Extraction lexicale bilingue à partir de textes médicaux comparables : application à la recherche d'information translangue

Yun-Chuang Chiao

THÈSE DE DOCTORAT DE L'UNIVERSITÉ PARIS 6

Spécialité

Informatique Médicale

Présentée par

YUN-CHUANG CHIAO

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ PARIS 6

Sujet de la thèse :

Extraction lexicale bilingue à partir de textes médicaux comparables : application à la recherche d'information translangue

soutenue le 30 juin 2004

devant le jury composé de :

Christian Fluhr	Rapporteur	Directeur de recherche, CEA et Professeur, INSTN
Éric Gaussier	Examinateur	Ingénieur, Rank Xerox Center Europe
Patrick Gallinari	Examinateur	Professeur, Université Paris VI
Benoît Habert	Rapporteur	Professeur, Université Paris X et LIMSI-CNRS
Pierre Le Beux	Examinateur	Professeur, Université de Rennes
Jean-David Sta	Co-directeur	Chercheur, EDF R&D
Pierre Zweigenbaum	Directeur	Ingénieur, AP-HP et Professeur, INaLCO

Résumé

L'accroissement explosif des connaissances dans le domaine médical et l'inflation textuelle et multilingue, notamment sur le Web, confèrent à l'accès, l'exploitation ou la traduction de ces informations un enjeu important. Ces traitements nécessitent des ressources lexicales multilingues qui font partiellement défaut. L'actualisation de ces ressources multilingues est donc une problématique clé dans l'accès à ces informations.

Les travaux présentés ici ont été réalisés dans le cadre de l'extraction de lexique bilingue spécialisé à partir de textes médicaux comparables. L'objectif est d'évaluer et de proposer un outil d'aide à l'actualisation de lexique bilingue spécialisé et à la recherche d'information translangue en s'appuyant sur l'exploitation de ressources bilingues provenant du Web dans le domaine médical.

Nous présentons un modèle fondé sur l'analyse distributionnelle en introduisant à cette occasion une nouvelle notion que nous nommons *symétrie distributionnelle*. En général, les modèles classiques d'extraction de lexique bilingue à partir de corpus comparables établissent la relation de traduction entre deux mots en calculant la ressemblance entre leurs distributions d'une langue vers l'autre (par exemple, du français vers l'anglais). L'hypothèse de symétrie distributionnelle postule que la ressemblance des distributions de deux mots dans les deux directions de langues est un critère fort du lien traductionnel entre ces mots.

Deux grandes applications de ce modèle ont été expérimentées afin de le valider. Il s'agit de l'extraction d'un lexique bilingue médical (français-anglais) et de la recherche d'information translangue. Dans le cas de l'extraction lexicale bilingue, les résultats montrent que la prise en compte de la symétrie distributionnelle améliore la performance de manière significative par rapport aux modèles classiques. Dans le cas de la recherche d'information translangue, notre modèle a été appliqué pour traduire et étendre les requêtes. Les résultats montrent que lorsque les propositions de traduction ou d'extension sont supervisées par l'utilisateur, il améliore la recherche d'information par rapport à une traduction basée sur un dictionnaire initial.

Abstract

In recent years, with a rapid expansion of online information available on medical web sites in different languages, one of the issues that have to be addressed is that of the access and the processing of this online information. It generally assumes that large, multilingual lexical resources are available for each language pair. How to update these multilingual resources becomes an important clue, especially in a rapidly evolving domain such as medicine.

This thesis focuses on domain-specific bilingual lexicon extraction from online medical texts. Our goal is to develop a translation method for bilingual lexicon acquisition from comparable corpora and for query translation in cross-language information retrieval (CLIR). We present here a novel approach based on *words distribution symmetry*.

Traditional approaches to bilingual lexicon extraction from comparable corpora are based on the assumption that words that are translations of each other will have similar distributional profiles across languages. However, they proposed one direction extraction, only from the source to the target language. The basic intuition of the symmetrical distribution is that the reciprocal distribution similarity between two words of different languages is an effective criterion for identifying the translational affinity between words.

On the one hand, we evaluated our model for a French-English medical lexicon extraction. On the other hand, the extracted lexicon is used for query translation and expansion in CLIR. The results show that our approach exploring symmetrical distribution performs better than the traditional approach to bilingual lexicon extraction. For query translation and expansion tasks, our model improves the retrieval results only in a semi-supervised mode when compared with the dictionary-based method.

Table des matières

Chapitre 1

Introduction

1.1 Problématique générale

L'inflation documentaire et notamment textuelle est une des caractéristiques du Web depuis ses débuts. Aujourd'hui, le Web est la première source d'information du monde professionnel. De plus en plus de textes sont disponibles sous forme électronique et l'Internet est le réseau privilégié d'échange et de communication des communautés spécialisées : scientifiques, techniques, etc.

Parallèlement à cette inflation, des genres textuels nouveaux apparaissent (le mail, le forum, le chat...) et l'on voit émerger un multilinguisme dont l'ampleur est corrélée à 'l'internetisation' grandissante du monde. L'information textuelle électronique disponible est donc volumineuse, diversifiée et multilingue.

Le domaine médical n'échappe pas à ce phénomène. L'importance croissante du développement des réseaux internes dans les hôpitaux et les cliniques puis de l'Internet participent à l'émergence d'un enjeu important : l'accès et l'exploitation des informations médicales (Degoulet & Fieschi, 1991) de plus en plus nombreuses dans un contexte bilingue voire multilingue.

Cette augmentation tant en volume qu'en nature des informations a des conséquences sur les disciplines liées aux traitements des informations textuelles. De disciplines initialement cantonnées aux laboratoires de recherche universitaires ou à des applications très spécifiques (traduction automatique, indexation, etc.), on est passé à une véritable ingénierie des langues qui œuvre dans le monde de l'Internet et des entreprises. Cette évolution répond aux besoins de rechercher, classer, explorer, traduire l'information.

Or, l'accroissement de la production textuelle et l'évolution afférente des techniques de traitement de ces textes ont créé une véritable pénurie de ressources lexicales. Traiter automatiquement une information textuelle nécessite en effet des connaissances qui prennent la forme de lexiques, dictionnaires, ontologies... Malheureusement, le rythme de création de ces ressources est beaucoup plus faible que le rythme de création des néologismes lui-même corrélé au rythme effréné de production des textes. Ce

décalage entraîne une pénurie chronique en ressources lexicales. Cela s'explique par le fait que l'acquisition lexicale est encore aujourd'hui largement une opération manuelle donc longue et coûteuse. Ainsi de nombreux domaines ne disposent pas de thésauri spécialisés et le construction de ces derniers est confrontée aux difficultés encore mal résolues de l'acquisition automatique de connaissances (Grefenstette, 1994b).

Dans le domaine de la médecine, les ressources lexicales existent. Par exemple le metathésaurus UMLS[1] couvre plus de 800 000 concepts et plus de 2 000 000 de chaînes de caractères issus d'une centaine de terminologies biomédicales (*MeSH*, SNOMED, CIM, DSM...) (NLM, 2000)) dont beaucoup sont en langues autres que l'anglais. Cependant, le problème de l'actualisation des ressources lexicales médicales notamment multilingues reste entier. En particulier, les lexiques doivent rendre compte des néologismes en usage dans les textes et ceci en plusieurs langues pour alimenter les processus comme la traduction qu'elle soit automatique ou manuelle. Ceci est particulièrement vrai pour les domaines en constante évolution (médecine, informatique, intelligence artificielle, etc.) pour lesquels le rythme élevé d'apparition des néologismes nécessite une remise en question continuelle de leurs terminologies (Chiao & Sta, 2002).

C'est en partie en réponse à la pénurie de ressources lexicales bilingues et dans le but d'automatiser le plus possible l'actualisation des lexiques spécialisés dans un cadre multilingue (français-anglais) que ce travail a été effectué.

1.2 Plan général

Dans un premier temps, nous présentons au chapitre 1 les objectifs de ce travail en insistant sur ce qui l'a motivé et sur les hypothèses fondamentales qui le soutiennent. Nous y verrons en particulier l'importance de l'analyse distributionnelle, fondement de l'approche développée ici.

Le deuxième chapitre aborde l'acquisition de lexique en présentant le corpus comme élément essentiel pour y parvenir. Nous distinguons à cette occasion l'acquisition monolingue de l'acquisition multilingue.

Le chapitre 3 présente le modèle et la méthodologie développés dans ce travail. Il s'agit d'un modèle visant à extraire un lexique bilingue médical à partir de corpus comparables. Plus précisemment, ce modèle propose pour un mot donné en français, des candidats à sa traduction en anglais. La méthode avancée ici repose sur l'analyse distributionnelle et sur une de ses caractéristiques ignorée jusqu'ici, la symétrie distributionnelle entre les langues. Cette observation nous a amené à construire un modèle et à proposer une nouvelle mesure, la similarité croisée répondant à cette symétrie et rapprochant un mot et sa traduction.

Le chapitre 4 met en œuvre notre modèle à travers plusieurs expériences d'acquisition lexicale médicale bilingue (français - anglais). Ces expériences ont été construites

1. Unified Medical Language System (http://www.nlm.nih.gov/research/umls).

dans le but d'expérimenter plusieurs paramètres (taille de fenêtre de mots, taille du corpus...) et de valider notre modèle.

Le chapitre 5 aborde la recherche d'information translangue en insistant sur la reformulation de la requête.

Une expérience a été menée sur ce sujet à partir des données de la collection OHSUMED (décrite dans la section 6.3) proposée par TREC[2] (Text REtrieval Conference) et est relatée dans le chapitre 6. Elle met en évidence l'intérêt des ressources lexicales bilingues et de leur acquisition en corpus pour la recherche d'information.

Enfin, le dernier chapitre conclut ce travail en en présentant les perspectives.

1.3 Objectifs

L'objectif de ce travail n'est pas de réaliser un outil de traduction automatique mais bien de développer et d'évaluer un modèle d'aide à l'acquisition de lexique bilingue dans le domaine médical. Ce modèle vise à traduire en anglais un mot français du domaine médical dans le but d'alimenter un lexique bilingue. Il s'agit plus précisément d'un modèle qui propose un ensemble de candidats à la traduction pour un mot donné à partir de textes rapatriés du Web et plus particulièrement à partir de corpus comparables, notion que nous développons dans le chapitre 2. Précisons que ce modèle repose sur l'exploitation de ressources bilingues de différentes natures : publications en ligne, dictionnaires de langues spécialisées et générales, etc.

Cet ensemble de candidats a pour vocation première à être soumis au jugement d'un ou plusieurs experts pour en extraire la traduction correcte. L'objectif du modèle est donc de produire une liste suffisamment réduite de candidats à la traduction qui ait de fortes chances de contenir la traduction correcte.

Ce modèle est fondé sur l'approche distributionnelle et sur l'analyse statistique de corpus comparables et plus particulièrement sur la notion de cooccurrence. Avant de préciser dans ce chapitre le cadre théorique retenu, notons d'ores et déjà que l'avantage de l'approche distributionnelle et des techniques statistiques associées est multiple : indépendance relative vis-à-vis des langues mises en jeu, facilité d'implémentation et robustesse face aux données bruitées ou partielles. Cette dernière qualité est particulièrement intéressante pour exploiter les ressources textuelles du Web dont on ne peut pas toujours garantir la qualité.

Nous nous limitons à l'acquisition de traduction entre mots simples, c'est-à-dire entre mots constitués d'unités linguistiques simples. Comme on le verra plus avant dans ce chapitre, la prise en compte de mots complexes pose des problèmes mal résolus aujourd'hui et est moins centrale pour le domaine médical. D'autre part, la plupart des systèmes d'extraction de lexique bilingue à partir de corpus comparables travaillent à ce niveau.

2. http://trec.nist.gov

Ce travail est aussi un travail d'évaluation du modèle proposé. La construction ou l'actualisation de lexiques bilingues médicaux est la première grande application sur laquelle il est évalué (chapitre 4). Le processus d'acquisition peut être décomposé en trois phases consécutives : détection des mots dont on cherche la traduction, proposition de candidats à la traduction, validation manuelle de la bonne traduction. Nous ne traitons pas tout le processus mais uniquement la phase de proposition de candidats à la traduction. La phase de détection des mots à traduire relève du domaine de l'extraction de termes à partir de corpus mais est tout de même une piste intéressante pour aller au delà des mots simples.

La deuxième grande application sur laquelle est évalué ce modèle est la recherche d'information translangue (chapitre 6). Le principe consiste à partir d'une requête exprimée en langue source, à trouver les documents en langue cible, répondant à cette requête. Pour cela, une solution notamment mise en œuvre dans (Fluhr *et al.*, 1998) est la reformulation de la requête avec des ressources lexicales multilingues. Notre modèle a été appliqué dans ce cadre à la reformulation des mots constituant une requête en les traduisant. Le modèle propose par ailleurs plusieurs candidats à la traduction pour un mot constituant une requête. Or parmi ces candidats, il peut y avoir en plus de la bonne traduction, des mots sémantiquement proches du mot initial de la requête. Cette caractéristique a été exploitée pour étendre la requête afin de tenter d'améliorer les résultats de la recherche.

Notre intention est donc de construire et d'évaluer un modèle qui se propose d'exploiter les ressources textuelles provenant du Web dans le domaine médical en constituant des corpus spécialisés dans différentes langues (français et anglais), d'en extraire ou d'actualiser un lexique bilingue à l'aide de techniques à partir de corpus, et d'en faire usage pour traduire ou étendre les requêtes en recherche d'information translangue.

1.4 Le corpus comme source d'acquisition lexicale bilingue

Avant d'exposer dans la section suivante le cadre théorique dans lequel nous nous situons, cette section aborde certains choix effectués en les justifiant.

Le corpus de textes s'impose comme source d'acquisition lexicale car il est le lieu où les termes de spécialité sont en usage. Le déploiement de l'Internet, la disponibilité de grandes quantités de publications et documentations diverses sous format électronique font depuis quelque temps du corpus une ressource privilégiée et indispensable pour construire et valider les applications d'informatique documentaire.

Nous nous plaçons dans le cadre des langages de spécialité et de la linguistique de corpus (Habert *et al.*, 1997). Dans ce cadre, le corpus doit être aussi représentatif que possible du domaine pour assurer la portée des résultats de l'extraction. La couverture lexicale du corpus pour le domaine considéré est essentielle. L'idée est d'exploiter les

informations contextuelles disponibles dans un corpus, souvent utilisées pour l'alignement de termes dans un contexte monolingue (Deerwester *et al.*, 1990; Grefenstette, 1994b; Rajman *et al.*, 2000) et multilingue (Fung & Yee, 1998; Rapp, 1999; Déjean & Gaussier, 2002; Teubert, 2001).

On distingue généralement les corpus alignés ou parallèles des corpus comparables. Les premiers sont des traductions l'un de l'autre, les seconds traitent du même thème. La notion de comparabilité des corpus sera examinée ultérieurement dans la section 2.4.3. Retenons que les corpus alignés souffrent d'un manque de disponibilité et que les techniques d'extraction à partir de telles ressources sont aujourd'hui éprouvées mais de portée limitée à cause de ce déficit.

Les corpus comparables sont plus courants mais le niveau de performance des techniques d'extraction de lexique bilingue appliquées à ce type de corpus est encore insuffisant. La difficulté est en effet plus grande car l'espace de recherche de la bonne traduction d'un mot est beaucoup plus vaste que pour un corpus aligné où il est réduit à des segments alignés.

C'est donc dans la perspective de l'acquisition lexicale bilingue à partir de corpus comparables que nous nous situons. Il s'agit d'une problématique encore mal résolue aujourd'hui mais fructueuse à terme étant donné la disponibilité de telles ressources.

Termes simples et termes complexes

Dans le cadre des domaines de spécialité, on parle de mot de spécialité ou de terme. On distingue généralement les termes simples composés d'une seule unité lexicale ou mot plein, des termes complexes composés de plusieurs unités lexicales (Jacquemin, 1997b). Cette distinction n'est pas effectuée pour des raisons de différence de statut mais pour des raisons plus pragmatiques de difficulté à les identifier.

En effet, les termes simples peuvent être ambigus faute de contexte mais ont une structure syntaxique simple en raison de leur réduction à une unité syntaxique. Les termes complexes, au contraire, posent moins de problèmes de polysémie mais exigent une analyse syntaxique plus profonde et complexe pour les identifier. La difficulté provient de l'absence de critères purement linguistiques permettant de délimiter l'ensemble des termes complexes (Habert & Jacquemin, 1993). Ceci est particulièrement vrai pour les langues comme le français, le chinois, etc., pour lesquelles la variabilité des termes complexes dans les textes est plus importante.

Les outils développés pour identifier et extraire automatiquement les termes complexes d'un texte diffèrent en général au niveau de la méthode d'analyse du corpus employée. Les méthodes statistiques consistent à déceler les cooccurrences préférentielles dans le corpus (Church & Hanks, 1990; Dunning, 1993). L'approche linguistique (ou plus précisément l'analyse morphosyntaxique) s'appuie sur les caractéristiques linguistiques concernant la structure des termes complexes (David & Plante, 1990; Bourigault *et al.*, 1996). Les méthodes hybrides associent des modèles statistiques et structurels (Enguehard, 1994; Daille, 1996). Quelle que soit la stratégie adop-

tée, les résultats ne sont pas parfaits. Tous ces outils d'extraction de termes nécessitent un traitement ultérieur conséquent de validation manuelle des termes par des experts du domaine étudié.

C'est par manque d'outils robustes que nous nous limitons au cas du terme simple. Soulignons aussi que les termes simples, dont la fréquence dans un corpus est en général plus élevée que celle des termes complexes, sont ainsi plus appropriés aux calculs statistiques (Bourigault & Jacquemin, 1999). Une raison moins pragmatique nous a également amené à nous focaliser sur les mots simples. Le problème de l'actualisation des lexiques est lié aux néologismes. Il s'avère que dans le domaine médical, le mode de construction des néologismes produit essentiellement des dérivés morphologiques (Jacquemin & Zweigenbaum, 2000) : *diabète, diabétique* ; des acronymes : *DDT, HBV, ARN* ; ou des composés simples : *hépatosplénomégalie, hépato-porto-enterostomie*. Cette remarque semble aussi valable pour les autres langues que le français. Cette observation nous a amené à donner au terme simple un rôle central.

Enfin, le problème de la traduction de nombreux mots complexes peut se réduire à la traduction de leurs unités constituantes simples à certaines variations près : la variation de l'ordre des mots dans le terme (*acide folique*/*folic acid*) ou encore la variation de la catégorie grammaticale (*arrêt cardiaque*/*heart arrest*). Par exemple traduire *hémoglobinurie paroxystique* revient à traduire *hémoglobinurie* par *hemoglobinuria* puis *paroxystique* par *paroxysmal* ; ou encore *acide* et *folique* dans *acide folique* sont traduits par *acid* et *folic* respectivement pour obtenir *folic acid*.

