il	1,95	het	2,60
in	1,61	a	2,26	à	2,00	che	1,94	in	2,07
war	1,46	to	1,99	les	1,88	a	1,24	een	1,88
zu	1,32	is	1,81	l'	1,61	un	1,13	den	1,25
sich	1,30	as	0,87	il	1,41	non	1,11	ik	1,23
nicht	1,17	it	0,81	un	1,13	in	1,09	die	1,23
es	1,14	with	0,80	d'	1,05	del	1,04	te	1,14

Tab. B.2: Scores des 10 mots les plus fréquents de chaque langue obtenus à partir des données d'apprentissage.

Pour les besoins d'un TP de programmation, ce module a également été programmé en JAVA sous une forme simplifiée. Le sujet (identification automatique de la langue d'un texte) a été proposé aux étudiants en deuxième année de Licence MIASS (Mathématiques et Informatique Appliquées aux Sciences Sociales) de l'Université Pierre Mendès France de Grenoble.

B.3 Programmes d'analyse morphologique non supervisée

Nous avons développé deux programmes d'analyse non supervisée. Le premier procède par segmentation et le second par classification. Le fonctionnement de ces programmes est détaillé dans les chapitres 4 et 5.

Le système d'analyse morphologique par segmentation utilise les librairies externes suivantes, afin notamment d'optimiser les temps de traitement ainsi que l'utilisation mémoire :

– Graphes : `graph_lib` de Nathan Denny (`http://www.ece.arizona.edu/~denny/python_nest/graph_lib_1.0.1.html`).

– Arbres de recherche : `trie` de James Tauber (`http://jtauber.com/`).
– Arbres des suffixes : `SuffixTree` de Danny Yoo (`http://hkn.eecs.berkeley.edu/~dyoo/python/suffix_trees/`).
– Arbres de recherche ternaire : `pytst` de Nicolas Lehuen (`http://nicolas.lehuen.com/download/pytst/`).

Au total, le processus d'apprentissage complet, pour le corpus volcano-fr qui comprend près de 60 000 mots différents, avec N = 5, a = 0.8, b = 0.1 dure 4 minutes 47 secondes et consomme au plus 100 Mo de mémoire (au cours de l'étape 2), pour un ordinateur équipé d'un processeur à 1,73 GHz et 1 Go de RAM.

Annexe C

Glossaires pour l'évaluation des mots clés

Id	URL	Date d'accès	Nombre de termes
1	http ://www.breastcancer.org	05/06/06	1089
2	http ://www.komen.org	05/06/06	252
3	http ://www.breastcancercare.org.uk	05/06/06	41
4	http ://www.oncolink.com (Types of cancer)	05/06/06	46
5	http ://bca.ns.ca	05/06/06	508
6	http ://www.oncolink.com (Resources)	05/06/06	2048
7	http ://www.healthtalk.com	05/06/06	65
8	http ://www.meds.com	05/06/06	227
9	http ://www.medicinenet.com	05/06/06	127
10	http ://www.rd.com	05/06/06	66

Tab. C.2: Caractéristiques des glossaires pour l'évaluation des mots clés du corpus cancer-en.

Id	URL	Date d'accès	Nombre de termes
1	http ://vulcan.wr.usgs.gov	02/06/06	145
2	http ://volcano.und.edu	02/06/06	209
3	http ://www.swisseduc.ch	02/06/06	61
4	http ://www.volcanolive.com	06/06/06	421
5	http ://volcano.expert-answers.net	06/06/06	142
6	http ://www.avo.alaska.edu	02/06/06	138
7	http ://www.mvo.ms	05/06/06	46
8	http ://library.thinkquest.org	05/06/06	25
9	http ://interactive2.usgs.gov	05/06/06	42
10	http ://www.gns.cri.nz	05/06/06	50

TAB. C.4: Caractéristiques des glossaires pour l'évaluation des mots clés du corpus volcano-en.

Id	URL	Date d'accès	Nombre de termes
1	http ://www2.ligue-cancer.asso.fr	05/06/06	133
2	http ://www.europadonna.fr	05/06/06	155
3	http ://www.swisscancer.ch	05/06/06	338
4	http ://www.cancer.ca	05/06/06	23
5	http ://www.cancerdusein.org	05/06/06	27
6	http ://www.passeportsante.net	05/06/06	37
7	http ://www.hanys.org	05/06/06	49
8	http ://www.univ-lille3.fr	05/06/06	104
9	http ://www.canceronet.com	05/06/06	146
10	http ://www.ligue-cancer.net	05/06/06	11

TAB. C.6: Caractéristiques des glossaires pour l'évaluation des mots clés du corpus cancer-fr.

Id	URL	Date d'accès	Nombre de termes
1	http ://volcans.free.fr	22/05/06	29
2	http ://www.er.uqam.ca	22/05/06	282
3	http ://www.terminalf.net	22/05/06	81
4	http ://services.vulcania.com	22/05/06	81
5	http ://www.lave-volcans.com	22/05/06	98
6	http ://www.brgm.fr	02/06/06	43
7	http ://www.volcans-ardeche.com	02/06/06	60
8	http ://www.activolcans.info	06/06/06	76
9	http ://php.educanet2.ch	02/06/06	58
10	http ://www.volcan.dufouraubin.com	02/06/06	15

Tab. C.8: Caractéristiques des glossaires pour l'évaluation des mots clés du corpus volcano-fr.

Bibliographie

[Adamson et Boreham, 1974] Adamson, G. W. et Boreham, J. (1974). The use of an association measure based on character structure to identify semantically related pairs of words and document titles. *Information Storage and Retrieval*, 10:253–260.

[Amsili, 2003] Amsili, P. (2003). L'antonymie en terminologie : quelques remarques. *In Actes des cinquièmes rencontres Terminologie et Intelligence Artificielle (TIA 2003)*, pages 31–40.

[Ananiadou, 1994] Ananiadou, S. (1994). A methodology for automatic term recognition. *In Proceedings of the 15th conference on Computational linguistics*, pages 1034–1038, Morristown, NJ, USA. Association for Computational Linguistics.

[Argamon *et al.*, 2004] Argamon, S., Akiva, N., Amir, A. et Kapah, O. (2004). Efficient Unsupervised Recursive Word Segmentation Using Minimum Description Length. *In Proceedings of The 20th International Conference on Computational Linguistics (COLING)*, pages 1058–1064, Geneva, Switzerland.

[Baayen *et al.*, 1995] Baayen, R. H., Piepenbrock, R. et Gulikers, L. (1995). *The Celex Lexical Database (Release 2) [CD-ROM]*. Linguistic Data Consortium, Philadelphia, PA.

[Baayen et Schreuder, 2000] Baayen, R. H. et Schreuder, R. (2000). Towards a psycholinguistic computational model for morphological parsing. *Philosophical Transactions of the Royal Society (Series A : Mathematical, Physical and Engineering Sciences)*, pages 1–13.

[Baroni, 2003] Baroni, M. (2003). Distribution-driven morpheme discovery : A computational/experimental study. *Yearbook of Morphology 2003*, pages 213–248.

[Baroni, 2005] Baroni, M. (2005). Extracting only editorial content from a HTML page. Corpora List message.

[Baroni et Bernardini, 2004] Baroni, M. et Bernardini, S. (2004). BootCaT : Bootstrapping Corpora and Terms from the Web. *In* Lino, M. T., Xavier, M. F., Ferreira, F., Costa, R. et Silva, R., éditeurs : *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, pages 1313–1316, Lisbon, Portugal.

[Baroni *et al.*, 2002a] Baroni, M., Matiasek, J. et Trost, H. (2002a). Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. *In Proceedings of the ACL Workshop on Morphological and Phonological Learning 2002*, pages 48–57.

[Baroni *et al.*, 2002b] Baroni, M., Matiasek, J. et Trost, H. (2002b). Wordform- and class-based prediction of the components of German nominal compounds in an AAC system. *In Proceedings of the 19th international conference on Computational linguistics*, pages 1–7, Morristown, NJ, USA. Association for Computational Linguistics.

[Baroni et Sharoff, 2005] Baroni, M. et Sharoff, S. (2005). Creating Specialized and General Corpora Using Automated Search Engine Queries. Slides of the talk given at the Web as Corpus Workshop at Corpus Linguistics 2005.

[Baroni et Ueyama, 2006] Baroni, M. et Ueyama, M. (2006). Building general- and special-purpose corpora by Web crawling. *In Proceedings of the 13th NIJL International Symposium, Language Corpora : Their Compilation and Application*, pages 31–40.

[Bauer, 1998] Bauer, L. (1998). Is there a class of neoclassical compounds, and if so is it productive ? *Linguistics*, 36(3):403–422.

[Berger *et al.*, 2004] Berger, H., Dittenbach, M. et Merkl, D. (2004). An Accommodation Recommender System Based on Associative Networks. *In* Frew, A. J., éditeur : *Proceedings of the 11th International Conference on Information and Communication Technologies in Tourism (ENTER 2004)*, pages 216–227, Cairo, Egypt. Springer-Verlag.

[Bernhard, 2005] Bernhard, D. (2005). Segmentation morphologique à partir de corpus. *In Actes de TALN & RÉCITAL 2005*, volume 1, pages 555–564, Dourdan, France. ATALA. `http://taln.limsi.fr/site/talnRecital05/tome1/P65.pdf`.

[Bernhard, 2006a] Bernhard, D. (2006a). Automatic Acquisition of Semantic Relationships from Morphological Relatedness. *In* Salakoski, T., Ginter, F., Pyysalo, S. et Pahikkala, T., éditeurs : *Advances in Natural Language Processing, Proceedings of the 5th International Conference on NLP, FinTAL 2006*, volume 4139/2006 de *LNAI*, pages 121 – 132, Turku, Finland. Springer Berlin / Heidelberg. `http://dx.doi.org/10.1007/11816508_14`.

[Bernhard, 2006b] Bernhard, D. (2006b). Multilingual Term Extraction from Domain-specific Corpora Using Morphological Structure. *In Conference Companion of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, pages 171–174, Trento, Italy. `http://acl.ldc.upenn.edu/eacl2006/companion/pd/22_bernhard_11.pdf`.

[Bernhard, 2006c] Bernhard, D. (2006c). Unsupervised Morphological Segmentation Based on Segment Predictability and Word Segments Alignment. *In* Kurimo, M., Creutz, M. et Lagus, K., éditeurs : *Proceedings of the Pascal Challenges Workshop on the Unsupervised Segmentation of Words into Morphemes*, pages 19–23, Venice, Italy. `http://www.cis.hut.fi/morphochallenge2005/P03_Bernhard.pdf`.

[Bertels *et al.*, 2006] Bertels, A., Speelman, D. et Geeraerts, D. (2006). Analyse quantitative et statistique de la sémantique dans un corpus technique. *In Actes de TALN 2006*, pages 73–82, Leuven, Belgique.

[Bilotti *et al.*, 2004] Bilotti, M. W., Katz, B. et Lin, J. (2004). What Works Better for Question Answering : Stemming or Morphological Query Expansion. *In Proceedings of the Information Retrieval for Question Answering (IR4QA) Workshop at SIGIR 2004*, Sheffield, England.

[Bodenreider *et al.*, 2001] Bodenreider, O., Burgun, A. et Rindflesch, T. C. (2001). Lexically-suggested hyponymic relations among medical terms and their representation in the UMLS. *In Actes de la Quatrième rencontre Terminologie et Intelligence Artificielle (TIA'01)*, pages 11–21, Nancy, France.

[Bordag, 2005] Bordag, S. (2005). Unsupervised Knowledge-Free Morpheme Boundary Detection. *In Proceedings of RANLP (Recent Advances in Natural Language Processing) 2005*, Borovets, Bulgaria.

[Bordag, 2006] Bordag, S. (2006). Two-step Approach to Unsupervised Morpheme Segmentation. *In Proceedings of the Pascal Challenges Workshop on the Unsupervised Segmentation of Words into Morphemes*, pages 37–40, Venice, Italy.

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

[Brent et Cartwright, 1996] Brent, M. et Cartwright, T. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, 61:93–125.

[Brent, 1999] Brent, M. R. (1999). An Efficient, Probabilistically Sound Algorithm for Segmentation and Word Discovery. *Machine Learning*, 34:71–105.

[Buitelaar et Sacaleanu, 2002] Buitelaar, P. et Sacaleanu, B. (2002). Extending Synsets with Medical Terms. *In Proceedings of the First International WordNet Conference*, Mysore, India.

[Caraballo, 1999] Caraballo, S. A. (1999). Automatic Acquisition of a Hypernym-Labeled Noun Hierarchy from Text. *In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 120–126. Association for Computational Linguistics.

[Cederberg et Widdows, 2003] Cederberg, S. et Widdows, D. (2003). Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction. *In Proceedings of the Conference on Natural Language Learning (CoNLL)*, pages 111–118, Edmonton, Canada.

[Church et Hanks, 1990] Church, K. W. et Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1):22–29.

[Church et Mercer, 1993] Church, K. W. et Mercer, R. L. (1993). Introduction to the Special Issue on Computational Linguistics Using Large Corpora. *Computational Linguistics*, 19(1):1–24.

[Claveau et L'Homme, 2005] Claveau, V. et L'Homme, M.-C. (2005). Apprentissage par analogie pour la structuration de terminologie - Utilisation comparée de ressources endogènes et exogènes. *In Actes de la conférence Terminologie et Intelligence Artificielle, TIA'05*, Rouen, France.

[Colé *et al.*, 1989] Colé, P., Beauvillain, C. et Segui, J. (1989). On the representation and processing of prefixed and suffixed derived words : A differential frequency effect. *Journal of Memory and Language*, 28:1–13.

[Cottez, 1984] Cottez, H. (1984). *Dictionnaire des structures du vocabulaire savant. Éléments et modèles de formation.* Le Robert, Paris, 3rd édition.