Il faut néanmoins nuancer cette remarque en précisant que la traduction d'un mot simple dépend souvent du ou des mots associés dans le mot complexe : c'est notamment le cas pour les mots polysémiques. Ainsi les mots *banque* et *bloc* sont traduits respectivement par *library* et *rooms* dans les termes *banque de gènes* et *bloc opératoire*, alors qu'ils sont respectivement traduits par *banks* et *block* dans *banque de sang* et *bloc cardiaque*. Cette nuance ne remet néanmoins pas en cause l'argument développé pour autant que le modèle propose toutes les alternatives de traduction comme candidats. Le choix de la bonne traduction est alors le résultat d'une validation manuelle ou d'heuristiques en aval.

1.5 Hypothèses et méthodologie

1.5.1 Sémantique distributionnelle et acquisition lexicale

Nous nous plaçons dans le cadre des approches distributionnelles d'acquisition de connaissances sémantiques. La sémantique distributionnelle ((Rajman & Bonnet, 1992) et (Harris, 1988; Harris, 1991)) postule que le sens d'une unité lexicale peut être défini par ses contextes. Plus précisément, le sens d'un mot est décrit par sa distribution sur un ensemble de contextes.

La nature et la taille du segment contextuel utilisé jouent un rôle important dans

l'analyse distributionnelle (Fung & McKeown, 1997). La notion de contexte sera explicitée au chapitre 3 (sections 3.1 et 3.2). Retenons que le contexte peut être positionnel, syntaxique, phrastique ou documentaire selon ce qui définit les frontières et les contraintes appliquées sur les mots le constituant.

Définir le sens d'un mot par sa distribution contextuelle induit aussi l'idée que deux mots ayant la même distribution sont sémantiquement proches ou tout au moins sémantiquement liés. Ainsi, la sémantique distributionnelle propose de rapprocher ou d'opposer les unités lexicales sur la base de contextes partagés. Notons que pour deux mots, avoir les mêmes contextes ne signifie pas nécessairement apparaître conjointement dans les mêmes contextes mais plutôt avoir les mêmes mots comme voisins dans un corpus.

A partir de cette idée, certains travaux ont appliqué un point de vue statistique à l'hypothèse de base de la sémantique distributionnelle, avançant l'idée que plus souvent deux mots ont les mêmes contextes dans un corpus, plus la liaison sémantique entre ces deux mots est forte. Ces approches relèvent de la linguistique de corpus à base de statistiques, tendance dans laquelle nous nous situons. Il faut noter que rien ne révèle la nature de la liaison sémantique entre les mots ainsi dévoilée. Il peut s'agir de synonymie, d'antonymie, d'hyperonymie, d'hyponomie, de méronymie... Il semble que la nature des contextes ait une incidence sur la nature de la liaison sémantique sans que les choses soient bien établies.

Dans un cadre bilingue, l'hypothèse de base de la sémantique distributionnelle reste valide mais demande à être précisée. Elle se reformule de la façon suivante : deux mots de langues différentes ayant les mêmes distributions sont sémantiquement liés. Dans la pratique, afin de pouvoir comparer les distributions de contextes de deux mots de langues différentes, le passage d'une langue à l'autre est nécessaire et s'effectue par l'intermédiaire de ressources bilingues disponibles (lexiques, dictionnaires ou thésaurus). Il y a ici transfert des distributions vers un espace de contextes réduit [3] et commun aux deux langues. C'est sur la base de cet espace commun que sont comparées les distributions. La nécessité d'établir un pont entre les deux langues conduit à adopter plutôt l'hypothèse suivante : deux mots de langues différentes ayant des distributions proches sur un ensemble commun de contextes sont sémantiquement liés. Pratiquement, cet ensemble commun est une liste de mots qui sont obtenus à l'aide d'un lexique bilingue ou d'un thesaurus.

Dans le cadre de l'acquisition lexicale bilingue, la relation sémantique que l'on vise est la relation de traduction. On peut donc finalement formuler l'hypothèse retenue de la façon qui suit.

Hypothèse : un mot de la langue source et un mot de la langue cible ayant des distributions similaires sur un ensemble commun de contextes sont traductions l'un de l'autre.

3. Le fait que l'espace soit réduit est une conséquence de la limitation aux contextes communs.

D'un point de vue statistique, on retient l'idée que plus deux mots ont des distributions proches sur un espace commun de contextes, plus ils ont de chances d'être traductions l'un de l'autre.

1.5.2 Approche vectorielle

A partir des hypothèses de la sémantique distributionnelle, tenter de rapprocher deux mots de langues différentes revient à comparer deux distributions. Il convient donc tout d'abord :

- de disposer ou de construire les corpus permettant ;
- de construire les contextes ;
- de calculer les distributions des mots sur ces contextes et enfin ;
- de comparer ces distributions.

Une des façons privilégiées de représenter les distributions est le modèle vectoriel que nous adoptons ici. Il s'agit d'un modèle utilisé en recherche d'information (Salton *et al.*, 1974) : une requête et les documents recherchés sont décrits dans un espace vectoriel d'un grand nombre de dimensions. Une dimension correspond à un mot. Dans le cadre de l'acquisition de traduction à partir de corpus, ce sont les mots à traduire ainsi que les mots candidats à la traduction qui sont décrits dans l'espace vectoriel des contextes. Ici aussi une dimension correspond à un mot. La distribution contextuelle d'un mot est donc représentée par un vecteur de contexte.

Une fois décrits les mots à traduire et les mots candidats dans l'espace vectoriel, pour comparer les vecteurs de contexte, plusieurs mesures de similarité notamment utilisées en recherche d'information, sont disponibles (Jaccard, cosinus, Dice...). Une telle mesure permet pour un mot donné d'ordonner les candidats à la traduction et ainsi de proposer une liste réduite à la validation manuelle en coupant cette liste à un seuil généralement choisi empiriquement.

Cette représentation vectorielle d'une distribution présente des avantages. En particulier, tout l'arsenal du calcul vectoriel est disponible pour opérer sur les distributions. Néanmoins, il s'agit bien d'une approximation qui a des limites. Le modèle vectoriel fait notamment l'hypothèse d'orthogonalité des dimensions entre elles donc d'indépendance des mots entre eux. Or, cette hypothèse est invalidée dans la pratique car bon nombre de mots entretiennent des relations de nature diverse. Cette hypothèse sera toutefois retenue ici comme par ailleurs dans la plupart des autres travaux. Notons toutefois plusieurs tentatives pour contrer ce problème du modèle vectoriel classique. La méthode LSI (Latent Semantic Indexing) (Littman *et al.*, 1998; Oard & Dorr, 1996; Brown, 1998) que nous détaillons dans la section 5.4.2 permet de réduire la taille de l'espace vectoriel initial (souvent de l'ordre de plusieurs milliers) en construisant et

ne retenant que quelques centaines de dimensions, combinaisons linéaires des dimensions initiales. Par ailleurs, le modèle vectoriel généralisé (Wong *et al.*, 1985) quant à lui propose de représenter chaque mot dans un espace réduit de mots initiaux.

1.5.3 Symétrie distributionnelle : une nouvelle notion

Nous utilisons ici une nouvelle notion que nous nommons 'symétrie distributionnelle'. Cette notion a été introduite dans (Chiao & Zweigenbaum, 2002b; Chiao *et al.*, 2004)[4].

La sémantique distributionnelle appliquée dans un cadre bilingue repose sur l'idée d'identité des distributions et dans un but opératoire, de proximité des distributions. Deux mots ayant des distributions proches ont toutes les chances d'être traductions l'un de l'autre. Or, lorsqu'on parle ici de proximité des distributions on parle implicitement de proximité par rapport aux autres mots candidats.

Pour un mot donné, il s'agit bien de comparer sa distribution avec celle de tous les mots candidats puis de ne retenir que le ou les plus proches. On dira donc plus précisément que la distribution d'un mot source est proche de celle d'un mot cible au sens d'une mesure de sa proximité relative à l'ensemble des mots cibles. Pour un couple de langues (A et B), il existe donc deux mesures de la proximité entre mots : la première dans une direction de traduction (langue $A \rightarrow$ langue B) et la seconde dans l'autre direction de traduction (langue $B \rightarrow$ langue A).

Ces deux proximités sont-elles équivalentes ? En d'autres termes, si deux mots sont proches au sens de la proximité dans une direction de traduction, le sont-ils aussi pour l'autre direction de traduction ? La réponse est négative dans un certain nombre de cas. Notamment, pour deux mots traductions l'un de l'autre dont l'un est polysémique (*temps*) et l'autre ne l'est pas (*weather*), les deux proximités ne seront pas équivalentes car le mot polysémique sera proche de mots associés à ses multiples significations alors que le mot non polysémique ne sera proche que des mots associés à son unique signification.

Nous faisons néanmoins l'hypothèse que pour certains mots la réponse est positive. Nous avançons l'hypothèse ici que si deux mots sont proches dans une direction de traduction ainsi que dans l'autre (langue $A \leftrightarrow$ langue B) alors ils ont de plus fortes chances d'être traductions l'un de l'autre que s'ils ne sont proches que pour une seule direction de traduction. C'est ce que nous appelons la symétrie distributionnelle qui est plus exactement une symétrie des mesures de proximité des distributions.

C'est sur cette hypothèse que nous avons construit notre modèle d'acquisition lexicale bilingue développé au chapitre 3 et validé expérimentalement au chapitre 4. Cette hypothèse peut être traduite intuitivement par cette image. Dans un dictionnaire bilingue, lorsqu'on cherche la traduction d'un mot a en langue A et que l'on trouve

4. Dans ces articles, la notion de symétrie distributionnelle a été introduite sous le nom de 'reverse'. Dans le cadre des corpus parallèles étiquetés, une notion proche a été mise en œuvre récemment dans (Ozdowska, 2004).

le mot b en langue B, on s'attend inversement en cherchant la traduction du mot b à trouver le mot a.

Il faut noter que cette hypothèse faite ici dans un cadre bilingue peut être généralisée avec la notion de corpus. Elle peut être formulée ainsi : si deux mots issus de deux corpus ont des distributions proches au sens de chacun de ces deux corpus, alors ils ont plus de chances d'être proches sémantiquement que s'ils ont des distributions proches au sens d'un seul corpus.

Cette généralisation s'applique à des relations sémantiques symétriques comme la synonymie, la collocation, la traduction mais a priori pas aux relations asymétriques comme l'hyperonymie ou la méronymie. Elle peut donner des moyens (qui restent à construire) pour établir de telles relations dans un cadre monolingue. La symétrie distributionnelle est en effet d'emblée assurée dans le cadre monolingue lorsqu'un seul corpus est utilisé et que la mesure de ressemblance entre les mots est symétrique. Dans ce cas, la mesure de ressemblance et la mesure inverse sont relatives au même corpus et donc sont symétriques. Cette hypothèse n'offre donc pas d'intérêt lorsque sont mis en œuvre un seul corpus et une mesure de ressemblance symétrique. Il est néanmoins envisageable de construire artificiellement des corpus distincts pour exploiter cette symétrie. Une autre possibilité consiste à n'utiliser qu'un seul corpus mais à utiliser des mesures de ressemblances asymétriques d'une langue vers l'autre.

1.6 Conclusion

Rappelons l'objectif, le périmètre et les choix théoriques et méthodologiques de ce travail. Il s'agit de construire un modèle d'acquisition lexicale bilingue dans le domaine médical pour répondre au déficit en ressources lexicales bilingues.

Applications et expérimentations

Le champ d'application de notre modèle est multiple. Deux grands types d'application sont examinés ici et servent de base pour expérimenter notre modèle. Le premier est l'acquisition lexicale bilingue (chapitre 4) et le second la recherche d'information translangue (chapitre 6).

Mots simples

Etant donné d'une part le rôle plus central des termes simples par rapport aux termes complexes pour le domaine médical et d'autre part la difficulté à identifier les termes complexes, nous avons limité la portée de notre modèle aux mots simples[5].

5. Il s'agit plus exactement des mots simples et des mots composés de plusieurs mots simples liés par le trait d'union (Catach, 1963).

Corpus comparables

Ce modèle propose pour un mot donné, des candidats à la traduction à partir de corpus et de ressources lexicales bilingues existantes (section 2.4.3 et chapitre 3). Le manque de corpus alignés et la meilleure disponibilité de corpus comparables font de ces derniers la ressource privilégiée pour l'acquisition lexicale bilingue.

Sémantique distributionnelle et modèle vectoriel

Le fondement théorique de cette approche est la sémantique distributionnelle et le modèle vectoriel associé. Nous nous inscrivons ainsi dans la lignée des travaux de la linguistique de corpus. L'avantage principal de cette approche est sa robustesse. En particulier, son application à d'autres domaines et à de nouvelles langues est envisageable sans le remettre en cause fondamentalement. Il s'applique directement à des corpus de textes bruts, sans que ceux-ci soient préalablement étiquetés (chapitre 4). L'approche vectorielle permet en outre d'aligner les mots à partir de textes bruts sans recours à l'analyse syntaxique des phrases (section 3.2.3). Sans nécessiter de gros dictionnaires de langue générale ou spécialisée, elle requiert néanmoins des ressources lexicales bilingues partielles préexistantes pour établir le passage entre les langues.

Symétrie distributionnelle

L'originalité de notre approche est la prise en compte d'une propriété des distributions de contextes de certains mots, propriété que nous nommons la 'symétrie distributionnelle' (section 3.3). Nous avançons ici l'idée que la symétrie distributionnelle est une caractéristique dont la prise en compte améliore l'extraction de la relation de traduction entre mots et expérimentons cette hypothèse (chapitre 4).

Chapitre 2

Acquisition lexicale à partir de corpus

2.1 Introduction

Suite à l'informatisation des hôpitaux, au développement d'Internet et à l'internationalisation des échanges, le nombre d'informations médicales sous forme électronique augmente sans cesse. Afin de pouvoir créer, rechercher, exploiter, et traduire ces informations, les outils de gestion de l'information ont besoin de données lexicales qui facilitent l'accès aux bases de connaissances (Zweigenbaum *et al.*, 2003). Or la constitution de ressources lexicales à grande échelle est toujours un objectif et un problème majeur pour la plupart des groupes de recherche, des industriels, développant des systèmes de traitement de l'information multilingue (Boitet, 2001; Arnold *et al.*, 1994).

En effet, la pénurie de telles ressources est directement liée à la façon manuelle de les constituer. La méthodologie classique d'acquisition lexicale consiste à d'abord définir la structure des dictionnaires pour chaque langue, de façon à établir un langage symbolique ('métalangage') ; ensuite les lexicographes créent le dictionnaire entrée par entrée en attribuant à chacune les informations correspondantes à l'aide de corpus ; puis enfin ils procèdent à la révision finale des entrées du dictionnaire.

L'acquisition manuelle de ressources lexicales s'avère donc une tâche laborieuse et coûteuse à cause de la masse et de la variété des informations à construire. Ces ressources sont souvent incomplètes ou obsolètes notamment lorsqu'il s'agit de domaines de spécialité vastes et en évolution comme la médecine, qui sont composés de sous domaines, et de micro-domaines (section 2.2.3) dans lesquels les concepts émergent avant d'avoir pu être institutionnalisés sous une forme linguistique privilégiée par une communauté.

L'acquisition manuelle est beaucoup trop longue pour s'appliquer à un nouveau domaine ou pour répondre à la nécessité pour les lexiques établis de coller au discours. Ce phénomène est exacerbé dans les domaines médicaux en constante évolution (*e.g.*, génétique moléculaire, épidémiologie, etc.) pour lesquels l'apparition fréquente de néologismes nécessite une actualisation continuelle de leurs lexiques. Par actuali-

sation d'un lexique multilingue, nous entendons l'enrichissement du vocabulaire du lexique en ajoutant de nouvelles paires de mots qui sont des traductions l'un de l'autre.

Plusieurs travaux de recherche s'intéressent à l'acquisition terminologique bilingue ou multilingue en corpus Web (Nie *et al.*, 2000; Fung & Yee, 1998; Déjean & Gaussier, 2002; Cancedda *et al.*, 2003), visant deux types d'application : la recherche d'information translangue et la construction de lexique spécialisé.

Dans la section suivante (2.2), nous montrons l'intérêt et le rôle central que prend le corpus comme source d'acquisition lexicale. Nous abordons ensuite l'acquisition lexicale monolingue d'une part et l'acquisition multilingue d'autre part. Dans le cadre monolingue (section 2.3), la problématique est liée à celles de l'extraction et de la structuration de termes, tandis que l'acquisition lexicale multilingue (section 2.4) s'apparente à l'alignement des termes, c'est-à-dire à l'extraction de traductions à partir de corpus bilingues.

2.2 Corpus comme ressource lexicale

2.2.1 Pourquoi le corpus ?

Depuis des décennies, de nombreux travaux s'intéressent à l'acquisition automatique des informations lexicales. Deux sources peuvent être exploitées pour les identifier : les dictionnaires existants et les corpus (ensemble de textes choisis pour illustrer l'usage). L'utilisation des dictionnaires pour l'acquisition lexicale est limitée par deux phénomènes : d'une part l'incompatibilité des relations lexicales entre les entrées d'un dictionnaire à l'autre et d'autre part l'incomplétude des informations lexicales dans les domaines de spécialité.

La réutilisation des dictionnaires pose des problèmes complexes et n'est pas qu'une affaire de compatibilité de formats de stockage ou de présentation des informations (Véronis & Ide, 1991). Le problème majeur réside dans une incompatibilité plus profonde des dictionnaires : ils diffèrent souvent par leurs niveaux de coordination et de couverture (critères de sélection des sens, critères de décomposition en entrées et sous-entrées dans le cas d'homographes, liens syntaxiques ou sémantiques entre les entrées...). Le problème est encore exacerbé lorsque plusieurs langues sont en jeu.

Le second problème est également lié à la nature du dictionnaire. Une telle base lexicale est toujours obsolète. Les informations présentes ne sont pas toujours suffisantes pour obtenir le sens précis des termes dans le cas d'une application spécifique à un domaine particulier. Dans les domaines de spécialité, l'usage d'une base lexicale générale préexistante comme *WordNet* (Miller *et al.*, 1990) n'est plus valable hors contexte. En effet, l'utilisation systématique de ce genre de données par une application de traitement automatique des langues est sujette à caution : dans quelle mesure un modèle sémantique conçu *a priori* s'avère-t-il adéquat pour représenter le fonctionnement de domaines particuliers (Claveau *et al.*, 2001) ?

C'est pourquoi de plus en plus de travaux en acquisition lexicale utilisent les textes dans lesquels les termes sont en usage, en l'occurrence le discours écrit. L'extraction de ressources lexicales à partir de corpus est devenue un thème majeur dans les domaines comme la lexicographie, la terminologie, la traduction automatique, etc. Par exemple, le projet *COBUILD* (Sinclair, 1991) a constitué des textes tirés de sources et de genres très divers, écrits et oraux afin d'obtenir un corpus (*Bank of English*) destiné à l'usage pédagogique et lexicographique de l'anglais général.

2.2.2 Corpus spécialisés et notion de sous-langage

Aujourd'hui, le déploiement du Web et la disponibilité croissante de publications et documentations diverses sous format électronique donnent au corpus un rôle central dans le domaine du traitement automatique des langues. Le corpus présente en effet l'avantage d'une plus grande souplesse : la profusion des textes spécialisés numérisés permet d'en extraire les termes et de suivre l'évolution du langage utilisé dans le domaine étudié en rendant compte des néologismes.