[Creutz, 2003] Creutz, M. (2003). Unsupervised Segmentation of Words Using Prior Distributions of Morph Length and Frequency. *In Proceedings of ACL-03, the 41st Annual Meeting of the Association of Computational Linguistics*, pages 280–287, Sapporo, Japan.

[Creutz, 2006] Creutz, M. (2006). *Induction of the Morphology of Natural Language : Unsupervised Morpheme Segmentation with Application to Automatic Speech Recognition.* Thèse de doctorat, Helsinki Univeristy of Technology, Espoo, Finland.

[Creutz et Lagus, 2002] Creutz, M. et Lagus, K. (2002). Unsupervised Discovery of Morphemes. *In Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02*, pages 21–30.

[Creutz et Lagus, 2004] Creutz, M. et Lagus, K. (2004). Induction of a Simple Morphology for Highly-Inflecting Languages. *In Proceedings of the 7th Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, pages 43–51, Barcelona.

[Creutz et Lagus, 2005] Creutz, M. et Lagus, K. (2005). Inducing the Morphological Lexicon of a Natural Language from Unannotated Text. *In Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*, Espoo, Finland.

[Creutz et Lagus, 2006] Creutz, M. et Lagus, K. (2006). Morfessor in the Morpho Challenge. *In Proceedings of the Pascal Challenges Workshop on the Unsupervised Segmentation of Words into Morphemes*, pages 13–18, Venice, Italy.

[Creutz et Lindén, 2004] Creutz, M. et Lindén, K. (2004). Morpheme Segmentation Gold Standards for Finnish and English. Publications in computer and information science, report a77, Helsinki University of Technology.

[Cruse, 2000] Cruse, D. A. (2000). *Meaning in Language. An Introduction to Semantics and Pragmatics*. Oxford : University Press.

[Curran et Moens, 2001] Curran, J. R. et Moens, M. (2001). Scaling context space. *In ACL '02 : Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 231–238, Morristown, NJ, USA. Association for Computational Linguistics.

[Daille, 1996] Daille, B. (1996). Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. *In* Klavans, J. et Resnik, P., éditeurs : *The Balancing Act : Combining Symbolic and Statistical Approaches to Language*, pages 49–66. The MIT Press, Cambridge, Massachusetts.

[Daille, 2003] Daille, B. (2003). Conceptual structuring through term variations. *In* Bond, F., Korhonen, A., MacCarthy, D. et Villacicencio, A., éditeurs : *Proceedings of the ACL 2003 Workshop on Multiword Expressions : Analysis, Acquisition and Treatment*, pages 9–16.

[Daille *et al.*, 2002] Daille, B., Fabre, C. et Sébillot, P. (2002). *Many Morphologies*, chapitre Applications of Computational Morphology, pages 210–234. Cascadilla Press.

[Dal *et al.*, 2005] Dal, G., Hathout, N. et Namer, F. (2005). Morphologie Constructionnelle et Traitement Automatique des Langues : le projet MorTAL. *Lexique*, 16.

[de Chalendar et Grau, 2000] de Chalendar, G. et Grau, B. (2000). Svetlan' ou comment classer des noms en fonction de leur contexte. *In Actes de la 7è conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN 2000)*, pages 81–90, Lausanne, Suisse.

[de Saussure, 1916] de Saussure, F. (1995, première édition 1916). *Cours de linguistique générale*. Payot et Rivages.

[Déjean, 1998] Déjean, H. (1998). Morphemes as Necessary Concept for Structures Discovery from Untagged Corpora. *In* Powers, D., éditeur : *Proceedings of the CoNLL98 Workshop on Paradigms and Grounding in Language Learning*, pages 295–298.

[Demberg, 2006] Demberg, V. (2006). Letter-to-Phoneme Conversion for a German Text-to-Speech System. Mémoire de D.E.A., Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung. `http://homepages.inf.ed.ac.uk/s0455377/Diplomarbeit.pdf` [Accédé le 20/09/2006].

[Denhière et Lemaire, 2003] Denhière, G. et Lemaire, B. (2003). Modélisation des effets contextuels par l'analyse de la sémantique latente. *In* Bastien, J., éditeur : *Actes des Deuxièmes Journées d'étude en Psychologie Ergonomique (EPIQUE 2003)*, Roquencourt : INRIA.

[Drouin, 2003] Drouin, P. (2003). Acquisition des termes simples fondée sur les pivots lexicaux spécialisés. *In Actes des cinquièmes rencontres Terminologie et Intelligence Artificielle*, pages 183–186.

[Drouin, 2004] Drouin, P. (2004). Spécificités lexicales et acquisition de la terminologie. *In Actes des 7e Journées internationales d'analyse statistique des données textuelles (JADT-2004)*, pages 345–352, Louvain-la-Neuve, Belgique.

[Dunning, 1994] Dunning, T. (1994). Statistical identification of language. Techical Report MCCS-94-273, Computing Research Lab (CRL), New Mexico State University.

[Dunning, 1993] Dunning, T. E. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74.

[Déjean, 1998] Déjean, H. (1998). *Concepts et algorithmes pour la découverte des structures formelles des langues*. Thèse de doctorat, Université de Caen.

[Eiken *et al.*, 2006] Eiken, U. C., Liseth, A. T., Witschel, H. F., Richter, M. et Biemann, C. (2006). Ord i Dag : Mining Norwegian Daily Newswire. *In* Salakoski, T., Ginter, F., Pyysalo, S. et Pahikkala, T., éditeurs : *Advances in Natural Language Processing, 5th International Conference on NLP, FinTAL 2006, Turku, Finland, August 23-25, 2006, Proceedings*, volume 4139/2006 de *LNAI*, pages 512–523, Turku, Finland. Springer Berlin / Heidelberg.

[Elkateb-Gara, 2005] Elkateb-Gara, F. (2005). *L'organisation des connaissances, approches conceptuelles*, chapitre Extraction d'entités nommées pour la recherche d'informations précises, pages 73–82. L'Harmattan.

[Enguehard, 1992] Enguehard, C. (1992). *ANA, Apprentissage Naturel Automatique d'un Réseau Sémantique*. Thèse de doctorat, Université de Technologie de Compiègne.

[Enguehard, 1993] Enguehard, C. (1993). Acquisition de terminologie à partir de gros corpus. *In Informatique & Langue Naturelle, ILN'93, Nantes*, pages 373–384.

[Enguehard et Pantéra, 1995] Enguehard, C. et Pantéra, L. (1995). Automatic Natural Acquisition of a Terminology. *Journal of Quantitative Linguistics*, 2(1):27–32.

[Faure et Nédellec, 1999] Faure, D. et Nédellec, C. (1999). Knowledge Acquisition of Predicate Argument Structures from Technical Texts Using Machine Learning : the System ASIUM. *In* Fensel, D. et Studer, R., éditeurs : *Proceedings of the 11th European Workshop EKAW'99*, pages 329–334.

[Ferret, 2004] Ferret, O. (2004). Découvrir des sens de mots à partir d'un réseau de cooccurrences lexicales. *In Actes de TALN 2004*, Fès, Maroc.

[Finkelstein-Landau et Morin, 1999] Finkelstein-Landau, M. et Morin, E. (1999). Extracting Semantic Relationships between Terms : Supervised vs. Unsupervised Methods. *In Proceedings of the International Workshop on Ontological Engineering on the Global Information Infrastructure*, pages 71–80, Dagstuhl Castle, Germany.

[Finn *et al.*, 2001] Finn, A., Kushmerick, N. et Smyth, B. (2001). Fact or fiction : Content classification for digital libraries. *In Joint DELOS-NSF Workshop on Personalisation and Recommender Systems in Digital Libraries (Dublin)*.

[Flenner, 1994] Flenner, G. (1994). Ein quantitatives Morphsegmentierungssystem für spanische Wortformen. *Computatio Linguae II*, pages 31–62.

[Freitag, 2005] Freitag, D. (2005). Morphology Induction from Term Clusters. *In Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 128–135, Ann Arbor, Michigan. Association for Computational Linguistics.

[French et Labiouse, 2002] French, R. M. et Labiouse, C. (2002). Four Problems with Extracting Human Semantics from Large Text Corpora. *In Proceedings of the 24th Annual Conference of the Cognitive Science Society*.

[Gaussier, 1999] Gaussier, E. (1999). Unsupervised learning of derivational morphology from inflectional lexicons. *In Proceedings of the Workshop on Unsupervised Methods in Natural Language Processing*. University of Maryland.

[Gendner, 2002] Gendner, V. (2002). Racine / radical / base. `http://www.limsi.fr/Individu/gendner/LG.2001/base_radical_racine.html`.

[Ghani *et al.*, 2001] Ghani, R., Jones, R. et Mladenic, D. (2001). Mining the web to create minority language corpora. *In CIKM '01 : Proceedings of the tenth international conference on Information and knowledge management*, pages 279–286, New York, NY, USA. ACM Press.

[Giraudo et Grainger, 2000] Giraudo, H. et Grainger, J. (2000). Effects of prime word frequency and cumulative root frequency in masked morphological priming. *Language and Cognitive Processes*, 15(4/5):421–444.

[Goldsmith, 2001] Goldsmith, J. (2001). Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics*, 27(2):153–198.

[Goldwater et McClosky, 2005] Goldwater, S. et McClosky, D. (2005). Improving Statistical MT Through Morphological Analysis. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 676–683, Vancouver, British Columbia, Canada.

[Grabar et Hamon, 2006] Grabar, N. et Hamon, T. (2006). Terminology Structuring Through the Derivational Morphology. *In* Salakoski, T., Ginter, F., Pyysalo, S. et Pahikkala, T., éditeurs : *Advances in Natural Language Processing, 5th International Conference on NLP, FinTAL 2006, Turku, Finland, August 23-25, 2006, Proceedings*, volume 4139/2006 de *LNAI*, pages 652 – 663, Turku, Finland. Springer Berlin / Heidelberg.

[Grabar et Zweigenbaum, 1999] Grabar, N. et Zweigenbaum, P. (1999). Acquisition automatique de connaissances morphologiques sur le vocabulaire médical. *In* Amsili, P., éditeur : *Actes de TALN 1999*, pages 175–184, Cargèse.

[Grabar et Zweigenbaum, 2002a] Grabar, N. et Zweigenbaum, P. (2002a). Lexically-based terminology structuring : a feasibility study. *In Proceedings of the LREC Workshop on Using Semantics for Information Retrieval and Filtering*, pages 73–77, Las Palmas, Canaries.

[Grabar et Zweigenbaum, 2002b] Grabar, N. et Zweigenbaum, P. (2002b). Lexically-based terminology structuring : Some inherent limits. *In* Chien, L.-F., Daille, B., Kageura, K. et Nakagawa, H., éditeurs : *Proceedings of the Second International Workshop on Computational Terminology (COMPUTERM 2002)*, pages 36–42.

[Grefenstette, 1995] Grefenstette, G. (1995). Comparing Two Language Identification Schemes. *In Proceedings of the 3rd International Conference on the Statistical Analysis of Textual Data (JADT 95)*.

[Grefenstette, 1996] Grefenstette, G. (1996). Evaluation Techniques for Automatic Semantic Extraction : Comparing Syntactic and Window Based Approaches. *Corpus Processing for Lexical Acquisition*, pages 205–216.

[Grefenstette, 1998] Grefenstette, G. (1998). Re : Corpora : Sentence splitting. Corpora List. `http://torvald.aksis.uib.no/corpora/1998-4/0035.html`.

[Grefenstette et Tapanainen, 1994] Grefenstette, G. et Tapanainen, P. (1994). What is a word, What is a sentence ? Problems of Tokenization. *In 3rd International Conference on Computational Lexicography (COMPLEX'94)*, Budapest, Hungary.

[Habert et Zweigenbaum, 2002] Habert, B. et Zweigenbaum, P. (2002). Contextual acquisition of information categories : what has been done and what can be done automatically ? *The Legacy of Zellig Harris : Language and information into the 21st Century - Vol. 2. Mathematics and computability of language*, pages 203–231.

[Hacioglu *et al.*, 2003] Hacioglu, K., Pellom, B., Ciloglu, T., Ozturk, O., Kurimo, M. et Creutz, M. (2003). On lexicon creation for Turkish LVCSR. *In Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech)*, pages pp. 1165–1168., Geneva, Switzerland.

[Hafer et Weiss, 1974] Hafer, M. A. et Weiss, S. F. (1974). Word segmentation by letter successor varieties. *Information Storage and Retrieval*, 10:371–385.

[Hahn *et al.*, 2003] Hahn, U., Honeck, M. et Shulz, S. (2003). Subword-Based Text Retrieval. *In Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03)*, Big Island, Hawaii.

[Harris, 1955] Harris, Z. (1955). From phoneme to morpheme. *Language*, 31(2):190–222.

[Hathout, 2002] Hathout, N. (2002). From WordNet to CELEX : acquiring morphological links from dictionaries of synonyms. *In Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 1478–1484, Las Palmas de Gran Canaria, Spain. ELRA.

[Hathout, 2005] Hathout, N. (2005). Exploiter la structure analogique du lexique construit : une approche computationnelle. *Cahiers de Lexicologie*, 87(2).

[Hayes, 1999] Hayes, B. (1999). The Web of Words. *American Scientist*, 87(2):108–112.

[Hearst, 1992] Hearst, M. A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. *In Proceedings of the Fourteenth International Conference on Computational Linguistics, Nantes, France, July 1992*, pages 539–545.

[Heid, 1998] Heid, U. (1998). A linguistic bootstrapping approach to the extraction of term candidates from German text. *Terminology*, 5(2):161–181.

[Heyer *et al.*, 2001] Heyer, G., Läuter, M., Quasthoff, U., Wittig, T. et Wolff, C. (2001). Learning Relations Using Collocations. *In Proceedings of the IJCAI Workshop on Ontology Learning*, Seattle / WA.