Néanmoins, le corpus pose un problème de représentativité. En effet, un corpus peut contenir des textes de natures très diverses et n'aborde pas la totalité des notions d'un domaine. Il en aborde à la fois moins et plus, mélangeant concepts de disciplines transversales et concepts de la langue générale. La représentativité des corpus est d'autant plus cruciale pour les disciplines transversales qui nécessitent des sources textuelles provenant de multiples domaines. L'acquisition lexicale à partir de corpus reste donc dépendante du corpus utilisé. La portée d'un terme devra donc être étendue avec prudence d'un corpus vers un domaine.

Marcus (Marcus *et al.*, 1994) suggère que le terme *corpus* soit réservé à l'ensemble de textes choisis de façon très précise pour répondre à des besoins et critères particuliers. D'après la définition de Sinclair (Sinclair, 1996), un corpus est une collection de données qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon de la langue. En résumé, un corpus est un produit artificiel qui contient des textes sélectionnés en fonction de critères précis pour être utilisés comme un échantillon de la langue.

En général, les corpus spécialisés se caractérisent par les objectifs du travail à accomplir. Pour des chercheurs qui s'intéressent aux aspects lexicaux, comme les terminologues, les lexicographes, les textes spécialisés sont marqués par leur vocabulaire spécifique, c'est-à-dire par l'usage fréquent des termes techniques dans les textes. Ceux qui s'intéressent aux aspects grammaticaux font souvent appel à la notion de sous-langage (Kittredge & Lehrberger, 1982; Harris, 1988; Harris, 1991). Dans ce cas, un texte spécialisé est caractérisé non seulement par son vocabulaire mais aussi éventuellement par sa structure grammaticale différente de celle de la langue générale.

La grammaire des textes spécialisés est limitée en structures utilisées et celles-ci n'obéissent pas aux règles de la grammaire générale : utilisation d'ellipses, constitution de termes composés longs et compliqués, etc. (Pearson, 1999). De plus, ceux qui

étudient la typologie des textes distinguent les textes en fonction de leur fonction et de leur organisation. Par exemple parmi les différents types de documents médicaux, on distingue des articles de recherche, des dossiers patient, des guides de bonnes pratiques, des rapports de laboratoires, etc. (Habert *et al.*, 2001).

Selon Harris (Harris, 1988; Harris, 1991), les sous-langages se caractérisent par un lexique spécifique et une structure de phrase particulière. Les mots dans un sous-langage ont une signification plus restreinte que les mots de la langue courante et leur utilisation l'est aussi. Un sous-langage de domaine se distingue donc du langage naturel par ses restrictions lexicales et syntaxiques. L'expérience menée par Harris et son équipe sur le discours du domaine de l'immunologie montre que l'analyse syntaxique des phrases conduit à la création de formules identifiant des types de phrases, les classes de mots constituant les éléments primitifs.

Habert (Habert, 1998) démontre en citant Sager que « Si l'on applique à un corpus de textes d'un secteur scientifique des méthodes de linguistique descriptive similaires à celles utilisées pour le développement d'une grammaire d'une langue dans son ensemble, on obtient des motifs précis de cooccurrences de mots à partir desquels on peut définir des sous-classes de mots et des séquences de ces sous-classes qui sont caractéristiques (c'est à dire une grammaire)... . La grammaire d'un sous-langage doit attraper les restrictions d'occurrences qui distinguent un champ de discours scientifique d'un autre. ». La distribution d'une unité linguistique est en effet l'ensemble des contextes de cette unité dans le corpus. Dans un corpus du domaine, les mots appartenant au lexique du sous-langage forment la majorité des mots utilisés. L'étude des mots sémantiquement pleins les plus fréquents dans un corpus permet donc de former un lexique 'spécialisé' du domaine.

L'analyse distributionnelle de contextes syntaxiques est la méthodologie principale pour l'étude des sous-langage (Sager, 1987; Harris *et al.*, 1989; Daladier, 1990). Son application automatique reste toutefois difficile à grande échelle ; il s'agit souvent d'une analyse partielle qui peut correspondre au manque des outils d'analyse robustes ou au choix délibéré de ne s'intéresser qu'aux composants d'une certaine nature syntaxique, *i.e.*, l'extraction de termes qui se concentrent sur les syntagmes nominaux dans lesquels figurent les unités polylexicales du domaine.

2.2.3 Choix du corpus de domaine pour l'acquisition lexicale

Afin de pouvoir construire un corpus représentatif d'un domaine sans courir le risque d'être trop restreint aux disciplines aussi bien qu'aux genres, il faut d'abord analyser ses propres besoins et définir les critères de choix des textes. Des domaines différents emploient des approches linguistiques différentes, et à l'intérieur des domaines vastes comme la médecine se trouvent des sous-domaines voire des micro-domaines spécialisés. Par exemple :

Domaine : médecine

Sous-domaines : anatomie, maladie, etc.

Micro-domaines : parasitologie, épidémiologie, acupuncture, etc.

Selon Pearson (Pearson, 1999), le rapport entre l'auteur et le lecteur est un aspect primordial dans le choix des textes qui constituent le corpus à des fins terminologiques (extraction de termes, de définitions et de relations grammaticales ou sémantiques, etc.). Elle distingue trois types de rapports auteur-lecteur qui ont un impact sur la constitution de corpus :

Communication entre experts : riche en termes techniques, vocabulaire de domaine, mais probablement pauvre en information explicite comme les définitions.

Communication entre experts et initiés : beaucoup de termes techniques et plus de connaissances explicites lorsque les experts estiment que le terme utilisé n'est pas connu par le lecteur.

Communication entre experts et non-initiés : moins dense en termes techniques de domaine et plus riche en éléments explicites.

Plusieurs critères peuvent entrer en jeu lors de la sélection de textes. A part l'équilibre de domaine que nous avons mentionné dans la section précédente (corpus spécialisé et la notion de sous-langage), le choix des textes doit aussi respecter un équilibre pragmatique qui inclut des textes appartenant aux différents cadres de communication comme par exemple le cadre professionnel (comptes rendus de laboratoire) ou éducatif (publications pédagogiques, vulgarisation). La date de publication est très importante lorsqu'il s'agit de domaines en constante évolution (informatique, génétique moléculaire, pharmacologie, etc.).

Disposer de l'intégralité du texte est important afin d'avoir une distribution des termes régulière pour l'étude des sous-langages. Des corpus de résumés de la base de données bibliographiques MEDLINE (section 6.3) par exemple sont beaucoup utilisés pour les applications du traitement automatique des langues (création et évaluation des outils de TAL). L'avantage est qu'ils sont plus faciles à récupérer et libres d'accès. Cependant ils ne sont pas tout à fait représentatifs du point de vue distributionnel. Le contexte des termes dans un résumé n'est pas nécessairement identique à celui d'un texte complet.

Dans le cadre de notre travail, nous nous intéressons aux aspects lexicaux du corpus et nous avons construit un corpus spécialisé pour l'acquisition automatique de lexique bilingue du domaine médical (section 4.2.1), conçu uniquement à partir de pages Web indexées par des portails de santé : CISMeF et CliniWeb (la plupart des textes sont entiers). Un autre corpus provenant de la collection fermée *OHSUMED* (Hersh *et al.,* 1994) est destiné à l'expérience de recherche d'information translangue (section 6.3).

2.3 Acquisition terminologique monolingue

En principe, le processus d'acquisition terminologique automatique comporte deux étapes : l'extraction et la structuration des termes. L'extraction terminologique permet de dresser une nomenclature des termes, appelés aussi candidats termes (souvent des termes complexes) du domaine. La structuration terminologique permet de structurer et d'organiser les listes de candidats termes. Cette étape est réalisée dans l'optique de regrouper les termes en classes en fonction des différents types de relations liant les termes entre eux. Ces deux étapes ne sont pas toujours successives mais peuvent être imbriquées. En effet, la structuration d'autant plus nécessaire que difficile donne des points d'entrée et permet de centrer le travail de validation sur des zones terminologiques particulières (Nazarenko & Hamon, 2002).

Dans cette section, nous présentons un ensemble de travaux d'acquisition terminologique en corpus. Nous débutons par des travaux d'extraction de terminologie (section 2.3.2). Nous abordons ensuite des travaux sur la structuration de terminologie (section 2.3.3).

La plupart des travaux présentés dans cette section ne se limitent en fait pas à un des aspects de la terminologie. En effet, la richesse des informations provenant des textes et la diversité des relations que les termes entretiennent entre eux en corpus peuvent servir à la création de différentes ressources terminologiques. Nous allons tout d'abord étudier (section 2.3.1) ces différents types de relations sur lesquelles reposent les travaux que nous présentons plus loin.

2.3.1 Relations entre les termes

Selon Grefenstette (Grefenstette, 1994a), trois niveaux de relations entre les termes peuvent être identifiés dans un corpus.

Les relations du premier ordre ('first-order affinity')

Elles désignent les associations typiques des mots, qui sont différentes de celles qui seraient observées si la distribution des mots était aléatoire dans un corpus. Ces relations sont appelées 'collocation', pour signifier qu'elles sont associées à une répétition de cooccurrences (Benson, 1989; Halliday & Hasan, 1976). McKeown et Radev (McKeown & Radev, 2000) indiquent que le fait que les collocations sont plus fréquentes dans les corpus de domaines spécialisés constitue un indice souvent exploité pour l'extraction automatique de collocations. En fonction des types de contextes étudiés, deux catégories de collocations sont observées dans les études statistiques : grammaticale et sémantique.

Des collocations du type grammatical, *e.g.*, *save from* jouent un rôle important en acquisition lexicale pour la construction de dictionnaires collocatifs (Sinclair *et al.*, 1987), et celles du type sémantique , *e.g.*, *signes oculaires*, sont plus exploitables pour

l'accès au contenu des documents. L'une des techniques les plus utilisées pour identifier les collocations dans un corpus est la mesure de l'information mutuelle (Church & Hanks, 1990) que nous détaillons au chapitre 3.

Il s'agit de comparer la probabilité de cooccurrence de deux mots i et j dans une fenêtre de n mots par rapport à leur probabilité d'occurrence indépendante. La valeur de n permet de jouer sur le type de relation contextuelle étudié : les petites valeurs de n ont tendance à favoriser les relations syntagmatiques (verbe+préposition, nom+préposition, nom+adjectif, etc.) tandis que les grandes valeurs de n ont tendance à repérer les relations thématiques (doctor-nurse, doctor-dentist, etc.).

Les relations du deuxième ordre ('second-order affinity')

Ce sont des relations qui relient les termes qui partagent les mêmes relations du premier ordre dans un contexte similaire, par exemple les noms qui apparaissent en argument des mêmes verbes. L'hypothèse qui sous-tend l'analyse distributionnelle (Harris *et al.*, 1989) est que des mots ayant une proximité sémantique tendent à remplir des fonctions syntaxiques semblables en termes de prédicats ou d'arguments et donc, à avoir des distributions semblables. On peut s'en servir ensuite pour regrouper les mots en classes de mots.

Les mots ayant des contextes grammaticaux similaires sont regroupés en classes ayant une forte proximité sémantique, à l'aide d'une analyse de contextes sur la base de répétition de cooccurrences dans des contextes syntaxiques. L'étiqueteur grammatical ou l'analyseur syntaxique sont dans ce cas associés à l'analyse statistique afin d'avoir accès à un niveau supérieur d'information qu'une simple cooccurrence et d'identifier des relations syntaxiques à partir des différentes relations ou étiquettes grammaticales. Ainsi, il est possible d'exploiter les contextes syntaxiques de mots (sujet, verbe, objet, etc.) pour évaluer la proximité décrivant les relations sémantiques ou conceptuelles qui relient les mots entre eux.

Cette proximité des mots produit outre des synonymes, des antonymes et des mots complémentaires ayant des liens d'hyperonymie et de méronymie. Par contre, il est difficile de rendre compte de la polysémie des mots avec de telles relations hétérogènes. En effet, dans la quasi-totalité des approches, tous les mots ne peuvent appartenir qu'à une seule des classes calculées. La diversité des classes sémantiques issues de regroupement par contextes syntaxiques similaires est soulignée par (Church & Hanks, 1990).

Les relations du troisième ordre ('third-order affinity')

Ce sont des relations désambiguïsantes qui associent un terme polysémique aux différents termes qu'il recouvre. En partant des classes de termes issues du regroupement en relations du deuxième ordre, on effectue une analyse plus fine sur leurs contextes à l'aide d'une taxinomie, par exemple afin de dégager les différentes classes

sémantiques correspondantes. On passe ainsi des contextes de mots à des classes de contextes et on remplace les mots des contextes par leur classes sémantiques. De cette façon, il est possible d'avoir plusieurs classes sémantiques associées à un mot, *mémoire* par exemple, dont l'une désigne la faculté cognitive et l'autre désigne le travail personnel ou le disque dur en l'informatique.

Ces relations servent aussi à la constitution de réseaux de liens conceptuels entre les termes du domaine en affinant des synonymes, en séparant des antonymes, et en dégageant des hyperonymes, des hyponymes, etc.

Ces trois niveaux de représentation des relations entre les termes donnent une référence pour catégoriser, en fonction des types de relations mises en évidence, les différents travaux d'acquisition automatique de terminologie. Les relations du premier ordre concernent les associations au sein de structures syntaxiques telles que les termes complexes. Elles peuvent aussi présenter les relations d'association thématique si le contexte étudié est plus large. Elles permettent d'identifier et d'extraire des candidats termes qui servent à dresser une nomenclature des termes d'un domaine spécialisé.

Les relations du second et du troisième ordres, mettant en évidence les liens sémantiques ou conceptuel entre les termes, servent à structurer les termes collectés et à les regrouper en classes.

2.3.2 Extraction de termes

Les différents travaux présentés dans cette section utilisent deux approches : l'analyse structurelle (linguistique) et l'analyse statistique. La première repose sur les connaissances linguistiques, *e.g.*, les structures morphologiques ou syntaxiques des termes. La deuxième étudie les contextes d'emploi et les distributions des termes en corpus. Certains travaux mélangeant des techniques de natures différentes constituent des modèles dits 'hybrides'.

La première étape dans l'acquisition terminologique à partir de corpus est l'extraction de candidats termes, c'est à dire de mots ou groupes de mots susceptibles d'être retenus comme termes par un expert.

Dans les textes scientifiques et techniques, les termes et leurs variantes ont une expression linguistique privilégiée sous forme de syntagmes nominaux restreints dont la structure est apparentée à celle des noms composés. Il est donc possible de les détecter et de les extraire en exploitant cette forme syntaxique privilégiée du terme, c'est-à-dire le groupe nominal, et en prenant en compte sa variabilité en corpus.

Les différentes variations des termes complexes en corpus relèvent d'un phénomène reconnu et quantifié. 25% des occurrences des termes en français (15% en anglais) sont ainsi des variantes du type morphosyntaxique (Jacquemin, 1997b). Ces taux varient selon les textes, le domaine évoqué par les textes et les types de variantes cherchées.

Analyses structurelles

L'extraction par sélection de patrons syntaxiques part du constat que la structure contrainte des termes et de ses variantes en corpus est celle du terme composé[1].

En anglais, le terme se présente essentiellement sous la forme *modifieur + tête*, la *tête* étant principalement un substantif et le *modifieur* un substantif ou un adjectif. Cette forme peut ensuite s'enchaîner à l'aide d'une préposition ou non en constituant un terme à structure de nom composé.

En français, les structures des termes sont plus proches de la syntaxe du syntagme nominal qu'en anglais puisqu'il s'agit soit de syntagmes nominaux, soit de synapsies[2] et de syntagmes nominaux sans déterminant. Ce phénomène favorise l'existence de variantes et augmente alors le nombre de patrons syntaxiques à prendre en compte. L'application d'un patron catégoriel sur un segment de texte nécessite que celui-ci soit étiqueté et que les ambiguïtés concernant la catégorie grammaticale soient levées. Le recours à des analyses mophosyntaxiques et syntaxiques rend cette méthode très dépendante de la langue car elle nécessite des connaissances linguistiques approfondies de la langue, donc des ressources comme les règles de grammaire, dictionnaires, etc.

On peut citer l'outil *Termino* de David et Plante (David & Plante, 1990) qui a pour but l'identification des structures synapsiques en s'appuyant sur une grammaire des syntagmes nominaux. Il repose sur une analyse complète de la phrase en ses constituants afin de n'en dégager que les synapsies. Néanmoins, décrire les termes par leur structure interne seule est insuffisant.

L'outil *Lexter* de Bourigault (Bourigault *et al.*, 1996) propose une technique d'analyse syntaxique locale pour le repérage de syntagmes nominaux candidats à être des termes du domaine que représente le corpus. L'approche repose sur des marques de frontières externes de syntagmes telles que les verbes, les pronoms, les conjonctions et les séquences préposition+déterminant qui permettent d'isoler les entités complexes de type nominal, qui sont ensuite décomposées en couple (*tête, expansion*). Le résultat ainsi obtenu conduit à former un réseau de termes où chaque terme est relié à sa tête, son expansion, ainsi qu'aux termes dont il est la tête ou l'expansion. En revanche, les termes extraits présentent un bruit important dû à la complexité des phrases qui font l'objet de nombreuses combinaisons syntaxiques. De plus, les relations exprimées dans le réseau sont de type syntaxique et ne suffisent pas à définir des associations conceptuelles. Pour pallier à cela, une analyse des fréquences de termes qui consiste à observer dans le corpus les paires candidates (tête, expansion) les plus fréquentes permet de filtrer les structures obtenues par l'analyse de surface.

Faster[3] de C. Jacquemin est un analyseur syntaxique dédié à la reconnaissance en corpus de termes appartenant à une liste contrôlée fournie au système (Jacquemin,

1. Les termes qui comportent au moins deux mots pleins, c'est à dire des mots dont la catégorie est nom, adjectif ou verbe (Jacquemin, 1997b).

2. Il s'agit des syntagmes ayant une structure syntaxique complexe Nom Préposition Nom (Émile Benveniste, 1966).

3. Pour Filtrage et Acquisition Syntaxique de TeRmes.

1997b). Les travaux réalisés pour le développement du système ont pour objectif de créer un outil d'indexation automatique contrôlée et d'enrichir une terminologie existante. Pour l'auteur, les termes n'ayant pas toujours en corpus la même forme linguistique, le principal enjeu est de pouvoir identifier leurs variantes qui permettent ensuite de repérer les occurrences d'un terme en discours même s'il a fait objet d'une élision, d'une coordination, etc. Il définit ainsi un ensemble de métarègles pour générer les diverses variations potentielles des termes en contexte. Par exemple (extrait de (Bourigault & Jacquemin, 2000)) :

Variantes syntaxiques : "mesure de volume et de flux" est une variante de "mesure de flux".

Variantes morphosyntaxiques : "flux de sève mesurés" est une variante verbale de "mesure de flux".

Variantes sémantico-syntaxiques : "évaluation du flux" est une variante sémantico-syntaxique de "mesure de flux" qui fait appel à la proximité sémantique de "mesure" et "évaluation".