[Heyer *et al.*, 2006] Heyer, G., Quasthoff, U. et Wittig, T. (2006). *Text Mining : Wissensrohstoff Text. Konzepte, Algorithmen, Ergebnisse.* W3L -Verlag, Bochum.

[Hindle, 1990] Hindle, D. (1990). Noun Classification from Predicate-Argument Structures. *In Proceddings of ACL-80, Pittsburgh, PA*, pages 268–275.

[Honkela *et al.*, 1995] Honkela, T., Pulkki, V. et Kohonen, T. (1995). Contextual Relations of Words in Grimm Tales, Analyzed by Self-Organizing Map. *In* Fogelman-Soulie, F. et Gallinari, P., éditeurs : *Proceedings of the International Conference on Artificial Neural Networds, ICANN-95*, pages 3–7, Paris.

[Ibekwe-SanJuan, 1998] Ibekwe-SanJuan, F. (1998). Terminological variation, a means of identifying research topics from texts. *In Proceedings of the Joint International Conference on Computational Linguistics (COLING-ACL'98)*, pages 564–570, Montréal, Québec.

[Ibekwe-SanJuan, 2005] Ibekwe-SanJuan, F. (2005). Inclusion lexicale et proximité sémantique entre termes. *In Actes des Sixièmes rencontres Terminologie et Intelligence Artificielle (TIA'05)*, pages 45–57, Rouen, France.

[Ibekwe-Sanjuan et Sanjuan, 2004] Ibekwe-Sanjuan, F. et Sanjuan, E. (2004). Mining Textual Data through Term Variant Clustering : the TermWatch system. *Recherche d'Information Assistée par Ordinateur (RIAO 2004) "Coupling approaches, coupling media and coupling languages for information retrieval"*, 26:487–503.

[Jacquemin, 1997] Jacquemin, C. (1997). Guessing morphology from terms and corpora. *In Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 156 – 165.

[Jacquemin et Bourigault, 2003] Jacquemin, C. et Bourigault, D. (2003). *The Oxford Handbook of Computational Linguistics*, chapitre Term Extraction and Automatic Indexing. Oxford University Press.

[Janssen, 1992] Janssen, A. (1992). Segmentierung französicher Wortformen in Morphe ohne Verwendung eines Lexikons. *Computatio Linguae*, pages 74–95.

[Ji, 2004] Ji, H. (2004). *Étude d'un modèle computationnel pour la représentation du sens des mots par intégration des relations de contexte*. Thèse de doctorat, Institut des Sciences Cognitives, Institut National Polytechnique de Grenoble - INPG.

[Kageura et Umino, 1996] Kageura, K. et Umino, B. (1996). Methods of Automatic Term Recognition - A Review. *Terminology*, 3(2).

[Kazakov, 1997] Kazakov, D. (1997). Unsupervised learning of naive morphology with genetic algorithms. *In* Daelemans, W., van den Bosch, A. et Weijters, A., éditeurs : *Workshop Notes of the ECML/MLnet Workshop on Empirical Learning of Natural Language Processing Tasks*, pages 105–112, Prague.

[Keshava et Pitler, 2006] Keshava, S. et Pitler, E. (2006). A Simpler, Intuitive Approach to Morpheme Induction. *In Proceedings of the Pascal Challenges Workshop on the Unsupervised Segmentation of Words into Morphemes*, pages 31–35, Venice, Italy.

[Kilgarriff et Grefenstette, 2003] Kilgarriff, A. et Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29(3):333–347.

[Kit et Wilks, 1999] Kit, C. et Wilks, Y. (1999). Unsupervised Learning of Word Boundary with Description Length Gain. *In Proceedings od the CoNLL99 (Computational Natural Language Learning) ACL Workshop*, Bergen.

[Kittredge, 2003] Kittredge, R. I. (2003). *The Oxford Handbook of Computational Linguistics*, chapitre Sublanguages and Controlled Languages, pages 430–447. Oxford University Press.

[Kleiber et Tamba, 1990] Kleiber, G. et Tamba, I. (1990). L'hyperonymie revisitée : inclusion et hiérarchie. *Langages*, 98:7–32.

[Klenk, 1992] Klenk, U. (1992). Verfahren morphologischer Segmentierung und die Wortstruktur im Spanischen. *Computatio Linguae*.

[Kohonen *et al.*, 2000] Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V. et Saarela, A. (2000). Self organization of a massive document collection. *IEEE Transactions on neural networks*, 11(3):574–585.

[Koskenniemi, 1984] Koskenniemi, K. (1984). A general computational model for word-form recognition and production. *In Proceedings of the 22nd annual meeting on Association for Computational Linguistics*, pages 178–181, Morristown, NJ, USA. Association for Computational Linguistics.

[Krovetz, 1993] Krovetz, R. (1993). Viewing morphology as an inference process. *In SIGIR '93 : Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 191–202, New York, NY, USA. ACM Press.

[Kurimo *et al.*, 2006] Kurimo, M., Creutz, M., Varjokallio, M., Arisoy, E. et Saraclar, M. (2006). Unsupervised segmentation of words into morphemes – Challenge 2005 : An Introduction and Evaluation Report. *In Proceedings of the Pascal Challenges Workshop on the Unsupervised Segmentation of Words into Morphemes*, pages 1–11, Venice, Italy.

[Labbé et Labbé, 2001] Labbé, C. et Labbé, D. (2001). Que mesure la spécificité du vocabulaire ? *Lexicometrica*, 3.

[Lafon, 1980] Lafon, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, 1:127–165.

[Lafourcade et Prince, 2001] Lafourcade, M. et Prince, V. (2001). Synonymies et vecteurs conceptuels. *In Actes de TALN 2001*, pages 233–242, Tours, France.

[Lagus *et al.*, 2002] Lagus, K., Airola, A. et Creutz, M. (2002). Data analysis of conceptual similarities of Finnish Verbs. *In Proceedings of Cog Sci 2002, the 24th annual meeting of the Cognitive Science Society.*

[Landauer *et al.*, 1998] Landauer, T. K., Laham, D. et Foltz, P. (1998). Learning Human-like Knowledge by Singular Value Decomposition : A Progress Report. *Advances in Neural Information Processing Systems*, 10:45–51.

[Le Pesant et Mathieu-Colas, 1998] Le Pesant, D. et Mathieu-Colas, M. (1998). Introduction aux classes d'objets. *Langages*, (131):6–33.

[Le Priol, 2001] Le Priol, F. (2001). Identification, interprétation et représentation de relations sémantiques entre concepts. *In Actes de TALN 2001*, Tours, France.

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

[Lee, 2004] Lee, Y.-S. (2004). Morphological Analysis for Statistical Machine Translation. *In* Susan Dumais, D. M. et Roukos, S., éditeurs : *HLT-NAACL 2004 : Short Papers*, pages 57–60, Boston, Massachusetts, USA. Association for Computational Linguistics.

[Lemaire et Denhière, 2004] Lemaire, B. et Denhière, G. (2004). Incremental Construction of an Associative Network from a Corpus. *In Proceedings of the 26th Annual Meeting of the Cognitive Science Society (CogSci'2004), Chicago*, pages 825–830.

[Léon et Millon, 2005] Léon, S. et Millon, C. (2005). Acquisition semi-automatique de relations lexicales bilingues (français-anglais) à partir du Web. *In Actes de RECITAL 2005*, volume 1, pages 595–604.

[Lepage, 1998] Lepage, Y. (1998). Solving analogies on words : an algorithm. *In Proceedings of the 17th international conference on Computational Linguistics*, volume 1, pages 728–734, Morristown, NJ, USA. Association for Computational Linguistics.

[Lepage, 2003] Lepage, Y. (2003). *De l'analogie rendant compte de la commutation en linguistique.* Hdr, Université Joseph Fourier - Grenoble 1.

[Levy et Bullinaria, 2001] Levy, J. P. et Bullinaria, J. A. (2001). Learning Lexical Properties from Word Usage Patterns : Which Context Words Should be Used ? *In* French, R. et Sougne, J., éditeurs : *Connectionist Models of Learning, Development and Evolution : Proceedings of the Sixth Neural Computation and Psychology Workshop*, pages 273–282.

[Li *et al.*, 2000] Li, P., Burgess, C. et Lund, K. (2000). The Acquisition of Word Meaning through Global Lexical Co-occurrences. *In* clark, E. V., éditeur : *Proceedings of the Thirtieth Annual Child Language Research Forum*, pages 167–178.

[Light, 1996] Light, M. (1996). Morphological cues for lexical semantics. *In Proceedings of the 34th annual meeting of the Association for Computational Linguistics*, pages 25–31. Association for Computational Linguistics.

[Lin, 1998] Lin, D. (1998). Automatic retrieval and clustering of similar words. *In Proceedings of the 17th international conference on Computational linguistics*, pages 768–774, Morristown, NJ, USA. Association for Computational Linguistics.

[Lin et Murphy, 2001] Lin, E. L. et Murphy, G. L. (2001). Thematic Relations in Adults' Concepts. *Journal of Experimental Psychology*, 130(1):3–28.

[Liu et Curran, 2006] Liu, V. et Curran, J. R. (2006). Web Text Corpus for Natural Language Processing. *In Proceeding of EACL 2006, 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 233–240, Trento, Italy.

[Lovis *et al.*, 1995] Lovis, C., Michel, P.-A., Baud, R. et Scherrer, J.-R. (1995). Word Segmentation Processing : A Way to Exponentially Extend Medical Dictionaries. *In* Greenes, R. A., Peterson, H. E. et Protti, D. J., éditeurs : *Proc 8th Word Congress on Medical Informatics*, pages 28–32.

[Lund et Burgess, 1996] Lund, K. et Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments & Computers*, 28(2):203–208.

[Manning et Schütze, 1999] Manning, C. D. et Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

[Martinez, 2000] Martinez, W. (2000). Mise en évidence de rapports synonymiques par la méthode des cooccurrences. *In Actes de JADT 2000 : 5es Journées Internationales d'Analyse Statistique des Données Textuelles*.

[Matsumoto, 2003] Matsumoto, Y. (2003). *The Oxford Handbook of Computational Linguistics*, chapitre Lexical Knowledge Acquisition, pages 395–413. Oxford University Press.

[McKinnon *et al.*, 2003] McKinnon, R., Allen, M. et Osterhout, L. (2003). Morphological decomposition involving non-productive morphemes : ERP evidence. *Cognitive Neuroscience and Neuropsychology*, 14(6):883–886.

[McNamee, 2005] McNamee, P. (2005). Language identification : a solved problem suitable for undergraduate instruction. *Journal of Computing Sciences in Colleges*, 20(3):94–101.

[Meunier, 2003] Meunier, F. (2003). La notion de productivité morphologique : modèles psycholinguistiques et données expérimentales. *Langue Française*, 140:24 – 37.

[Meunier et Segui, 1999] Meunier, F. et Segui, J. (1999). Frequency Effects in Auditory Word Recognition : The Case of Suffixed Words. *Journal of Memory and Language*, 41:327–344.

[Miller, 1995] Miller, G. A. (1995). WordNet : a lexical database for English. *Communications of the ACM*, 38(11):39–41.

[Moreau et Claveau, 2006] Moreau, F. et Claveau, V. (2006). Extension de requêtes par relations morphologiques acquises automatiquement. *In Actes de la Troisième Conférence en Recherche d'Informations et Applications CORIA 2006*, pages 181–192.

[Morin, 1998] Morin, E. (1998). Prométhée, un outil d'aide à l'acquisition de relations sémantiques entre termes. *In Actes de la conférence TALN 1998*, pages 172 ?– 181, Paris, France.

[Naets, 2005] Naets, H. (2005). La Déclaration Universelle des Droits de l'Homme : 329 langues pour la constitution automatique de corpus et de lexique. *In Actes de TALN 2005*, volume 2, pages 261–268.

[Namer, 2002] Namer, F. (2002). Acquisition automatique de sens à partir d'opérations morphologiques en français. *In Actes de TALN 2002*, pages 235–244, Nancy.

[Namer, 2003] Namer, F. (2003). Automatiser l'analyse morpho-sémantique non affixale : le système DériF. *Cahiers de Grammaire*, 28:31–48.

[Namer, 2005] Namer, F. (2005). Morphosémantique pour l'appariement de termes dans le vocabulaire médical : approche multilingue. *In Actes de TALN 2005*, pages 63–72.

[Namer et Zweigenbaum, 2004] Namer, F. et Zweigenbaum, P. (2004). Acquiring meaning for French medical terminology : contribution of morphosemantics. *In Proceedings of Medinfo. 2004*, volume 11, pages 535–539, San Francisco CA.

[National Cancer Institute, Office of Communications and Center for Bioinformatics, 2006] National Cancer Institute, Office of Communications and Center for Bioinformatics (2006). NCI Thesaurus. `ftp://ftp1.nci.nih.gov/pub/cacore/EVS`. [Online ; accessed 23 March 2006].

[Neuvel, 2002a] Neuvel, S. (2002a). Vive la différence ! *Folia Linguistica*, 35(3-4):313–320.

[Neuvel, 2002b] Neuvel, S. (2002b). Whole Word Morphologizer. Expanding the Word-Based Lexicon : A non-stochastic computational approach. *Brain and Language*, 81:454–463.

[Neuvel et Fulop, 2002] Neuvel, S. et Fulop, S. A. (2002). Unsupervised Learning of Morphology Without Morphemes. *In Proceedings of the ACL Workshop on Morphological and Phonological Learning 2002*, pages 31–40.

[Nguyen et Murphy, 2003] Nguyen, S. P. et Murphy, G. (2003). An Apple is More Than Just a Fruit : Cross-Classification in Children's Concepts. *Child Development*, 74:1783–1806.

[Ninova *et al.*, 2005] Ninova, G., Nazarenko, A., Hamon, T. et Szulman, S. (2005). Comment mesurer la couverture d'une ressource terminologique pour un corpus ? *In Actes de TALN - RÉCITAL 2005*, pages 293–302, Dourdan, France.