Analyses statistiques

Les méthodes statistiques sont souvent utilisées pour l'exploitation de corpus massifs (étiquetés ou non) qui constituent de réelles bases de connaissances linguistiques. Elles se distinguent des approches structurelles par le fait qu'elles requièrent moins de traitements linguistiques complexes[4].

Les expériences menées par Choueka (Choueka *et al.*, 1983; Choueka, 1988) pour l'acquisition de collocations sont parmi les premières à l'époque qui travaillent sur un corpus de taille importante (11 million de mots extraits des articles du journal *New York Times*). Selon Choueka, une collocation est une séquence de deux ou plusieurs mots non interrompus qui apparaissent plus d'une fois au sein d'un corpus. Etant donné que les techniques utilisées sont fondées sur la fréquence des cooccurrences dans le corpus sans aucune analyse linguistique, la pertinence des résultats dépend alors de la taille du corpus utilisé.

Les travaux de Lebart et Salem (Lebart & Salem, 1988; Lebart & Salem, 1994), comme ceux de Choueka, consistent à étudier les *segments répétés* dans un corpus dans le but d'en extraire un sous-ensemble jugé particulièrement pertinent. L'objectif est de représenter un texte, qui est habituellement vu comme un enchaînement de formes simples, comme un ensemble de formes simples et de segment répétés. Un segment répété désigne une séquence de deux ou plusieurs formes graphiques consécutives, non séparées, qui apparaissent plus d'une fois dans le corpus. Cette méthode

4. La statistique aborde l'acquisition de termes sous plusieurs angles. On peut distinguer deux tendances : l'une ne considère que les mots simples et étudie leur liaison à partir de leurs occurrences et cooccurrences dans le corpus ; l'autre considère le terme complexe comme un tout et étudie le comportement de ses occurrences en corpus.

privilégie les cooccurrences et met en évidence l'importance du contexte pour l'identification des termes. Elle utilise des textes ayant été préalablement lemmatisés afin de pouvoir regrouper des segments qui diffèrent d'un point de vue purement graphique, par exemple : *le problème financier* et *les problèmes financiers*. La sélection des segments répétés pertinents dépend fortement des objectifs de recherche visés. Dans une perspective d'acquisition terminologique par exemple, seuls les segments qui constituent des groupes nominaux bien formés sont conservés. Dans le but de créer un système d'aide à la construction de terminologie d'un domaine précis, Oueslati (Oueslati, 1999) utilise des critères morphosyntaxiques pour filtrer et structurer les résultats du calcul des segments répétés. Il utilise ensuite l'analyse distributionnelle pour identifier les relations sémantiques entre les termes propres au domaine étudié et pour constituer des classes de termes.

Ana est un outil dédié à l'apprentissage de concepts sans recours à l'analyse morphologique, syntaxique ou aux dictionnaires (Enguehard *et al.*, 1992). La reconnaissance des termes s'effectue au moyen d'égalités entre séquences de mots et d'une observation de répétition de séquences. A partir d'une liste de termes collectés à partir du corpus étudié, le système établit un noyau initial de structures utilisé comme amorce pour la recherche des structures similaires par enrichissement incrémental. Des composés sont ensuite repérés par l'exploitation des cooccurrences de deux mots pleins du noyau initial puis sont ajoutés au noyau. Les séquences d'un mot de schémas syntaxiques (*de, du, etc.*) et d'un terme reconnu permettent de repérer un mot inconnu qui est ajouté au noyau si ce dernier s'associe de façon répétitive aux premiers. Les termes repérés sont ensuite organisés en réseau dans lequel sont liés les termes partageant les mêmes têtes ou les mêmes arguments.

Approches hybrides

L'association d'analyses statistiques et linguistiques conduit à la réalisation de modèles hybrides. Ces deux analyses peuvent s'appliquer d'une manière alternative. Les études présentées ici adoptent un ordre qui varie. En effet, certains auteurs préfèrent filtrer les résultats de l'analyse statistique avec des critères linguistiques afin de ne conserver que les données qui sont informativement pertinentes. Alors que d'autres préfèrent effectuer le calcul statistique sur des données sélectionnées, par exemple par l'application de patrons syntaxiques.

Le logiciel *Acabit* (Daille, 1994; Daille, 1999) extrait des candidats termes à partir de corpus préalablement étiquetés. Il utilise des automates pour sélectionner des structures de compositions nominales candidates à décrire des termes. L'emploi de différents critères statistiques permet de déterminer si le lien entre les mots associés dans un composé est du type des relations du premier ordre décrit précédemment. A partir d'un corpus de référence et d'une liste de termes valides associée, la mesure *loglike* (Dunning, 1993) semble retenir le mieux les termes candidats. De la même façon, les travaux de (Church, 1988; Church *et al.*, 1991) utilisent un corpus préalablement éti-

queté, et un analyseur partiel *Fidditch* (Hindle, 1989) pour désambiguïser les contextes syntaxiques étudiés avant de calculer un score d'information mutuelle pour déterminer l'association entre mots.

Le système *Xtract* (Smadja, 1993) consiste à repérer les collocations en trois étapes. Partant d'un corpus étiqueté avec l'outil développé par Church (Church, 1988), il extrait les couples de mots (*bigrammes*) ayant une forte association entre eux selon une mesure basée sur l'information mutuelle. Les deuxième et troisième étapes sont effectuées parallèlement. L'une produit des collocations composées de plus de deux mots (*n-grammes*) en étudiant le contexte des couples de mots extraits en première étape. L'autre filtre les collocations produites par la première étape à l'aide de l'analyseur syntaxique *Cass* (Abney, 1991) qui permet de catégoriser les différents types de collocations, *i.e.*, nom+nom, adjectif+nom, sujet+verbe, verbe+objet, etc.

2.3.3 Structuration de la terminologie

Nous avons signalé dans la section précédente que l'organisation et la structuration des candidats termes extraits est utile pour la validation des travaux terminologiques. La structuration de termes se fait par la classification automatique de termes et le repérage de relations entre les termes. Cette tâche est difficile du fait que les relations que les termes entretiennent entre eux peuvent refléter des informations de nature variée : les relations de cooccurrence, de similarité, et d'équivalence (voir la section 2.3.1, les trois niveaux de relations selon Grefenstette). Un premier principe consiste à fouiller par des méthodes statistiques la distribution des ensembles de mots afin de déterminer les relations qui les lient.

Classification de termes : regroupement de termes en classes sémantiques

Partant de l'hypothèse que des termes qui ont des distributions comparables ont souvent un élément de sens commun (Harris *et al.*, 1989), dans les travaux de Toussaint (Toussaint *et al.*, 1998), des candidats termes extraits par *Acabit* et *Faster* sont regroupés en classes en fonction de leur cooccurrence dans le corpus. Cette classification est effectuée à l'aide d'un coefficient d'association entre des termes, basé sur leurs cooccurrences dans un même texte. Les classes de termes ainsi constituées sont ensuite présentées aux experts du domaine. Il leur est alors possible de structurer les éléments des classes à l'aide de relations sémantiques, *e.g.*, synonymie, hyponymie, ou encore méronymie.

En fait, comme nous l'avons déjà mentionné dans la section précédente, les approches purement statistiques mettent en évidence des associations entre les termes qui correspondent à plusieurs types de relations conceptuelles. Elles n'extraient pas véritablement de relations entre les termes mais des nuages de points dans lesquels seul un expert peut retrouver et identifier des relations (Sta & Chiao, 2001). Les résultats ainsi obtenus doivent donc être validés systématiquement.

Les expériences de Grefenstette sur la génération automatique de thésaurus (Grefenstette, 1994b) sont basées aussi sur la cooccurrence en y ajoutant des contraintes de sélection de contexte. A l'aide d'un analyseur syntaxique robuste *Sextant*, il exploite les distributions des contextes de type nom-adjectif, nom-verbe ou nom-nom, qui entretiennent des relations sujet-objet, verbe-objet et forme des classes de mots ayant une bonne probabilité de contenir les mêmes traits sémantiques. Les techniques utilisées consistent à rapprocher les termes qui ont des distributions syntaxiques analogues.

Dans le même but de construire automatiquement un thésaurus, les expériences de Lin (Lin, 1997; Lin, 1998) exploitent la dépendance syntaxique entre deux mots de différentes catégories (nom, verbe, adjectif/adverbe), et observent leur occurrences, *i.e.*, leurs distributions dans le corpus afin de calculer la similarité entre eux et de les classer ensuite en fonction de leur similarité.

Dans la continuité des travaux inspirés de la linguistique harrissienne, un ensemble d'études se sont intéressées à la composition des syntagmes nominaux et à la distribution de leurs unités dans le corpus.

L'outil *Zellig* (Habert *et al.*, 1996) exploite les relations de dépendances syntaxiques entre les composants des groupes nominaux d'un corpus. Il simplifie les arbres d'analyse syntaxique fournis par l'extracteur *Lexter* en se limitant à des patrons (nom-prep-nom) et (nom-adjectif). Ces arbres élémentaires permettent ensuite de constituer des classes de contextes (à gauche et à droite) afin de mettre en évidence les comportements syntaxico-sémantiques et de calculer la proximité de deux mots sur la base du nombre de contextes partagés. Les résultats montrent que les classes contextuelles permettent de repérer rapidement des mots qui sont fortement associés mais les relations ainsi obtenues sont variées : antonymie (*résiduel/sévère*), synonymie (*lésion/sténose*), etc. Elles permettent cependant d'associer aux mots des champs d'attributs communs et caractéristiques, ce qui permet à un expert d'interpréter et de valider facilement les classes proposées.

Selon le même principe, l'outil *Lexiclass* (Assadi, 1997) exploite la distribution interne des termes pour mesurer leur proximité sémantique. A l'aide de méthodes de classification hiérarchique ascendante, il regroupe les unités (noms, adjectifs ou syntagmes) qui se trouvent dans des positions syntaxiques analogues au sein des syntagmes nominaux extraits par *Lexter*. Les termes *système d'alimentation nucléaire* et *système de réfrigération nucléaire* sont ainsi regroupés en fonction de la distribution de leurs contextes adjectivaux.

Les travaux que nous venons de citer permettent de regrouper des termes qui partagent de façon significative les mêmes contextes. Les techniques employées visent à relever des traits pertinents qui fondent l'appartenance à une classe sémantique. Ces outils proposent des hypothèses de relations, sans permettre d'identifier le type précis de ces liens qui nécessitent d'être interprétés. L'interprétation de la nature des relations entre les termes d'une classe est assistée par l'analyse de la distribution contextuelle de l'ensemble des termes de la classe. Ces méthodes dépendent en fait directement de la distribution de fréquence des termes dans le corpus. Plus les termes sont fréquents, plus

nombreux sont les indices contextuels qu'ils présentent, ce qui facilite le traitement (Grefenstette, 1996).

Classification de termes : regroupement de termes en classes de concepts

La diversité des classes sémantiques issues de regroupement par contextes syntaxiques a conduit certains à l'exploitation des sources de connaissances préalables comme des dictionnaires ou thésaurus. Les expériences de Resnik (Resnik, 1993; Resnik, 1995) exploitent ainsi l'analyse distributionnelle vue précédemment en remplaçant les mots de contexte par leurs classes sémantiques afin de mettre en évidence les types de relations sémantiques associées. On peut ainsi déterminer dans une structure argumentale (verbe+nom), la classe sémantique la plus pertinente pour l'argument, en fonction des contraintes sémantiques établies par le verbe sur leurs arguments[5]. Ces classes sont obtenues en remontant les liens génériques de *WordNet* (Miller *et al.*, 1990).

Selon le même principe, dans le cadre de l'extraction d'informations, les méthodes développées par Riloff (Riloff, 1993) s'attachent à générer automatiquement des patrons syntaxiques discriminants des instances d'une classe conceptuelle dans un corpus de domaine spécialisé. A partir d'un dictionnaire contenant un ensemble de syntagmes associés à un concept, on cherche à acquérir des schémas d'extraction de membre de classes conceptuelles. Étant donnée une classe conceptuelle (*cibles des terroristes*) et un terme instanciant cette classe (*ambassade*), une analyse syntaxique est ensuite effectuée sur les phrases qui contiennent ce terme, *e.g., L'ambassade a été bombardé*, pour vérifier si les patrons ainsi repérés, correspondent aux heuristiques prédéfinies dans le dictionnaire, telles que *sujet-verbe au passif* auquel est associé les termes appartenant à une classe, *e.g., cibles terroristes*. L'identification d'une heuristique permet de suggérer le futur schéma candidat et de repérer un terme comme instance de la classe conceptuelle. Par exemple, le schéma ainsi défini permet de déduire que le terme *Maison Blanche* dans la phrase *Maison Blanche a été bombardée...* appartient à la classe des *cibles des terroristes*.

Repérage de relations sémantiques

Il existe des approches qui travaillent au niveau des occurrences elles-mêmes afin de repérer les différents types de relations entre les termes. Elles exploitent la structure interne de termes répertoriés comme susceptibles de marquer un type de relation précis entre deux éléments. Ces travaux sont basés sur l'hypothèse que, pour une relation sémantique observable dans un texte, il peut y avoir des formules linguistiques

5. Par exemple, la classe sémantique du mot *baseball* dans les associations *hit baseball, play baseball, watch baseball*, qui est respectivement OBJET, GAME, RECREATION, est déterminée selon les restrictions d'association des verbes *hit, play, watch*.

récurrentes, *i.e.*, marqueurs, qui expriment cette relation. L'un des enjeux principaux concerne la généralité des relations, et celle des marqueurs décrivant ces relations.

D'un côté, les travaux qui exploitent des relations jugées toujours pertinentes, *e.g.*, les relations d'hyperonymie, de synonymie ou d'hyponymie, pour décrire un domaine de connaissance visent à définir des marqueurs ainsi considérés généraux pour extraire ces relations (Kavanagh, 1995; Garcia, 1998).

A l'opposé de ces travaux qui consistent à mettre au point des marqueurs a priori pour les projeter sur un corpus de domaine spécialisé, se trouvent les travaux qui supposent que la spécificité des domaines techniques se manifeste au niveau du langage observable dans ces domaines et donc dans leurs corpus. Chaque corpus ayant ainsi sa spécificité linguistique, il est possible que les marqueurs susceptibles d'identifier une relation considérée générale diffèrent d'un corpus à l'autre.

L'apprentissage inductif de ces marqueurs est alors nécessaire pour découvrir les relations vraiment présentes dans un corpus. Il s'agit plus précisément d'acquérir des marqueurs à partir du corpus étudié. Les travaux de Hearst (Hearst, 1992) sur l'extraction automatique des liens d'hyperonymie font figure de référence. Partant d'une liste pré-établie de patrons lexico-syntaxiques pertinents pour cette relation, l'observation des occurrences de ces patrons dans un corpus permet d'extraire de nouveaux patrons qui seront utilisés pour trouver d'autres occurrences. Les travaux qui suivent adoptent tous le principe d'une recherche itérative alternée dans le corpus à la fois des patrons d'une relation donnée et des couples de termes qui sont liés par cette relation (Condamines & Rebeyrolle, 1997; Morin, 1998; Séguéla & Aussenac, 1999).

2.3.4 Application à la recherche d'information

L'objectif de la recherche d'information est d'identifier de façon efficace dans une collection de documents ceux qui correspondent à des requêtes exprimées par des mots clefs ou en langue naturelle. L'approche générale consiste à indexer préalablement tous les documents de la base et à élaborer ensuite des stratégies d'appariement entre une requête et les index d'un document.

Les techniques développées pour l'acquisition de terminologie présentées précédemment sont de plus en plus exploitées en recherche d'information, notamment pour l'extension de requête (*Query Expansion*). Celle-ci a pour but d'améliorer la performance des systèmes de recherche, c'est-à-dire de trouver plus de documents pertinents pour une requête donnée en utilisant des termes associés à cette requête. Les termes associés sont utilisés dans l'appariement.

Les expériences dans ce domaine se sont surtout intéressées aux méthodes de regroupement de termes similaires. Le principe de ces méthodes consiste à établir une représentation signifiante pour chaque terme et à les classer ensuite à l'aide des mesures de similarité. Cette représentation de terme peut être définie soit par l'analyse de cooccurrences, soit par l'analyse de contextes.

L'analyse de cooccurrences regroupe les mots sur la base de leur présence si-

multanée répétée dans des documents (Choueka *et al.*, 1983; Choueka, 1988; Oueslati, 1999; Lebart & Salem, 1988; Lebart & Salem, 1994), tandis que l'analyse de contextes les regroupe en fonction de leur cooccurrence dans des contextes grammaticaux (Grefenstette, 1994b; Habert *et al.*, 1996; Lin, 1998; Assadi, 1997). Les résultats obtenus par ces méthodes nécessitent un post-traitement qui consiste en une interprétation car les regroupements sont trop larges, entretenant des relations variées syntaxiques ou sémantiques (section 2.3.3). Ces approches sont utilisées pour l'extension de requête en recherche d'information monolingue ou translangue (abordée plus en détail au chapitre 5) et peuvent effectivement améliorer la performance de ces systèmes (Sparck Jones & van Rijsbergen, 1975; Qiu & Frei, 1993; Xu & Croft, 1996; Schütze & Pedersen, 1997; Carbonell *et al.*, 1997; Brown, 1998; Yang *et al.*, 1998).

2.4 Acquisition terminologique bilingue

L'objectif de l'extraction de lexique bilingue à partir de corpus consiste à identifier et extraire les termes et leurs traductions. Il s'agit plus précisément de repérer les termes des textes sources et des textes cibles, puis de les mettre en correspondance. Les traitements effectués peuvent ainsi se décomposer en deux étapes. La première correspond à celle de l'extraction de terminologie monolingue et la seconde à celle de l'alignement ou plutôt de l'appariement[6].

Selon Véronis (Véronis, 2000a), ces deux étapes ne peuvent pas être totalement modularisées dans la pratique : "la détermination des unités dans la langue source est dépendante de la langue cible (par exemple, il faut aligner d'un bloc *demande de brevet* et *patentanmeldung* alors que l'alignement peut se fractionner avec *demanda di brevetto*).".

L'utilité des corpus pour l'acquisition automatique de données terminologiques bilingues a été mentionnée et validée depuis un certain temps (Atkins, 1990) et les projets de dictionnaires comme *Oxford-Hachette French Dictionary* (Grundy, 1996) ou *Dictionnaire Canadien Bilingue* (Roberts & Montgomery, 1996) ont fait appel aux corpus bilingues.

2.4.1 Corpus parallèles et corpus comparables

Deux types de corpus bilingues sont exploités pour l'extraction de lexique bilingue : le corpus parallèle et le corpus comparable. Les corpus parallèles sont constitués de textes sources et de leurs traductions. Un corpus comparable désigne un ensemble de textes de langues différentes rassemblés selon des critères similaires, en ce

6. Pour ne pas confondre avec l'alignement de textes dont l'objectif consiste à aligner les textes à différents niveaux : sections, paragraphes, phrases, expressions ou mots. L'extraction de lexique bilingue correspond à l'alignement de textes au niveau des termes, quelle que soit leur forme : composée ou simple.

qui concerne le domaine, le genre, la date de publication, etc. Si l'extraction de lexique bilingue à partir de corpus parallèles est maintenant beaucoup étudiée et développée (van der Eijk, 1993; Smadja *et al.*, 1996; Dagan & Church, 1997; Resnik & Melamed, 1997; Fung & McKeown, 1997; Hiemstra *et al.*, 1997; Hiemstra, 1998; Gaussier, 1998), l'extraction de lexique bilingue à partir de corpus comparables est plus récente (Fung & Yee, 1998; Rapp, 1999; Déjean & Gaussier, 2002; Chiao & Zweigenbaum, 2002a).