[Oakes, 1998] Oakes, M. P. (1998). *Statistics for Corpus Linguistics*. Edinburgh : Edinburgh University Press.

[Pantel et Lin, 2001] Pantel, P. et Lin, D. (2001). A Statistical Corpus-Based Term Extractor. *AI 2001, Lecture Notes in Artificial Intelligence*, pages 36–46.

[Perruchet et Peereman, 2005] Perruchet, P. et Peereman, R. (2005). Apprendre sa langue maternelle, une question de statistique ? *Pour la science*, 327:82–85.

[Péry-Woodley, 1995] Péry-Woodley, M.-P. (1995). Quels corpus pour quels traitements automatiques ? *T.A.L.*, 36(1-2):213–232.

[Plaut et Gonnerman, 2000] Plaut, D. C. et Gonnerman, L. M. (2000). Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing ? *Language and Cognitive Processes*, 4/5(15):445–485.

[Ploux et Victorri, 1998] Ploux, S. et Victorri, B. (1998). Construction d'espaces sémantiques à l'aide de dictionnaires informatisés des synonymes. *TAL*, 39:161–182.

[Pointal, 2004] Pointal, L. (2004). Tree Tagger Wrapper. `http://www.limsi.fr/Individu/pointal/python/treetaggerwrapper-doc/`. [Online ; accessed 13 June 2006].

[Porter, 1980] Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

[Porter, 2001] Porter, M. F. (2001). Snowball : A language for stemming algorithms. `http://snowball.tartarus.org/texts/introduction.html`.

[Pratt et Pacak, 1969] Pratt, A. W. et Pacak, M. G. (1969). Automated processing of medical English. *In Proceedings of the 1969 conference on Computational linguistics*, pages 1–23, Morristown, NJ, USA. Association for Computational Linguistics.

[Quasthoff *et al.*, 2006] Quasthoff, U., Richter, M. et Biemann, C. (2006). Corpus Portal for Search in Monolingual Corpora. *In Proceedings of the fifth international conference on Language Resources and Evaluation, LREC 2006*, pages 1799–1802, Genoa, Italy.

[Rapp, 2003] Rapp, R. (2003). Discovering the Meaning of an Ambiguous Word by Searching for Sense Descriptors with Complementary Context Patterns. *In Actes des Cinquièmes Rencontres Terminologie et Intelligence Artificielle*, pages 145–155.

[Rastier, 1995] Rastier, F. (1995). Le terme : entre ontologie et linguistique. *La banque des mots*, (7):35–65.

[Rastle *et al.*, 2000] Rastle, K., Davis, M. H., Marslen-Wilson, W. D. et Tyler, L. K. (2000). Morphological and semantic effects in visual word recognition : A time-course study. *Language and Cognitive Processes*, 4/5(15):507–537.

[Rayson et Garside, 2000] Rayson, P. et Garside, R. (2000). Comparing corpora using frequency profiling. *In Proceedings of the workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000)*, pages 1–6, Hong Kong.

[Rebeyrolle, 2000] Rebeyrolle, J. (2000). Utilisation de contextes définitoires pour l'acquisition de connaissances à partir de textes. *In Actes de la conférence IC'2000, Journées Francophones d'Ingénierie de la Connaissance, Toulouse, IRIT*, pages 105–114.

[Rey-Debove, 1984] Rey-Debove, J. (1984). Le domaine de la morphologie lexicale. *Cahiers de lexicologie*, 45:3–19.

[Roche, 2005] Roche, C. (2005). Terminologie et Ontologie. *Langages*, 157:48–62.

[Roget, 1911] Roget, P. M. (1911). *Roget's Thesaurus*. `http://www.gutenberg.org/etext/22`.

[Rousselot, 2004] Rousselot, F. (2004). L'outil de traitement de corpus LIKES. *In Actes de TALN 2004*.

[Rytting, 2004] Rytting, C. A. (2004). Segment Predictability as a Cue in Word Segmentation : Application to Modern Greek. *In Proceedings of the Seventh Meeting of the ACL Special Interest Group in Computational Phonology*, pages 78–85, Barcelona, Spain. Association for Computational Linguistics.

[Saffran *et al.*, 1996] Saffran, J. R., Newport, E. L. et Aslin, R. N. (1996). Word Segmentation : The Role of Distributional Cues. *Journal of Memory and Language*, 35(4):606–621.

[Salton et McGill, 1983] Salton, G. et McGill, M. J. (1983). *Introduction to modern information retrival*. McGraw-Hill, New York.

[Schone et Jurafsky, 2000] Schone, P. et Jurafsky, D. (2000). Knowledge-Free Induction of Morphology Using Latent Semantic Analysis. *In Proceedings of the Fourth Conference on Computational Natural Language Learning and of the Second Learning Language in Logic Workshop*, Lisbon, Portugal.

[Schone et Jurafsky, 2001] Schone, P. et Jurafsky, D. (2001). Knowledge-Free Induction of Inflectional Morphologies. *In Proceedings of the Second meeting of the North American Chapter of the Association for Computational Linguistics*, pages 1–9.

[Schwab *et al.*, 2005] Schwab, D., Lafourcade, M. et Prince, V. (2005). Extraction semi-supervisée de couples d'antonymes grâce à leur morphologie. *In Actes de TALN 2005*, pages 73–82.

[Seidenberg, 1997] Seidenberg, M. S. (1997). Language Acquisition and Use : Learning and Applying Probabilistic Constraints. *Science*, 275.

[Seidenberg et Gonnerman, 2000] Seidenberg, M. S. et Gonnerman, L. M. (2000). Explaining Derivational Morphology As The Convergence of Codes. *Trends in Cognitive Sciences*, (4): 353–361.

[Singh, 1999] Singh, S. (1999). *Histoire des codes secrets. De l'Égypte des Pharaons à l'ordinateur quantique*. Jean-Claude Lattès.

[Sterling, 2005] Sterling, B. (2005). Order Out of Chaos. Wired Magazine, Issue 13.04. `http://www.wired.com/wired/archive/13.04/view.html?pg=4` [Online ; accessed 28-September-2006].

[Streiter *et al.*, 2003] Streiter, O., Zielinski, D., Ties, I. et Voltmer, L. (2003). Term Extraction for Ladin : An Example-based Approach. *In Actes de l'Atelier Traitement automatique des langues minoritaires et des petites langues à TALN*, Batz-sur-Mer, France.

[Taft et Forster, 1975] Taft, M. et Forster, K. (1975). Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior*, 14:638–647.

[ten Hacken et Lüdeling, 2002] ten Hacken, P. et Lüdeling, A. (2002). Word Formation in Computational Linguistics. *In Proceedings of TALN 2002*, volume 2, pages 61–87.

[Trost, 2003] Trost, H. (2003). *The Oxford Handbook of Computational Linguistics*, chapitre Morphology, pages 25–47. Oxford University Press.

[Turney, 2001] Turney, P. (2001). Answering Subcognitive Turing Test Questions : a Reply to French. *J. Experimental and Theoretical Artificial Intelligence*, 13:409–419.