La performance de l'extraction à partir de corpus parallèles dépend principalement de la qualité de l'alignement entre les textes. Les expériences en extraction de lexique bilingue à partir de corpus parallèles généralement alignés au niveau de phrases ont montré des résultats satisfaisants. Néanmoins l'utilisation de corpus parallèles présente quelques contraintes : outre la qualité de l'alignement, les couples de langues concernés et la taille pour l'instant modeste des corpus parallèles par rapport aux corpus monolingues, leur représentativité est généralement limitée aux domaines de spécialité (articles scientifiques, techniques, etc).

En revanche, il est plus facile d'accéder à un corpus comparable dans un domaine donné qu'à un corpus parallèle de bonne qualité (Fung & Yee, 1998). De plus du point de vue terminologique, les usages réels des termes dans les deux parties monolingues de corpus comparables sont bien conservés puisqu'ils n'ont pas subi de transformation due à la traduction : le vocabulaire de la langue source influence lors d'une traduction le choix d'équivalents en langue cible du traducteur. Les recherches récentes se sont ainsi intéressées à l'exploitation de corpus comparables.

Néanmoins à cause de certaines caractéristiques propres aux corpus comparables, il est plus difficile de leur appliquer les méthodes statistiques utilisées pour l'alignement de corpus parallèles (Fung, 1998). Par exemple les traductions d'un terme du corpus de langue source peuvent être absentes dans le corpus de langue cible, les fréquences et les positions d'occurrences ne sont pas homogènes et comparables dans les corpus comparables.

2.4.2 Acquisition de lexique bilingue en corpus parallèles

Les techniques d'alignement au niveau des termes que nous décrivons par la suite sont développées pour une application sur des corpus parallèles alignés au niveau des phrases. Il nous semble plus intéressant de les présenter pour certaines raisons pratiques. Tout d'abord, l'une des sources importantes de textes parallèles est la mémoire de traduction, constituée souvent des phrases qui sont traductions l'une de l'autre, stockées dans des systèmes de traduction automatique. La plupart des systèmes d'alignement de textes travaillent à ce niveau et le développement de techniques d'alignement de textes au niveau de phrases semblent parvenu à maturité : les systèmes atteignent plus de 98.5% d'efficacité (Véronis, 2000a).

En général, les techniques d'alignement exploitent différents types d'information : longueur des phrases, dictionnaires bilingues, distributions lexicales et cognats,

c'est à dire les occurrences qui sont identiques ou se ressemblent graphiquement, etc. Nous trouvons un état de l'art sur les systèmes d'alignement dans la thèse de Kraif (Kraif, 2001a) (ch.2) et l'évaluation des systèmes existants effectuée dans le cadre du projet *ARCADE* (Véronis & Langlais, 2000). Nous verrons que certains paramètres cités ci-dessus ont également alimenté les algorithmes d'appariement de termes dans les corpus parallèles.

Problématique de l'alignement

Aux alentours de 1990, une équipe de recherche chez IBM s'est penchée sur des modèles purement statistiques de traduction automatique basée sur l'apprentissage sur corpus parallèles (Brown *et al.*, 1990; Peter F. Brown & Mercer, 1994). Les autres recherches utilisent des textes parallèles pour extraire des lexiques bilingues, à l'aide de dictionnaires bilingues en combinaison avec l'analyse statistique (Catizone *et al.*, 1989) ou l'analyse syntaxique (Klavans & Tzoukermann, 1990).

Certaines études s'intéressent à l'alignement au niveau des mots simples par des méthodes statistiques (Dagan *et al.*, 1993; Wu & Xia, 1994; Resnik & Melamed, 1997). Cependant, ces techniques qui mettent en jeu des occurrences isolées ne prennent pas en compte des phénomènes courants présents dans un texte : termes complexes, collocations, expressions, etc. La plupart des causes d'erreur d'alignement concernent les variations linguistiques, *i.e.*, les flexions, les substitutions pronominales, les mots composés, les expressions semi-figées avec insertion ou suppression d'adjectif et d'adverbe, la passivation, etc.

Dans le cas de l'alignement d'unités complexes, le traitement s'effectue de deux façons : soit en les segmentant en mots simples soit les considérant comme une seule unité. On distingue en général trois méthodes principales d'extraction de terminologie bilingue à partir des corpus parallèles que nous présentons maintenant.

Extraction parallèle des termes dans les langues source et cible et alignement

L'objectif des expériences de Van der Eijk (van der Eijk, 1993) est d'automatiser l'acquisition de lexique bilingue. Son approche suit le schéma classique en effectuant d'abord l'extraction des syntagmes nominaux des textes de chaque langue. Les termes extraits de chacun des corpus sont alignés sur la base de cooccurrences de termes dans des phrases alignées. Les traductions ainsi obtenues pour chaque terme sont pondérées par une mesure qui tient compte de la position attendue du terme dans les données cibles. Cette position est attribuée au terme cible par rapport à la taille des couples de phrases alignées, en tenant compte de la position du terme source dans la phrase source.

L'outil *Termight* de Dagan et Church (Dagan & Church, 1997) se compose également d'une procédure d'extraction terminologique monolingue et d'une procédure d'alignement bilingue. Les candidats termes sont identifiés et extraits à partir de textes

étiquetés à l'aide des patrons syntaxiques. Ils sont validés ensuite par un expert terminologue. L'appariement bilingue se fait au niveau des mots. La traduction candidate est définie comme une séquence composée du premier jusqu'au dernier mot aligné avec n'importe quel mot du terme source.

Divers auteurs proposent la combinaison des méthodes statistiques et des analyses morphosyntaxiques pour extraire la traduction des unités complexes de textes parallèles (Kupiec, 1993; Daille *et al.*, 1994; Gaussier & Langé, 1995). Certains chercheurs, notamment dans le domaine de la traduction automatique basée sur les exemples, ont mis en avant des méthodes qui nécessitent des analyses syntaxiques monolingues sur les phrases alignées, et tentent de mettre en correspondance les unités composantes de structures similaires à l'aide des dictionnaires bilingues (Kaji *et al.*, 1992; Matsumoto *et al.*, 1993).

Extraction des termes dans une langue et alignement avec des séquences dans une autre langue

Cette approche est plus souple du point de vue pratique par rapport aux approches classiques citées ci-dessus. Elle peut être plus facilement appliquée dans le cas où l'extraction de termes monolingues n'est disponible ou pertinente que pour une des deux langues (la langue source ou cible). Par exemple pour le couple de langues anglais - français, les termes sont plus faciles à identifier en anglais qu'en français. Un autre intérêt est que cela permet de prendre en compte le fait que certaines unités complexes sont figées dans une langue et ne peuvent pas être traduites mot à mot dans une autre langue.

L'application de cette approche donne en général un taux de rappel plus élevé que la méthode classique (extraction parallèle des termes monolingues suivie par l'alignement bilingue). L'alignement des termes peut produire comme traduction des séquences de mots qui ne sont pas forcément des termes simples ou complexes.

Partant d'une liste de collocations dans la langue source (anglais), le système *Champollion* (Smadja *et al.*, 1996) tente d'identifier et regrouper les mots de la langue cible qui peuvent constituer une partie de la traduction d'une collocation donnée. La liste des mots retenue est calculée par la mesure de similarité *coefficient de Dice*[7] :

$$Dice(X, Y) = \frac{2f_{xy}}{f_x + f_y}$$

Les mots sont ensuite combinés de façon itérative en séquences de mots, et à chaque itération est conservée la séquence ayant le meilleur score de similarité. Parmi les séquences proposées, l'outil choisit comme traduction optimale la séquence qui a le score de similarité le plus élevé. Dans le cas de la traduction d'une collocation figée, un algorithme est appliqué après la sélection pour rétablir l'ordre des mots au sein de la collocation.

7. f_{xy} est ici la fréquence absolue de cooccurrences de deux variables X et Y, f_x et f_y sont la fréquence absolue d'occurrences de X et Y respectivement.

Le processus d'alignement proposé par Gaussier (Gaussier, 1998) est présenté comme un problème de flot dans un graphe. La technique utilisée consiste à rechercher des flots maxima dans un réseau de flots, à l'aide d'un modèle de traduction statistique qui mesure les probabilités d'alignement entre mots simples et composés.

Les études montrent l'importance de la combinaison des différentes sources d'informations (indices lexicaux, cognats, longueurs de segments) avec des méthodes efficaces de filtrage et de réduction de l'espace de recherche (Hiemstra, 1998; Melamed, 2000) afin d'améliorer la performance d'alignement.

Extraction et alignement simultanées

Le modèle *Inversion Transduction Grammars* (ITG) expérimenté dans les travaux de Wu (Wu, 1997; Wu, 2000) génère simultanément des paires de structures syntaxiques bilingues. Selon l'auteur, la différence entre son modèle et celui qu'il qualifie de *parse-parse-match* (Kaji *et al.*, 1992; Matsumoto *et al.*, 1993) est que dans le dernier modèle, l'analyse syntaxique s'effectue parallèlement sur chaque partie monolingue de textes alignés. Le principe des ITG est d'extraire des patrons syntaxiques bilingues dans lesquels l'ordre des constituants peut être inversé en fonction du couple des langues traitées[8]. Ces patrons sont ensuite utilisés pour trouver des équivalences de traduction au niveau des mots ou des syntagmes.

Dans ces expériences, différentes variantes des ITG sont proposées, notamment une version stochastique (SITG) dans laquelle à chaque règle de réécriture est associée une probabilité estimée à l'aide de la mesure du *maximum de vraisemblance*.

2.4.3 Acquisition de lexique bilingue en corpus comparable

Nous avons mentionné au début de la section 2.4.1 la distinction entre corpus parallèles et corpus comparables. Contrairement aux textes parallèles qui sont des traductions l'un de l'autre, les corpus comparables sont un ensemble de textes liés, sans qu'ils soient des traductions réciproques, par une relation d'identité. La notion d'identité ici est floue, mais elle rend bien compte du continuum qui existe entre des corpus parallèles bruités (Véronis & Langlais, 2000), des corpus comparables et des corpus non reliés. La relation d'identité peut être temporelle, par exemple des textes rédigés et publiés pendant une même période, ou encore liée au vocabulaire pour des corpus traitant des mêmes sujets ou domaines.

Nous avons aussi expliqué les raisons pratiques et théoriques pour lesquelles les travaux en acquisition lexicale commencent à s'intéresser à l'extraction de traductions à partir de corpus comparables (section 2.4). Pourtant, l'identification d'équivalents de traduction dans ce genre de textes présente de toute évidence une tâche plus complexe et ambitieuse que l'exploitation des textes parallèles. Elle reste pour l'instant du domaine de la recherche.

8. Il s'agit du couple anglais-chinois dans ces expériences.

Tous les travaux réalisés sont fondées sur la même hypothèse, celle de la sémantique distributionnelle. Cette hypothèse suppose que le sens d'un mot peut être décrit par la distribution de ses occurrences dans un ensemble de contextes (Rajman & Bonnet, 1992).

Dans un contexte multilingue, cette hypothèse peut être reformulée ainsi : un mot de la langue A dont la distribution est similaire à celle d'un mot de la langue B est, avec une forte probabilité, traduction de ce mot. La mise en correspondance entre deux mots de langues différentes est ainsi réalisée au niveau sémantique, et le corpus bilingue est considéré comme un objet d'acquisition de connaissances et de mise à jour de ressources lexicales existantes (Déjean & Gaussier, 2002).

Partant du principe que les mots qui ont une distribution similaire sont des traductions réciproques, l'approche suivie dans les travaux réalisés jusqu'à aujourd'hui (Fung & Yee, 1998; Rapp, 1999; Chiao & Zweigenbaum, 2002a; Déjean & Gaussier, 2002) consiste à d'abord établir les distributions des mots de la langue source et de la langue cible à partir de corpus comparables. Une distribution est définie par un vecteur de contexte constitué de mots qui cooccurrent avec le mot étudié. L'empan de cooccurrence est souvent une fenêtre graphique, *i.e.*, n mots à gauche et à droite.

Les vecteurs de contexte sont ensuite traduits d'une langue à l'autre à l'aide de lexiques bilingues partiels (Fung & Yee, 1998; Rapp, 1999; Chiao & Zweigenbaum, 2002a). Des méthodes statistiques (en général des mesures de similarité) sont utilisés pour calculer la ressemblance entre les vecteurs transférés d'une langue et originaux de l'autre langue. Les candidats à la traduction sont les termes dont les vecteurs de contexte ont les meilleurs scores de similarité.

Au lieu de traduire directement les vecteurs de contexte, Déjean et Gaussier (Déjean & Gaussier, 2002) exploitent la similarité distributionnelle au sein d'une même langue pour mesurer ensuite la correspondance entre les mots des langues différentes à l'aide d'un thésaurus bilingue.

Nous décrivons en détail dans le chapitre 3 certains des différents paramètres et les mesures de similarité utilisés dans ces travaux et par conséquent nous ne détaillons pas ici les formules des mesures employées.

Travaux de Reinhard Rapp

Les premières expériences de Rapp (Rapp, 1995) sur l'identification des traductions à partir de corpus comparables reposent sur l'analyse purement statistique des cooccurrences de mots dans un corpus comparable en allemand-anglais. Dans ses dernières expériences (Rapp, 1999), des matrices de cooccurrences sont définies par une fenêtre de n mots (n = 12). Les analyses linguistiques, *e.g.*, lemmatisation, suppression des mots grammaticaux, etc. sont effectuées pour nettoyer les corpus de taille très importante et réduire la complexité du calcul: 135 millions de mots pour le corpus allemand et 163 millions de mots pour le corpus anglais. Les corpus sont constitués des articles du journal allemand *Frankfurter Allgemeine Zeitung* couvrant la période

de 1993 à 1996 et du journal *Guardian* en anglais de 1990 à 1994. Il faut noter ici que la comparabilité entre les deux corpus est réduite du point de vue temporel puisqu'ils se chevauchent à peine.

L'approche consiste à construire d'abord les matrices de cooccurrences de tous les mots de la langue cible. Les cooccurrences sont filtrées par un dictionnaire bilingue de 16 380 entrées composées de mots simples et pondérées sur la base de leur force d'association dans le corpus. Pour un mot source à traduire, il construit ensuite le vecteur de cooccurrences à partir du corpus source et le traduit en langue cible avec le dictionnaire. Le vecteur du mot source traduit est ensuite comparé avec la matrice de tous les mots du corpus cible par la mesure de distance *city-block-metric,* aussi appelé *distance de Manhattan* que nous décrivons dans la section 3.2.3. Les résultats affichent une performance de 72% de mots bien traduits (par rapport à l'ensemble des mots à traduire) en ne considérant que le premier candidat et de 89% en considérant les 10 premiers candidats pour une évaluation sur une liste de 100 mots allemands fréquents dans le corpus.

Travaux de Pascale Fung

L'approche utilisée dans (Fung & McKeown, 1997; Fung & Yee, 1998) exploite elle aussi le modèle vectoriel (utilisé en recherche d'information) pour identifier les traductions de néologismes, *i.e.,* les nouveaux mots qui sont absents des ressources lexicales existantes. Elle part d'une liste bilingue de couples de mots pivots (*seed words*) établie à partir de dictionnaires en ligne en ne gardant que les entrées présentes et fréquentes dans les deux parties monolingues du corpus. Pour chaque mot inconnu de la langue A, il s'agit de créer un vecteur d'association en calculant un taux d'association à l'aide de *tf.idf* , utilisée notamment en recherche d'information, avec chaque mot pivot en langue A. Le même vecteur est construit pour les mots inconnus de la langue B, qui contient tous les mots pivots de la langue B. Les vecteurs d'association (langues A et B) sont ensuite comparés sur la base de la mesure de similarité $cosinus$ dans l'espace vectoriel des mots pivots.

Il faut mentionner que les corpus utilisés sont lemmatisés et désambiguïsés. Les mots pivots utilisés sont limités aux noms, verbes et adjectifs, lemmatisés et non polysémiques. Pour des expériences sur le couple de langues japonais-anglais, les auteurs utilisent le journal japonais *Nikkei Financial News Material* entre Janvier et Mars 1994, et le journal anglais *Wall Street Journal material* de même période (Fung & McKeown, 1997), sans préciser la taille de ce corpus. Ils obtiennent une moyenne de 30% de performance au premier rang (nombre de traductions trouvées sur nombre de mots à traduire) qui monte à 50% en considérant les 20 premiers candidats.

Dans des expériences sur le couple de langues chinois-anglais (Fung & Yee, 1998), la taille des corpus utilisés est de 3 millions d'octets pour les textes anglais tirés du journal *Hong Kong Standard* et de 8,8 millions d'octets pour les textes chinois du journal *Mingpao* de la même période. Selon les auteurs, ils sont comparables en

taille puisqu'un caractère chinois compte 2 octets, ce qui donne un corpus anglais de 3 millions de caractères et un corpus chinois de 4,4 millions de caractères[9].

Travaux de Hervé Déjean et de Éric Gaussier

L'expérience de Déjean et Gaussier (Déjean & Gaussier, 2002) repose sur l'analyse de cooccurrences et l'usage d'un thésaurus bilingue. Le principe est similaire à ceux que nous avons décrits précédemment. La différence se manifeste dans l'usage des vecteurs de cooccurrences.

Le vecteur de cooccurrences d'un mot que l'on cherche à traduire est d'abord comparé avec celui des entrées dans la même langue du thésaurus. La mesure de similarité est le cosinus (section 3.2.3). Au sein d'une même langue, un mot est ainsi lié à des mots proches au sens du cosinus et par conséquent aux classes conceptuelles associées du thésaurus. Nous sommes donc en présence de mots des deux langues proches de classes conceptuelles du thesaurus. L'étape suivante utilise un modèle probabiliste pour probabiliser la proximité d'un mot à une classe conceptuelle puis pour estimer la probabilité que deux mots de langues différentes soient traductions mutuelles. Les mots du corpus en langue cible partageant le plus grand nombre d'entrée du thésaurus avec le mot à traduire sont des candidats à la traduction avec de fortes probabilités.

Les résultats donnent un score de 79% (nombre de traductions trouvées sur nombre de mots à traduire) pour une liste de 180 mots fréquents dans un corpus général (9 millions de mots au total) en considérant les 10 premiers candidats. Avec un corpus médical, ils affichent un score de 63% pour une liste de 50 entrées du thésaurus médical *MeSH*.

2.5 Conclusion

Nous avons montré dans ce chapitre l'avantage des corpus pour l'acquisition de lexique à travers l'exploitation des notions de sous-langage et de corpus de domaine.

Dans le cadre monolingue, la problématique de l'acquisition de lexique rejoint celle de l'extraction terminologique. Celle-ci s'effectue en général en deux étapes : l'extraction et la structuration de termes. A partir des différents types de relations qu'entretiennent les termes en corpus, nous avons présenté les approches utilisées pour extraire les candidats termes à partir de corpus et pour organiser et structurer les termes extraits.

Dans le cadre multilingue, l'acquisition de lexique consiste à extraire des termes et leur traductions. On distingue deux types de corpus utilisés pour l'extraction lexicale bilingue : le corpus parallèle et le corpus comparable. Les corpus parallèles contiennent

9. La façon d'établir ici la taille des corpus est discutable car un mot chinois est en général composé de deux caractères (bigramme) alors qu'un mot anglais est en général composé de quatre ou cinq caractères.

des textes qui sont traductions mutuelles. Des textes comparables sont des textes de langues différentes regroupés selon des critères similaires concernant le domaine, le genre, la date de publication, etc.

La performance de l'acquisition lexicale bilingue en corpus parallèles dépend en principe de la qualité de l'alignement des textes. Les travaux réalisés dans ce cadre ont montré des résultats satisfaisants. Néanmoins l'utilisation de corpus parallèles présente des contraintes liées à leur taille plus modeste par rapport aux corpus monolingues, à leur manque de disponibilité, à la qualité de l'alignement des corpus et aux couples de langues concernés, etc.

Les corpus comparables présentent en revanche l'avantage d'être plus accessibles que les corpus parallèles de bonne qualité, et de mieux préserver l'usage réel des termes dans la langue étudiée sans avoir subi de transformation due à la traduction. Néanmoins la difficulté à traiter les corpus comparables est supérieure de plusieurs ordres de grandeurs à la difficulté à traiter des corpus alignés. Elle est principalement due à certaines caractéristiques propres aux corpus comparables. L'espace de recherche de la traduction d'un mot ne peut pas être réduit dans un corpus comparable, contrairement à un corpus parallèle où l'espace de recherche est réduit aux segments alignés.

C'est dans la perspective de l'acquisition lexicale bilingue à partir de corpus comparables dans le domaine médical que nous situons cette étude. Il s'agit d'une problématique encore mal résolue mais qui s'avère fructueuse étant donné la disponibilité de telles ressources.

Chapitre 3

Acquisition lexicale bilingue à partir de corpus comparables : une nouvelle approche

3.1 Introduction

Dans la pratique, l'extraction automatique de lexique bilingue à partir de corpus se compose de deux tâches : une étape de segmentation qui aboutit à la détermination des unités à traduire (mots sources) ainsi que des unités candidates à la traduction (mots cibles), et une étape d'appariement de ces unités. Après avoir présenté le problème de la segmentation ou plus précisément celui de l'extraction des termes du point de vue terminologique dans le chapitre 2, nous nous intéressons dans ce chapitre à l'appariement des mots sources et cibles.

Nous avons montré dans la section 2.4.3 que les travaux d'acquisition automatique de lexique bilingue à partir de corpus comparables reposent sur l'idée que la relation de traduction entre langues (comme les relations paradigmatiques au sein d'une langue, *i.e.*, les mots voisins qui sont sémantiquement proches) peut être mise en évidence par la comparaison des distributions des mots dans les corpus. L'hypothèse sous-jacente est que le sens d'un mot peut être déterminé en contexte et donc, en simplifiant, par l'ensemble des mots qui figurent dans ses contextes (Habert *et al.*, 1997).

Cette idée peut alors être exploitée pour dériver automatiquement la sémantique d'un mot à partir de l'ensemble de ses contextes dans un corpus. Partant de ce principe, la tâche consiste alors à définir d'abord un ensemble de contextes en fonction desquels la distribution de chaque mot sera calculée, puis à appliquer une mesure de similarité entre distributions. Notons qu'un élément essentiel de cette démarche est l'utilisation d'un lexique bilingue partiel et préexistant qui joue le rôle de pont entre les langues et notamment entre les distributions. La démarche est en général décomposée en trois

étapes (Habert *et al.*, 1997).

Définition du contexte d'un mot en fonction du corpus exploité et des relations sémantiques recherchées.

Représentation des mots par leur lien d'association, calculé en fonction du contexte défini.

Choix de la mesure de similarité entre les représentations des mots afin de construire des classes de mots en fonction de l'application visée.

Lorsque la relation sémantique que l'on cherche est celle d'équivalence traductionnelle à travers des langues, l'espace de recherche des unités lexicales à mettre en correspondance est réduite à la phrase dans le cas des corpus parallèles. Dans le cas des corpus comparables, aucune contrainte d'alignement ne réduit l'espace de recherche. Tous les segments possibles d'une langue peuvent être mis en correspondance avec n'importe quelle unité d'une autre langue. Pour pallier cette difficulté on exploite l'idée que les contextes dans lesquels apparaît le mot b traduction du mot a doivent être similaires à ceux dans lesquels apparaît le mot a. Dans ce cas, les contextes d'une langue doivent être traduits dans une autre langue afin de pouvoir reconstituer l'espace de recherche. Le passage d'une langue à une autre se fait par l'intermédiaire des ressources lexicales bilingues, *i.e.*, dictionnaires, thésaurus, etc.

Nous abordons dans un premier temps les différents paramètres (section 3.2) qui permettent de mettre en évidence la relation de traduction entre deux mots : contexte, pondération, similarité, etc. Nous développons ensuite (section 3.3) notre modèle d'extraction de lexique bilingue à partir de corpus comparables.

3.2 Mise en évidence de la relation de traduction à partir des contextes

3.2.1 Contexte de cooccurrence

Le choix du contexte de cooccurrence dépend de ce qu'il est censé mettre en évidence et aussi de la nature du corpus traité. Ainsi, la cooccurrence entre deux unités définie dans un texte n'a pas la même interprétation que celle définie dans une phrase.

Le contexte le plus grand est le document lui-même. Il pose en soi un certain nombre de problèmes. Dans un corpus, la taille des documents peut être extrêmement variable : du court résumé au document de plusieurs pages. Les textes longs affichent des cooccurrences peu signifiantes et les textes courts contiennent peu de cooccurrences.

Le paragraphe, repéré par un passage à la ligne, est intéressant pour deux raisons. Sa taille est en général homogène : de une à quelques phrases. Le découpage en paragraphes est aussi l'expression d'une homogénéité de contenu. Un paragraphe constitue

donc souvent une unité de sens homogène, ce qui laisse présager des liaisons fortes entre les mots.

La phrase est un bon contexte de cooccurrence. Elle est le lieu idéal où l'auteur met en rapport, notamment syntagmatique, les unités lexicales. La cooccurrence de deux mots dans la phrase peut être l'expression d'une relation syntagmatique stable comme d'une relation paradigmatique. Le contexte peut y être syntaxique, c'est-à-dire contraint par les relations syntaxiques (Nom prép. Nom...). Toutefois, l'utilisation de contextes syntaxiques pose des problèmes lorsque la relation sémantique recherchée n'est pas limitée à un cadre monolingue. C'est le cas notamment de la traduction pour laquelle un mot d'une catégorie grammaticale peut être traduit par un mot d'une autre catégorie grammaticale (*e.g.*, l'adjectif *cardiaque* dans *crise cardiaque* et traduit par le nom *heart* dans *heart attack*). Notons également que la segmentation automatique en phrases pose problème. Les algorithmes de segmentation ajoutent en général du bruit à cause de l'ambiguïté des marqueurs typographiques de séparation : ponctuation, majuscule. En particulier, la suite point-espace-majuscule apparaît aussi bien en fin de phrase que dans des abréviations.

Enfin, le contexte peut être réduit à une fenêtre de quelques mots ou même à deux mots. Dans ce dernier cas, la cooccurrence est utilisée pour la mise en évidence de bigrammes. L'ordre d'apparition ou la position peuvent aussi être utilisés : on parle alors de contexte droit ou gauche.

On le voit, la notion de contexte peut être définie de plusieurs manières. Plusieurs paramètres interviennent : la taille du contexte, l'ordre des mots, leurs positions, les contraintes syntaxiques... Il semble que la taille de la fenêtre a une influence sur le type de relation sémantique visée : « la taille de fenêtre dépend des relations sémantiques que l'on étudie, les cooccurrences à petite, moyenne et grande distance tendant respectivement à faire ressortir des expressions figées ou semi-figées [...], des contraintes de sélection [...] et des mots appartenant au même champ sémantique. » (Habert *et al.*, 1997).

Comme beaucoup d'autres travaux qui exploitent les corpus comparables pour extraire la relation de traduction, nous avons choisi la phrase comme unité opératoire. Ses frontières sont marquées par des indices typographiques de début et de fin. A l'intérieur de la phrase nous avons choisi la fenêtre graphique comme représentant du contexte d'un mot. A notre sens, une fenêtre de taille réduite prend en compte certaines relations syntaxiques du type 'verbe + actant' ainsi que la dépendance collocationnelle (section 2.3.1). Nous avançons l'hypothèse que la prise en compte de ces relations et le fait qu'elles se retrouvent associées aux contextes de deux mots de deux langues différentes est un signe de l'existence d'une relation de traduction entre ces mots. Autrement dit, notre modèle s'appuie donc sur l'hypothèse que la ressemblance des contextes peu distants de deux mots est un indice fort de leur relation de traduction. Dans l'expérience décrite au chapitre suivant nous faisons varier la taille de la fenêtre afin de déterminer expérimentalement dans quelle mesure elle a une influence sur la détermination de la relation de traduction.

3.2.2 Systèmes de pondération

Nous avons mentionné précédemment que l'approche générale adoptée pour l'extraction de lexique bilingue à partir de corpus comparables est de tout d'abord définir un ensemble de contextes dans lesquels la distribution de chaque mot dans chacune des langues est calculée ; de traduire ensuite chaque élément de chaque contexte dans une des deux langues à l'aide des ressources lexicales bilingues ; puis, de comparer les mots sur la base de leurs distributions.

Le calcul de la distribution contextuelle d'un mot fait appel à une pondération. En effet on ne peut se contenter de compter simplement les cooccurrences entre un mot et un mot de contexte. Des valeurs similaires de cooccurrences ne mettent pas toujours en évidence le même lien. En particulier pour deux mots de contexte ayant des fréquences très différentes dans un corpus (l'un de faible fréquence et l'autre de forte fréquence), des mesures de cooccurrences similaires entre un mot et ces deux mots de contextes ne sont pas comparables. Le nombre de cooccurrences observé pourra par exemple dans un cas être plus élevé que le nombre de cooccurrences estimé pour une distribution aléatoire, et dans l'autre cas moins élevé. Finalement deux nombres de cooccurrences similaires seront dans ce cas le reflet pour l'un d'une dépendance entre les deux mots et pour l'autre d'une absence de dépendance. Le simple dénombrement des cooccurrences ne suffisant pas, plusieurs systèmes de pondération ont été avancés reposant sur des mesures statistiques que nous examinons ici.

Information Mutuelle

L'information mutuelle s'inscrit dans le cadre de la théorie de l'information (Shannon, 1948). La quantité d'information apportée par un mot j sur la présence d'un autre mot i est l'information mutuelle $IM(i,j)$. Cette valeur s'exprime par le rapport de la probabilité d'observer i sachant que l'on a observé j sur la probabilité de i, soit :

$$IM(i,j) = \log \frac{P(i|j)}{P(i)}$$

Avec la formule de Bayes $P(i|j)$ s'écrit $P(i,j)/P(j)$ où $P(i,j)$ est la probabilité d'observer i et j simultanément. L'information mutuelle devient alors :

$$IM(i,j) = \log_2 \frac{P(i,j)}{P(i)P(j)}$$

La probabilité d'occurrence d'un mot i est calculée à partir de sa fréquence dans un corpus donné ($f(i)$), normalisée par le nombre total de mots dans le corpus soit N. La probabilité de cooccurrences de deux mots i et j est calculée à partir de leur cooccurrences dans le corpus $f(i,j)$, divisée par le nombre total de mots dans le corpus :

$$IM(i,j) = \log_2 N \times \frac{f(i,j)}{f(i) \times f(j)}$$

Cet indicateur est utilisé pour mesurer l'association entre des mots simples composant des collocations (Church & Hanks, 1990) ou des termes complexes de type nominal (Rapp, 1995) ainsi que pour les mettre en relation à partir de corpus comparables bilingues (Iram *et al.*, 1999).

Notons que l'information mutuelle est l'expression de liaisons récurrentes et exclusives de mots. Plus deux mots sont dans les mêmes contextes et rien que dans les mêmes contextes, plus l'information mutuelle a une valeur importante. Or beaucoup de mots cooccurrent avec des mots et ne sont pas liés de manière exclusive avec ces mots. L'information mutuelle a donc tendance à favoriser ces liaisons exclusives et notamment l'association entre les mots ayant de faibles occurrences.

Mesure du Chi 2

La statistique du χ^2 (Chi 2) mesure la dépendance entre deux mots à partir de la table de contingence du tableau 3.1 :

	j	$\neg j$	
i	a	c	$i_1 = a + c$
$\neg i$	b	d	$i_0 = b + d$
	$j_1 = a + b$	$j_0 = c + d$	$N = a + b + c + d$

TAB. 3.1 – *Table de contingence pour la dépendance de deux unités i et j.*

Dans cette table, a est le nombre de contextes dans lesquels i et j apparaissent tous les deux, b est le nombre de contextes où j est présent mais i est absent, etc. La mesure d'association χ^2 est alors définie comme suit :

$$\chi^2(i,j) = \frac{N(ad - cb)^2}{j_1 \ i_1 \ i_0 \ j_0}$$

La distance du χ^2 ainsi définie mesure le degré de dépendance des mots. Plus les deux mots i et j apparaissent ensemble dans les mêmes contextes, plus grande est la valeur ad. Plus ils sont absents tous les deux des mêmes contextes, plus la valeur cb est petite. Ainsi, deux mots indépendants ont un χ^2 nul. Quand le nombre des contexte est trop faible, l'emploi du Chi 2 rencontre ses limites (Muller, 1997).

Rapport de vraisemblance

D'après plusieurs auteurs (Dunning, 1993; Baayen, 2001; Habert & Jardino, 2003), les modèles basés sur l'hypothèse d'une distribution normale des occurrences ne sont pas adaptés à l'étude des événements rares. En effet, dans ces modèles, comme avec l'information mutuelle, la cooccurrence de deux hapax devient hautement improbable. Il propose alors d'associer une probabilité p constante d'obtenir un mot donné en choisissant une occurrence au hasard. L'apparition de k occurrences d'un

mot est alors considérée comme issue de tirages indépendants de probabilité p (http://helmer.hit.uib.no/corpora/1997-2/0148.html) :

$$\begin{aligned} loglike(i,j) &= \textstyle\sum_{ij} \log \frac{k_{ij}N}{C_i R_j} \\ &= k_{11} \log \tfrac{k_{11}N}{C_1 R_1} + k_{12} \log \tfrac{k_{12}N}{C_1 R_2} + k_{21} \log \tfrac{k_{21}N}{C_2 R_1} + k_{22} \log \tfrac{k_{22}N}{C_2 R_2} \end{aligned}$$

$$\begin{aligned} C_1 &= k_{11} + k_{12} \\ C_2 &= k_{21} + k_{22} \\ R_1 &= k_{11} + k_{21} \\ R_2 &= k_{12} + k_{22} \\ N &= k_{11} + k_{12} + k_{21} + k_{22} \end{aligned}$$

k_{11} correspond aux cooccurrences des deux mot i et j

k_{12} est la différence entre le nombre d'occurrences de i et k_{11}

k_{21} est la différence entre le nombre d'occurrences de j et k_{11}

k_{22} = nombre total d'occurrences dans le corpus - $k_{12} - k_{21} + k_{11}$

Cette mesure est couramment utilisée pour calculer la distribution des mots dans les corpus bilingues comparables (Fung & McKeown, 1997; Rapp, 1999; Déjean & Gaussier, 2002).

tf.idf

La valeur *tf.idf* (*term frequency* $\times$ *inverse document frequency*) est un indice classique utilisé en recherche d'information pour la pondération des termes. En général, $tf_{i,j}$ désigne la fréquence d'un terme i dans un document j, normalisée par la fréquence maximale des termes dans j :

$$tf_{i,j} = \tfrac{f_{i,j}}{\max_{l,j}}.$$

Pourtant, la fréquence ne peut pas être le seul critère utilisé, puisque deux mots peuvent avoir la même fréquence dans un document mais l'un peut apporter plus d'informations sur le document que l'autre s'il est plus rare dans le corpus. Pour exploiter ce phénomène, l'indice idf est pris en compte et met en évidence le pouvoir de discrimination d'un terme i :

$$idf_i = \log \tfrac{N}{n_i}.$$

N est le nombre total de documents dans le corpus et n_i est le nombre de documents dans lesquels se trouve le terme i.

Finalement le poids d'un terme i dans un document j est une combinaison de sa fréquence dans le document et de son pouvoir de discrimination dans le corpus :

$$p_{i,j} = tf_{i,j} \log \tfrac{N}{n_i}$$

Lorsque l'on applique ce système de pondération au cas des distributions de mots dans des contextes (Fung & Yee, 1998; Chiao & Zweigenbaum, 2002a), le document j est remplacé par le terme dont on calcule la distribution. Nous pouvons ainsi reformuler le $tf_{i,j}$ par la fréquence de cooccurrence d'un terme j dans le contexte d'un autre terme i (soit le nombre de cooccurrences de i et j : $cooc_{i,j}$), normalisée par la fréquence maximale de cooccurrence du terme j sur l'ensemble des contextes ($\max_{cooc_j}$) :

$$tf_{i,j} = \frac{cooc_{i,j}}{\max_{cooc_j}}$$

Le *idf* devient :

$$idf_i = 1 + \log \frac{\max_{cooc}}{cooc_i}$$

Où $cooc_i$ est le nombre total de mots pour lesquels i est un mot de contexte et $\max_{cooc}$ est le nombre maximal de mots de contexte dans le corpus. Pour un terme j, le poids d'un mot de contexte i est ainsi obtenu par :

$$p_{i,j} = tf_{i,j} \times idf_i$$

3.2.3 Modèle vectoriel et similarité entre vecteurs de contexte

Afin de représenter l'ensemble des contextes d'un mot, une idée est de considérer la matrice des associations mot/mot, en associant à chaque mot l'ensemble des mots de ses contextes. Cette idée introduit la notion de vecteur de contexte généralisant la notion de contexte simple constitué de cooccurrences. Le modèle vectoriel développé par Salton (Salton *et al.*, 1974) dans le cadre de la recherche d'information, que nous présentons dans le chapitre 5, peut ainsi être exploité. Pour un corpus donné, il s'agit de calculer la matrice des cooccurrences entre les mots (matrice carrée et symétrique). Un mot est donc considéré comme un vecteur dans l'espace des mots. La coordonnée d'un mot sur une dimension est la force d'association entre ce mot et le mot associé à la dimension en question (information mutuelle, $tf.idf$...). Notons que cette matrice est creuse (remplie essentiellement de zéros).

La relation entre deux mots est alors calculée par une mesure de similarité (cosinus, Jaccard) ou par une mesure de distance (distance de Manhattan, distance euclidienne) entre deux vecteurs de l'espace des mots.