[Tzoukermann *et al.*, 2003] Tzoukermann, E., Klavans, J. L. et Strzalkowski, T. (2003). *The Oxford Handbook of Computational Linguistics*, chapitre Information Retrieval, pages 529–544. Oxford University Press.

[van den Bosch et Daelemans, 1999] van den Bosch, A. et Daelemans, W. (1999). Memory-based morphological analysis. *In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 285–292, Morristown, NJ, USA. Association for Computational Linguistics.

[Vaufreydaz, 2002] Vaufreydaz, D. (2002). *Modélisation statistique du langage à partir d'Internet pour la reconnaissance automatique de la parole continue*. Thèse de doctorat, Université Joseph Fourier – Grenoble 1.

[Vergne, 2003] Vergne, J. (2003). Un outil d'extraction terminologique endogène et multilingue. *In Actes de TALN 2003*, volume 2, pages 139–148.

[Vergne, 2005] Vergne, J. (2005). Une méthode indépendante des langues pour indexer les documents de l'internet par extraction de termes de structure contrôlée. *In Actes de la Conférence Internationale sur le Document Électronique (CIDE 8)*, Beyrouth, Liban.

[Véronis, 2003] Véronis, J. (2003). Hyperlex : cartographie lexicale pour la recherche d'informations. *In Actes de la Conférence Traitement Automatique des Langues (TALN'2003)*, pages 265–274, Batz-sur-mer (France). ATALA.

[Véronis, 2004] Véronis, J. (2004). Hyperlex : lexical cartography for information retrieval. *Computer, Speech and Language*, 18(3):223–252.

[Véronis, 2006] Véronis, J. (2006). Le Nébuloscope. http://aixtal.blogspot.com/2006/01/outil-le-nbuloscope.html [Online ; accessed 21-September-2006].

[Vossen, 2003] Vossen, P. (2003). *The Oxford Handbook of Computational Linguistics*, chapitre Ontologies, pages 464–482. Oxford University Press.

[Wandmacher, 2005] Wandmacher, T. (2005). How semantic is Latent Semantic Analysis ? *In Actes de RECITAL 2005*, pages 525–534, Dourdan.

[Widdows et Dorow, 2002] Widdows, D. et Dorow, B. (2002). A Graph Model for Unsupervised Lexical Acquisition. *In Proceedings of the 19th International Conference on Computational Linguistics (COLING 19)*, pages 1093–1099, Taipei.

[Wikipedia, 2006a] Wikipedia (2006a). Folksonomy — Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/w/index.php?title=Folksonomy\&oldid=77787700 [Online ; accessed 28-September-2006].

[Wikipedia, 2006b] Wikipedia (2006b). Tag (metadata) — Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/w/index.php?title=Tag_%28metadata%29\&oldid=76258361 [Online ; accessed 21-September-2006].

[Wisniewski et Bassok, 1999] Wisniewski, E. J. et Bassok, M. (1999). What Makes a Man Similar to a Tie ? Stimulus Compatibility with Comparison and Integration. *Cognitive Psychology*, 39:208–238.

[Witschel, 2005] Witschel, H. (2005). Text, Wörter, Morpheme - Möglichkeiten einer automatischen Terminologie-Extraktion. *In* Fisseni, B., Schmitz, H.-C., Schröder, B. et Wagner, P., éditeurs : *Proc. of GLDV-Tagung 2005*, pages 659–672, Frankfurt. Peter Lang.

[Witschel et Biemann, 2006] Witschel, H. F. et Biemann, C. (2006). Rigorous dimensionality reduction through linguistically motivated feature selection for text categorization. *In* Werner, S., éditeur : *Proceedings of the 15th NODALIDA conference, Joensuu 2005*, volume 1 de *Ling@JoY : University of Joensuu electronic publications in linguistics and language technology*, pages 197–204, Joensuu, Finland.

[Xu et Croft, 1998] Xu, J. et Croft, W. B. (1998). Corpus-based stemming using cooccurrence of word variants. *ACM Transactions on Information Systems (TOIS)*, 16(1):61–81.

[Zipf, 1968] Zipf, G. K. (1968). *The Psycho-biology of Language. An Introduction to Dynamic Philology.* The M.I.T. Press, Cambridge, second paperback printing (first edition : 1935) édition.

[Zweigenbaum et Grabar, 2000] Zweigenbaum, P. et Grabar, N. (2000). Liens morphologiques et structuration de terminologie. *In Actes de IC 2000 : Ingénierie des Connaissances*, pages 325–334.

[Zweigenbaum *et al.*, 2003] Zweigenbaum, P., Hadouche, F. et Grabar, N. (2003). Apprentissage de relations morphologiques en corpus. *In* Daille, B., éditeur : *Actes de TALN 2003*, pages 285–294, Batz-sur-mer, France.

Résumé

Les ressources lexico-sémantiques, telles que les thésaurus, les terminologies ou les ontologies, visent à organiser les connaissances en rendant explicites divers types de relations sémantiques comme la synonymie ou la spécialisation. Le coût de la construction manuelle de telles ressources reste élevé, ce qui explique l'essor des méthodes d'acquisition automatique de connaissances, allant de l'extraction des termes représentant les unités de connaissance à l'identification des relations sémantiques qui les relient. Nous nous intéressons dans cette thèse au rôle que peut jouer la morphologie, c'est-à-dire la structure interne des mots, pour l'acquisition de telles connaissances à partir de corpus de textes de spécialité, essentiellement médicaux, et dans une perspective multilingue.

Nous présentons deux systèmes d'acquisition de connaissances morphologiques non supervisés, caractérisés par des approches différentes. Le premier procède par segmentation des mots, tandis que le second regroupe les mots dans des familles morphologiques.

Nous explorons ensuite les utilisations possibles de ce type d'informations pour l'acquisition de termes et de relations sémantiques. Nous proposons notamment une méthode de pondération et de visualisation des mots clés extraits de corpus de textes de spécialité en fonction de leur famille morphologique. Nous définissons également des schémas, basés sur les résultats de la segmentation morphologique, afin de découvrir des relations sémantiques telles que la spécialisation et la cohyponymie.

Mots-clés: Morphologie, domaines spécialisés, corpus, terminologie, relations sémantiques.

Abstract

Lexico-semantic resources, like thesauri, terminologies and ontologies, aim at organising knowledge by detailing semantic relationships such as synonymy or specialisation. The cost for manually building this kind of resources is high. Methods for the automatic acquisition of knowledge from text corpora are therefore widely used. These methods aim at automatically extracting terms and semantic relationships. In this thesis, we investigate the role which can be played by morphology, i.e. the internal structure of words, within such systems.

We describe two methods for the unsupervised acquisition of morphological knowledge. The first one segments words into sub-units while the other conflates words in morphological families.

We then explore possible uses for this kind of knowledge. We re-use morphological families to weight and visualise keywords. We also define patterns based on morphological segmentation which make it possible to discover semantic relationships such as hypernymy and co-hyponymy.

Keywords: Morphology, specialised domains, corpus, terminology, semantic relationships.