Un des problèmes de l'approche vectorielle est sa complexité élevée. En particulier, la comparaison des vecteurs entre eux nécessite la comparaison des mots deux à deux. La complexité est égale au nombre total des mots du corpus au carré (en $O(n^2)$). Sur des textes de taille importante le temps et la mémoire nécessaires deviennent rapidement rédhibitoires. Il est donc raisonnable de limiter la taille de la matrice pour réduire l'espace de recherche. La réduction du nombre de dimensions (qui est égal au nombre de mots dans le corpus) est possible en éliminant les mots vides[1] ou grammaticaux qui, même s'ils ne sont pas nombreux cooccurrent souvent avec les autres mots.

1. Ces mots ont des catégories grammaticales autres que le nom, l'adjectif et le verbe.

Les dictionnaires bilingues peuvent aussi être utilisés pour sélectionner uniquement les mots des contextes à comparer. Nous décrivons dans le chapitre 4, les procédés de réduction de l'espace vectoriel que nous avons mis en œuvre.

Une fois les contextes des mots d'une langue construits et pondérés, ils sont traduits à l'aide des ressources bilingues disponibles afin de permettre la comparaison avec les contextes des mots de l'autre langue. Cette comparaison est effectuée classiquement par des mesures de similarité ou des distances présentées ici. Pour comparer deux vecteurs v, w de longueur n, deux mesures de similarité sont très utilisées :

Cosinus

$$\cos(v,w) = \frac{\sum_{k=1}^{k=n} v_k w_k}{\sqrt{\sum_{k=1}^{k=n} v_k^2} \times \sqrt{\sum_{k=1}^{k=n} w_l^2}}$$

Coefficient de Jaccard [2]

$$\textit{Jaccard}(v,w) = \frac{\sum_{k=1}^{k=n} v_k w_k}{\sum_{k=1}^{k=n} v_k^2 + \sum_{k=1}^{k=n} w_l^2 - \sum_{k=1}^{k=n} v_m w_m}$$

Distance de Minkowski Au contraire des mesures de similarité, les distances accordent une valeur maximale à deux objets complètement différents et minimale (0) à deux objets identiques. Les mesures les plus utilisées sont la distance euclidienne ou celle de Manhattan, qui ne sont en fait que des cas particuliers de la mesure de Minkowski :

$$D_p(v,w) = \Big(\sum_{k=1}^{k=n} |v_k - w_k|^p \Big)^{\frac{1}{p}}$$

Où $p = 1$ donne la distance de Manhattan et $p = 2$ la distance euclidienne.

3.2.4 Similarité interlangue

L'hypothèse sur laquelle sont fondées les méthodes exposées jusqu'ici est que la ressemblance des distributions contextuelles de deux mots est un signe qu'une relation de traduction les lie. Une hypothèse supplémentaire peut être avancée dans un cadre bilingue. Si un mot de la langue a est proche de mots au sens d'une similarité de leurs distributions dans le corpus de langue a, qu'un mot de la langue b est proche de ces mêmes mots (à la traduction près par une ressources bilingue pré-existante) au sens d'une similarité de leurs distributions dans le corpus de langue b, alors ces deux mots ont de fortes chances d'être traductions l'un de l'autre. Par exemple, si *médecin*

2. Le coefficient de Jaccard est aussi connu sous le nom de coefficient de Tanimoto. Dans cette mesure, le poids des faibles cooccurrences est neutralisé, contrairement au coefficient de Dice présenté dans la section 2.4.2 qui attribue un double poids aux cooccurrences.

a une distribution similaire à celle de *infirmière* de la même façon que *doctor* a une distribution proche de celle de *nurse* et que *infirmière* est la traduction de *nurse* alors *médecin* a de grandes chances d'être la traduction de *doctor*.

Cette hypothèse a été mise en œuvre et expérimentée dans (Déjean & Gaussier, 2002). Nous avons vu dans la section 2.3.3 que pour une même langue et un même corpus la similarité des distributions entre les mots peut être exploitée pour le regroupement ou le rapprochement de ceux-ci. Le modèle vectoriel et la comparaison des distributions à l'aide d'une mesure de similarité sont utilisés dans ce modèle dans un cadre monolingue pour chacun des corpus. Ainsi, la première étape est la même que celle de l'approche classique et consiste à définir et à constituer les contextes de co-occurrences pour calculer les proximités sémantiques au sein d'une même langue. Il s'agit de rapprocher pour un même corpus d'une langue donnée les mots qui partagent les mêmes contextes de cooccurrence.

Les similarités résultantes dans les deux langues sont ensuite utilisées pour établir que deux mots sont traductions l'un de l'autre. Ici aussi les ressources lexicales bilingues participent au rapprochement. Les similarités calculées entre d'une part les deux mots de langues différentes et d'autre part les entrées de la ressource bilingue servent à estimer la probabilité que ces deux mots sont traductions l'un de l'autre.

Ce modèle nous semble mettre en œuvre une forme de la symétrie distributionnelle évoquée dans le chapitre 1. En effet l'idée développée dans ce modèle est que si deux mots de langues différentes ont les mêmes (en passant par une ressource bilingue) mots proches (au sens d'une mesure de similarité dans leur corpus d'origine) alors ils ont toutes les chances d'être traductions l'un de l'autre. Alors que nous avons défini la symétrie distributionnelle sur les proximités des distributions d'un corpus vers l'autre corpus, elle est ici mise en œuvre à partir des proximités des distributions dans chacun des corpus. Plus précisément, dans la définition de la symétrie distributionnelle que nous avons donnée, les similarités entre distributions sont calculées entre les mots des corpus de langues différentes puis la symétrie est prise en compte entre ces similarités. Ici, les similarités sont calculées entre mots d'un même corpus et la symétrie est prise en compte entre les similarités d'un corpus d'une langue et celles du corpus de l'autre langue.

3.2.5 Autres indicateurs

Plusieurs indicateurs ont été utilisés en appoint de l'approche distributionnelle pour en améliorer l'efficacité. Deux d'entre elles sont présentées ici. L'une est fondée sur la ressemblance graphique des mots de deux langues différentes, l'autre sur l'équivalence de leur fréquence.

Ressemblance graphique

Beaucoup de méthodes d'alignement de corpus parallèles exploitent la ressemblance graphique pour effectuer un pré-alignement grossier qui permet de repérer les segments de textes sur lesquels sont appliqués des algorithmes d'alignement plus précis. Deux types d'indices formels peuvent être distingués : la présence de *transfuges* et celle de *cognats*.

Les transfuges sont des chaînes de caractères identiques dans les deux langues par exemple, les chiffres, les noms propres comme De Gaulle, Churchill, ou les noms communs lorsqu'il s'agit du phénomène de l'emprunt : parking, Internet, etc. Dans le même ordre d'idée, les mots graphiquement apparentés, *i.e.*, *cognats* tels que *terminologie* et *terminology* peuvent aussi servir de points de repère pour l'alignement. Contrairement aux transfuges qui sont identiques du point de vue graphique et sémantique, la notion de cognat est plus vague. Ceci est dû aux différentes définitions possibles de la ressemblance. Il existe de plus des "faux amis", des mots qui se ressemblent graphiquement mais qui diffèrent sur le plan sémantique, *e.g.*, *habit* en français et *habit* en anglais, ou *caution* et *caution*, etc.

A part l'utilisation de dictionnaires de cognats, la plupart des méthodes d'identification automatique de cognats développées dans le cadre de l'alignement des corpus parallèles reposent sur la comparaison des longueurs des chaînes de caractères des mots, *i.e.*, la prise en compte de *n-grammes*. Le modèle *char_align* de Church (Church, 1993) considère comme cognats les mots ayant quatre caractères consécutifs en commun. Un filtrage des candidats s'effectue à l'aide de l'analyse de leur fréquence. Des couples de mots sont retenus comme cognats si leur fréquence ne dépasse pas un certain seuil prédéfini. La mesure du *rapport de la plus longue sous-chaîne commune*[3], une autre méthode plus élaborée, propose d'améliorer les résultats des méthodes basées sur les *n-grammes*, en calculant la similarité entre deux mots sur la base de la longueur maximale des sous-chaînes partagées par rapport à la taille des mots (Melamed, 1995; Davis *et al.*, 1995; Mann & Yarowsky, 2001; Kraif, 2001b).

L'utilisation des ressemblances graphiques trouve néanmoins ses limites pour les couples de langues non apparentées, comme le français-chinois par exemple. Dans le cas des langues apparentées, la majorité des mots qui sont traductions mutuelles ne rentrent pas dans la catégorie des transfuges ou des cognats mais les études récentes ont tout de même montré l'intérêt de combiner ces indices avec l'analyse distributionnelle afin d'améliorer les résultats de l'alignement des textes parallèles.

Dans la cadre de l'acquisition lexicale à partir de corpus comparables, nous accordons moins d'importance aux cognats pour une raison principale. La nature d'un corpus parallèle est différente de celle d'un corpus comparable. La traduction d'un mot dans les textes comparables ne peut pas être repérée par sa position textuelle comme c'est la cas dans des corpus parallèles.

Il est à noter que beaucoup de cognats sont déjà inclus dans les ressources bi-

3. *Longest common subsequence ratio*.

lingues que nous exploitons pour nos expériences (tableau 3.2). Dans le domaine de la médecine, beaucoup d'entrées sont des noms propres qui désignent des maladies, des médicaments, etc. (Bodenreider & Zweigenbaum, 2000). Comme nous l'avons déjà mentionné, les noms propres ou les chiffres sont des catégories dans lesquelles se trouvent en général plus de cognats ou de transfuges que dans d'autres catégories.

abarognosie	abarognosis
abarthrose	abarthrosis
abdomino-génital	abdominogenital
abcès	abscess
biologie	biology
bronchectasie	bronchiectasis
cancer	cancer
chlorurémie	chloruremia
iléo-caecostomie	ileocecostomy
pathologique	pathological
thrombo-embolie	thromboembolia
vitamine	vitamin
urine	urine

TAB. 3.2 – *Exemples de cognats dans le lexique médical*[4] *de mots simples.*

Distribution des fréquences des mots

Un autre indicateur peut être utilisé en appoint de l'approche distributionnelle. Il s'agit de la fréquence d'un mot.

La distribution des fréquences des mots dans un texte ou un corpus n'est pas uniforme. En général, les mots grammaticaux, *i.e.*, déterminants (le, la...), prépositions (de, à...) ont une fréquence plus élevée. Les mots les moins fréquents sont les hapax[5], qui sont souvent ici, du fait de la constitution de nos corpus, des fautes typographiques ou orthographiques.

La comparaison des distributions des fréquences des mots peut être intéressante pour l'exploitation des corpus parallèles et comparables. L'hypothèse sous-jacente est que les mots qui sont traductions réciproques sont distribués d'une façon similaire, *i.e.*, ont une fréquence similaire dans les corpus. Le problème est de trouver un modèle permettant de rapprocher les différentes distributions constatées en prenant en compte l'échelonnement des fréquences à l'intérieur d'une distribution. L'étude sur la distribution des fréquences la plus ancienne et la plus connue est celle proposée par Zipf (Zipf, 1949), connue sous le nom de *loi de Zipf*. Il a remarqué qu'en classant les mots d'un

4. Il s'agit du lexique médical décrit et utilisé dans le chapitre 4.

5. Les mots qui n'apparaissent qu'une seule fois.

Rang	Corpus FR	Corpus EN
1	*santé*	*patient*
2	*cas*	al
3	*patient*, *traitement*, personne	*disease*
4	risque, *maladie*, service	gene, cell
5	enfant, recherche	*health*, pubmed, id
6	*étude*, information, ans, soin, fait	home, mutation
7	médecin, *site*, jour,	provider, *page*, topic
8	travail, effet, rapport, *clinique*, donnée, centre, peuvent, groupe	medline, related, *case*
9	nombre, faire, produit, mise, *page*, *cancer*, prise, année, association, cour	hospital, textbook, *clinical*, *site*, syndrome, human
10	activité, niveau, programme, femme, article, type, mesure, france, résultat, vie, infection	*treatment*, *cancer*, *study*, neighbor, family

TAB. 3.3 – *Distribution des fréquences de mots dans les corpus*[6] *français et anglais ; exemple des 10 premiers rangs correspondant aux mots les plus fréquents dans les corpus ; les mots notés en italique ont également leur traduction dans les 10 premiers rangs.*

texte par ordre décroissant de fréquences f et en leur attribuant un rang r, le produit $f.r$ est une constante qui dépend uniquement du texte. Pour un mot i dans le corpus, cette relation s'écrit alors :

$$C = f_i.r_i$$

Nous remarquons que cette loi reflète bien le comportement général de la distribution des fréquences des mots d'un corpus : il existe un petit nombre de mots très fréquents et un grand nombre de mots très rares n'apparaissant qu'une fois ou deux ainsi qu'un ensemble de mots dont la fréquence se situe entre ces deux zones. Lorsqu'on observe de près les rangs des fréquences des mots dans un corpus bilingue, on s'aperçoit que parmi les mots les plus fréquents d'un des deux corpus monolingues, la traduction de beaucoup de mots se trouve classée parmi les mots les plus fréquents de l'autre corpus monolingue (les mots en italique dans le tableau 3.3). De la même façon, la traduction d'u mot rare possède elle-même une fréquence faible.

Comme pour les cognats, la mise en correspondance des distributions de fréquences n'est pas suffisante pour fournir des résultats pertinents. Par contre, il y a une forte corrélation entre la fréquence d'un mot et celle de sa traduction, notamment dans des corpus comparables, *e.g.*, la traduction d'un mot rare possède généralement elle-même une fréquence faible.

6. Il s'agit des corpus utilisés dans les expériences décrites au chapitre 4.

3.3 Un nouveau modèle d'extraction de lexique bilingue à partir de corpus comparables

3.3.1 Mise en œuvre de la symétrie distributionnelle

Nous avons passé en revue différents paramètres et mesures exploités pour l'extraction de lexique bilingue à partir de corpus comparables. De façon générale, la ressemblance des contextes de cooccurrences est la base de la détection de la relation de traduction entre les mots de différentes langues.

Nous proposons dans cette section un modèle qui repose principalement sur l'analyse de cooccurrence en exploitant des ressources bilingues de différentes natures : publications en ligne, dictionnaires de langues spécialisée et générale, etc. Ce modèle a été appliqué à l'acquisition de lexique bilingue (chapitre 5) ainsi qu'à la recherche d'information translangue (chapitre 6).

Jusqu'à présent, les méthodes d'alignement à partir de corpus comparables sont orientées en partant d'une langue A vers une autre langue B (Fung & Yee, 1998; Rapp, 1999; Déjean & Gaussier, 2002; Chiao & Zweigenbaum, 2002b). Pour un mot de la langue A, elles proposent comme candidats les mots de la langue B dont la distribution lui est la plus similaire. Or, comme évoqué dans le chapitre 1 on peut adjoindre à l'hypothèse distributionnelle sur laquelle reposent ces méthodes, l'hypothèse de symétrie distributionnelle. Nous proposons ainsi une nouvelle approche exploitant la symétrie distributionnelle dont nous rappelons la formulation ici : si deux mots sont proches dans une direction de traduction ainsi que dans l'autre alors ils ont de plus fortes chances d'être traductions l'un de l'autre que s'ils ne sont proches que pour une seule direction de traduction.

Nous proposons donc un modèle d'alignement symétrique, dans les deux directions de langues (langue $A \leftrightarrow$ langue B). Plus concrètement, il s'agit de procéder à une mise en correspondance croisée dans les deux directions et de mettre en évidence les mots les plus proches des deux côtés simultanément. Cela nous a amené à construire une 'similarité croisée' explicitée plus loin rendant compte de cette idée.

Comme l'illustre la figure 3.1, la mise en œuvre de notre modèle se décompose en quatre étapes :

1. Prétraitement du corpus (section 3.3.2).

2. Construction des vecteurs de contexte (section 3.3.3).

3. Transfert des vecteurs de contexte (section 3.3.4).

4. Calcul de la similarité croisée (section 3.3.5).

Nous détaillons dans les sections suivantes chaque étape du modèle proposé.

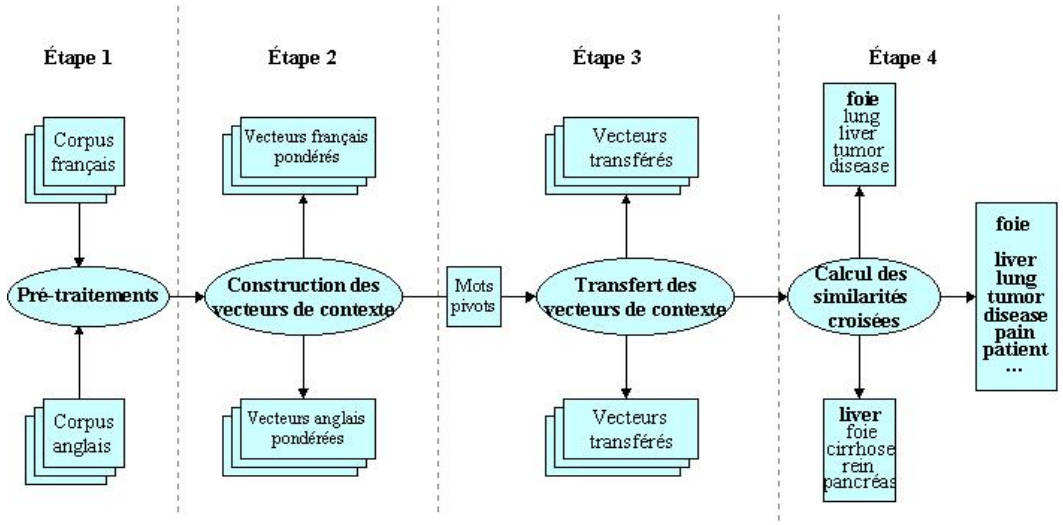


FIG. 3.1 – *Mise en œuvre de l'approche symétrique d'extraction de lexique bilingue.*

3.3.2 Prétraitement du corpus

Le prétraitement du corpus permet de formater les données textuelles et de les rendre directement exploitables pour les traitements ultérieurs : examen des cooccurrences pour construire les vecteurs, alignement, etc. Le prétraitement consiste essentiellement à repérer les mots et les contextes associés dans les corpus comparables. Nous avons décidé de travailler sur des textes bruts sans étiquetage pour plusieurs raisons. La première raison est d'assurer la portabilité de notre modèle sur de nouvelles paires de langues. Or l'étiquetage de corpus est une opération dépendante de la langue du corpus traité car il nécessite l'utilisation de connaissances linguistiques. Par ailleurs, l'étiquetage utile pour le problème de l'acquisition lexicale est essentiellement un étiquetage sémantique qui par exemple désambiguïse les mots polysémiques. Ce type d'étiquetage reste encore une opération qui demande à être validée manuellement car elle génère un bruit non négligeable. La possibilité de travailler sur des corpus analysés reste néanmoins une voie d'amélioration de notre modèle que nous évoquons dans les perspectives au dernier chapitre (7.3.4).

Découpage en phrases

Il s'agit de découper le texte en une suite de phrases en utilisant les marqueurs de ponctuations fortes : "!", "." et "?". Nous laissons de côté le cas particulier des abréviations avec la présence du point : *J.F.K.*, *F.N.O.R.S* (Fédération Nationale des Observatoires de la Santé), etc. en le considérant comme un problème lié à l'identification des noms propres, donc du ressort de l'acquisition de terminologie monolingue. Le découpage en phrases a pour but de constituer une liste des phrases du corpus permettant ensuite de fixer les frontières des fenêtres de mots. La phrase, marquant les frontière des contextes de cooccurrences, permet de conserver des relations syntaxiques entre les mots.

Prenons par exemple l'extrait d'un texte dans la figure 3.2. Avec le découpage

> ...Les rotavirus sont la première cause de *gastro-entérites* chez les enfants de moins de 5 ans. Les adultes peuvent également être infectés notamment lorsque leurs enfants présentent une diarrhée à rotavirus ou lorsqu'ils sont immunodéficients. ...

FIG. 3.2 – *Extrait de texte médical.*

en phrases, pour le mot *gastro-entérites*, les mots pleins qui constituent son contexte après segmentation en fenêtre de sept mots comme décrit ci-dessous sont *rotavirus, première, cause, enfants, ans*. Sans fixer la frontière en phrases, les cooccurrents pris en considération seraient *rotavirus, première, cause, enfants, ans, adultes*.

Repérage des mots

Le repérage des mots s'effectue par le découpage des textes en une suite de mots à l'aide des caractères séparateurs de mots (présence d'un blanc, d'une tabulation, d'un point, etc.). Le trait d'union est considéré comme partie intégrante d'un mot. Cela permet d'identifier une partie des mots complexes du domaine médical : *brachio-céphalique, beta-alanine, amino-acides, adéno-cellulite, bec-de-lièvre*, etc. Ces mots représentent environ 18% des entrées du lexique spécialisé que nous décrivons dans le chapitre 4. Les dates et nombres ne sont pas pris en compte dans le repérage des mots car nous estimons qu'ils rentrent dans la catégorie des transfuges (section 3.2.5).

Segmentation en fenêtres de mots

La segmentation en fenêtres de mots est une étape importante pour le calcul des cooccurrences. Un mot i est considéré comme cooccurrent d'un autre mot j s'il apparaît à moins de k mots de distance dans l'ordre linéaire du texte. La segmentation en fenêtres de mots permet d'associer ainsi, à chaque mot retenu par le repérage des

mots, ceux qui apparaissent dans ses contextes (cooccurrents) et le nombre de leurs apparitions dans ses contextes (cooccurrences).

Nous utilisons deux listes de mots vides, une pour le corpus français et une pour l'anglais (annexes A.3 et A.4). Ils servent à filtrer les mots qui sont estimés peu discriminants pour être inclus dans un contexte, et moins intéressants pour l'extraction de lexique bilingue. Nous appliquons aussi deux algorithmes pour la mise au singulier[7] de tous les mots repérés dans les deux corpus français/anglais respectivement. Le tableau 3.4 donne un exemple de la segmentation en fenêtre des mots.

Phrase	**7 mots**	**5 mots**
quantitative	quantitative	
or	—	
qualitative	qualitative	qualitative
deficiency	deficiency	deficiency
of	—	—
sialophorin	**sialophorin**	**sialophorin**
in	—	—
some	—	—
way	—	—
due	—	—
to	—	—
abnormal	abnormal	abnormal
Wiskott-Aldrich	wiskott-aldrich	wiskott-aldrich
Syndrome	syndrome	

TAB. 3.4 – *Exemple des contextes de fenêtre de n mots pour le mot **sialophorin** dans le corpus anglais ; ($n = \{5, 7\}$)*

3.3.3 Construction des vecteurs de contexte

La seconde étape de notre approche vise à établir une représentation des mots du corpus par des vecteurs de contexte. Elle consiste à construire, pour chaque mot m dans les deux corpus respectivement, un vecteur de contexte composé de tous les mots pleins qui cooccurrent avec m dans une fenêtre de n mots au sein de la même phrase p (section 3.3.2). On attribue ensuite un poids à chaque mot de contexte i associé à m. Parmi les différents systèmes de pondération présentés dans la section 3.2.2, trois sont évaluées dans le chapitre 4 : *tf.idf*, *IM*, et *loglike* (rapport de vraisemblance).

L'utilité de la pondération est de ne pas donner la même importance aux mots du contexte qui décrivent un mot. Par exemple, un mot du contexte qui cooccurre souvent doit avoir un poid plus fort qu'un mot qui cooccurre moins souvent. Les différentes mesures de pondération mettent en évidence différents phénomènes tels que l'effet discriminant d'un terme, son importance dans le corpus, etc.

7. Pour la langue anglaise, nous avons adapté l'algorithme S-stemmer (Harman, 1991) ; et pour la langue française, nous avons construit un simple algorithme pour traiter certains suffixes : *-s, -x*.

A la fin de cette étape, un mot plein m du corpus est alors représenté par un vecteur de contextes pondérés (w_1, w_2..., w_j), appelé le vecteur de contexte du mot. Le tableau 3.5 montre un exemple de vecteur pondéré dans lequel les mots sont classés en ordre décroissant de poids.

Mot	**Occ.**	**Co.**	**10 premiers éléments dans le vecteur de contexte pondéré par l'information mutuelle**
adénose	48	45	nucléole (17,4), hyperplasie (14,2), photo (13,9), prolifération (11,9), prostatique (11,9), adénome (11,8), prostate (9,1), cellule (8,9), lésion (8,8), test (5,9)

TAB. 3.5 – *Exemple de vecteur de contexte pondéré pour* adénose *; Occ. = nombre d'occurrences dans le corpus ; Co = nombre de cooccurrents.*

3.3.4 Tranfert des vecteurs de contexte

La troisième étape a pour but de transférer les vecteurs de contexte afin d'obtenir des vecteurs normalisés donc comparables d'une langue à l'autre.

Pour les vecteurs de la langue source, le transfert consiste à remplaçer les mots des vecteurs par leur traduction en langue cible. Cette opération de traduction se fait à l'aide de ressources bilingues partielles pré-existantes, que nous décrivons dans le chapitre 4. L'utilisation d'un lexique préexistant est primordial dans notre modèle car elle rend possible la mise en correspondance des vecteurs de langues différentes. Le transfert des vecteurs par la traduction permet alors d'effectuer ensuite la comparaison des vecteurs par une mesure de similarité dans la phase finale.

Selon le modèle vectoriel classique en recherche d'information, l'espace vectoriel est défini par l'ensemble des termes d'indexation. Dans le cas de l'acquisition lexicale bilingue, nous utilisons un ensemble de *mots pivots* pour définir l'espace vectoriel. Les mots pivots sont d'après le lexique bilingue initial des couples de mots qui sont des traductions mutuelles. Les critères utilisés pour établir la liste des mots pivots sont les suivants. Un mot fait partie des mots pivots si et seulement si :

- il apparaît dans le corpus de langue source ;
- il correspond à une entrée du lexique bilingue utilisé pour la traduction des vecteurs ;
- un de ses équivalents en langue cible apparaît dans le corpus cible.

Le vecteur de contexte illustré dans le tableau 3.5 après la traduction par des mots pivots est donné dans le tableau 3.6.

Un corpus de N mots est alors représenté par une matrice X de dimension $N \times \mid P \mid$, dans laquelle la ligne k correspond au vecteur de contexte du $k^{éme}$ mot et $\mid P \mid$ est le cardinal de l'ensemble des mots pivots.

$$\mathbf{X} = \left\{ \begin{array}{c} m_1 \\ m_2 \\ \vdots \\ m_N \end{array} \right\} = \left\{ \begin{array}{cccc} w_{11} & w_{12} & \dots & w_{1|P|} \\ w_{21} & w_{22} & \dots & w_{2|P|} \\ \vdots & \vdots & \ddots & \vdots \\ w_{N1} & w_{N2} & \dots & w_{N|P|} \end{array} \right\}$$

Mot	**Occ.**	**Co.**	**vecteur traduit avec la pondération information mutuelle**
adénose	48	45	hyperplasia (14,2), photograph (13,9), lesion (8,8), cell (8,9), proliferation (11,9), adenoma (11,8), test (5,9), prostatic (11,9), prostate (9,1), nucleolus (17,4)

TAB. 3.6 – *Exemple de vecteur traduit pour le mot **adénose** ; Occ. = occurrences and Co. = cooccurrents dans le corpus ; les mots dans le vecteur sont suivis par l'information mutuelle avec **adénose**.*

Le transfert ne consiste pas uniquement à traduire les mots de contexte de langue source. Les mots ayant la même traduction (dans une direction ou l'autre de langues) sont remplacés par un mot unique qui est en réalité la première entrée rencontrée. Ainsi les synonymes en langue source sont traduits par un même mot de langue cible. Les synonymes en langue cible sont remplacés aussi par un même mot. Le transfert s'applique donc aussi aux mots de contexte de langue cible. Le traitement des synonymes se justifie car on peut estimer qu'il s'agit de contextes équivalents. Le transfert traite aussi les mots polysémiques qui possèdent plusieurs traductions. Pour le traitement des mots polysémiques, plusieurs options sont possibles sans qu'aucune ne soit complétement satisfaisante. L'une consiste à étendre le vecteur de contexte avec l'ensemble des traductions du mot polysémique. Cela présente le désavantage de produire des vecteurs bruités. L'autre option consiste à choisir la plus fréquente des traductions soit la première rencontrée dans le lexique. C'est l'option retenue ici.

3.3.5 Calcul de la similarité croisée

Le calcul de la similarité croisée s'effectue en deux phases. La première consiste à calculer la similarité classique et à classer les candidats en fonction de leur score dans les deux directions de langues parallèlement (langue source $\rightleftarrows$ langue cible). On obtient alors deux classements : un pour chaque direction de langues. La seconde consiste à calculer la similarité croisée et à reclasser les candidats à partir des listes précédemment classées dont on ne retient que les n premiers candidats[8] .

8. Dans nos expériences, le reclassement s'effectue à partir des listes des 30 premiers candidats (section 4.3).

Classement des candidats selon la similarité classique

Pour chaque mot en langue source, son vecteur de contexte traduit issu du transfert (section 3.3.4) est comparé avec les vecteurs des mots en langue cible sur la base d'une mesure de similarité décrite dans la section 3.2.3. Les trois mesures présentées sont évaluées dans le chapitre 4 : cosinus, *Jaccard* et *Manhattan*[9] . Les scores de similarité ainsi obtenus pour chaque couple de mots (mot A_1 source/mot $B_{1...n}$ cible, où n est le nombre de mots différents dans le corpus en langue cible) servent ensuite à trier les couples des mots afin de constituer une liste ordonnée des candidats à la traduction du mot source A_1. Le tableau 3.7 présente un extrait des classements selon les valeurs de la similarité classique.

foie	lung	.270294
foie	liver	.231073
foie	pain	.174125
foie	patient	.162746
foie	tumor	.137852
foie	disease	.136998
foie	primary	.119938
foie	treatment	.119257
foie	brain	.109586
foie	cancer	.105038
foie	bone	.104870
foie	kidney	.104498

TAB. 3.7 – *Extrait de la liste des candidats ordonnés en fonction de leurs similarités décroissantes pour le mot français* ***foie*** *(direction : français → anglais).*

Les méthodes classiques d'extraction de lexique bilingue à partir de corpus comparables se contentent de proposer cette liste des candidats triée par leur valeur de similarité calculée de la langue source vers la langue cible. Nous pouvons remarquer dans cet exemple que pour le mot *foie*, la traduction attendue *liver* se trouve en 2ème position derrière *lung*.

Afin de pouvoir calculer la similarité croisée, nous avons besoin d'établir la liste des candidats à la traduction en langue source pour chaque mot de la langue cible. Les étapes précédentes sont donc appliquées sur le vecteur de chaque mot en langue cible dans la direction inverse (langue cible → langue source). Pour chaque mot en langue cible, on obtient ainsi une liste des mots en langue source ordonnée par similarité décroissante (Tableau 3.8).

Nous avons maintenant à notre disposition deux listes de couples de mots liés par une similarité (une liste par direction de langues).

9. La distance de *Manhattan* a été évaluée mais les résultats très peu satisfaisants n'ont pas été rapportés ici.

liver	foie	.365169
liver	rare	.309686
liver	associée	.292330
liver	alzheimer	.284989
liver	transmissible	.269096
liver	fréquente	.263598
liver	pathologie	.257709
liver	cardiovasculaire	.250468
liver	cardio-vasculaire	.248039
liver	creutzfeldt-jakob	.243688
liver	hépatique	.242475
liver	origine	.240563

TAB. 3.8 – *Extrait de la liste des candidats ordonnée en fonction de leurs similarités décroissantes pour le mot anglais* ***liver*** *(direction : anglais → français).*

Reclassement des candidats selon la similarité croisée

Une façon d'accorder autant d'importance aux deux corpus de langues différentes est, au lieu de considérer les valeurs des similarités, de prendre en compte les rangs associés. Ainsi à chaque couple de mots des deux listes, l'algorithme de reclassement associe d'abord sa position dans la liste ordonnée de mots proches du premier mot du couple. La similarité croisée de deux mots est alors calculée à partir du rang calculé dans une direction de traduction et du rang calculé dans l'autre direction de traduction.

Une des façons de combiner les rangs des deux listes est la *moyenne harmonique*[10] . Pour un couple de mots en langue source et cible (S, C), la moyenne harmonique de leurs rangs notée MH est :

$$MH(r_{sc}, r_{cs}) = \frac{1}{\frac{1}{2}(\frac{1}{r_{sc}} + \frac{1}{r_{cs}})} = \frac{2r_{sc}r_{cs}}{r_{sc} + r_{cs}}$$

Où r_{sc} est le rang calculé dans la direction (langue source → langue cible) et r_{cs} est le rang calculé dans l'autre direction (langue cible → langue source). L'intérêt de cette mesure est qu'elle favorise le fait d'être très bien classé au moins dans une direction de langues.

Reprenons l'exemple du mot *foie* dans le tableau 3.7. Les trois premiers couples de traductions classés selon la similarité classique calculée dans la direction (source → cible) sont *foie, lung* ($r_{sc} = 1$), *foie, liver* ($r_{sc} = 2$) et *foie, pain* ($r_{sc} = 3$). Le classement dans l'autre direction pour ces trois couples (cible → source) est *lung, foie* ($r_{cs} = 4$), *liver, foie* ($r_{cs} = 1$) et *pain, foie* ($r_{cs} = 31$). Leur nouveau classement selon

10. http://mathworld.wolfram.com/HarmonicMean.html

la similarité croisée est alors : *foie, liver* ($MH = 1.33$), *foie, lung* ($MH = 1.6$) et *foie, pain* ($MH = 5.48$). Les nouveaux rangs sont donnés dans le tableau 3.9.

Candidats	$\mathbf{r_{sc}}$	$\mathbf{r_{cs}}$	**MH**	**Nouveau rang**
lung	1	4	1,60	2
liver	2	1	1,33	1
pain	3	31	5,48	4

TAB. 3.9 – *Liste des mots candidats pour le mot* $foie$ *classés aux 3 premiers rangs dans les deux directions de langues et reclassés par la moyenne harmonique.*

3.3.6 Complexité des calculs

Nous abordons ici le calcul de la complexité en temps et en espace des algorithmes développés. Il s'agit de calculer d'un point de vue théorique les ressources nécessaires pour leur fonctionnement. Pour cela, pour chaque algorithme est calculée la complexité dans le pire des cas. Elle est une majoration de la complexité effective (notation O). Les différents processus mis en œuvre sont traités. Pour chacun des algorithmes, n représente le nombre d'entrées c'est-à-dire le nombre d'occurrences des mots dans les corpus. Un autre paramètre important est la taille de la fenêtre de contexte f. Même si dans notre modèle la taille de la fenêtre est réduite donc f est faible (5 ou 7 mots), d'autres travaux utilisent des fenêtres plus conséquentes de plusieurs dizaines de mots. La valeur de la taille de la fenêtre n'est donc pas anodine et constitue à nos yeux un paramètre à considérer dans le calcul de la complexité. Pour l'implémentation, notons que la plupart des traitements utilisent des tables de hachage dont la complexité en temps est constante tant pour la lecture d'un élément que pour son écriture.

Segmentation en mots

Le cas de la segmentation est simple. Sa complexité est en $O(n)$ car à chaque occurrence d'un mot en entrée correspond un nombre fini de traitements (repérage des frontières, passage au mot suivant...) qui ne dépendent pas de n. La complexité en espace est aussi en $O(n)$.

Construction des vecteurs pondérés de contexte

Le calcul des vecteurs de contexte prend en entrée les corpus segmentés. La complexité en temps est donc au moins en n. Pour chaque mot distinct provenant des n occurrences est construit un vecteur de contexte associé. Nous considérons trois paramètres, n le nombre total d'occurrences du corpus, p le nombre total de mots distincts et q le nombre maximum de mots de contexte. Pour calculer la complexité, il faut déterminer dans quelle mesure ces trois paramètres sont fonctions l'un de l'autre.

Il y a au plus et au pire n mots distincts pour n occurrences de mots. Pour chaque mot distinct, le vecteur de contexte contient au plus et au pire $(n-1)$ éléments. Pour chaque occurrence d'un mot, l'algorithme scrute son contexte proche dans une fenêtre de f mots. Pour chaque mot distinct, on alimente son vecteur de contexte donc au pire une matrice (n, $n-1$). La manipulation de cette matrice est en temps constant, mais chaque élément de cette matrice constitue un résultat du traitement et est donc écrit en sortie. On a donc une complexité en temps en $n.(n-1)$ soit $O(n^2)$. Pour la complexité en espace, il faut considérer la matrice utilisée précédemment. Elle est donc en $O(n^2)$.

Cette estimation est en réalité trop peu contraignante. En effet, deux grandeurs sont surestimées : le nombre maximum de mots distincts et le nombre maximum de mots des contextes. Considérons le cas limite ou chaque occurrence en entrée correspond à un mot distinct, le nombre de mots de contexte scruté est alors pour chaque occurrence, limité par la taille de la fenêtre utilisée f moins un, soit $(f-1)$ opérations par occurrence. Dans ce cas, la complexité en temps est $(f-1).n$ soit $O(f.n)$. L'autre cas limite est celui où chaque occurrence du corpus correspond au même et unique mot. Dans ce cas la complexité en temps est n soit $O(n)$.

Ces deux extrêmes ne sont pas satisfaisants et ne rendent pas compte de la complexité du cas général. Pour cerner au mieux la complexité nous nous appuyons sur la loi de Heaps (Heaps, 1978) qui postule que le nombre de mots distincts p est une fonction du nombre d'occurrences n de la forme $p = K.n^a$, où K est une constante et a une constante inférieure à 1. Le paramètre a dépend de la nature du corpus et de la langue considérée. Ainsi il est établi que pour l'anglais a est compris entre 0.4 et 0.6 (Araújo *et al.*, 1997).

Etudions maintenant le rapport entre le nombre de mots distincts p et le nombre de mots des contextes q. Dans une première approximation, nous avons dit que ces nombres étaient au pire égaux soit $p = q$. Or cette situation peut se présenter. Les mots grammaticaux (conjonctions, articles, auxiliaires...) ou généraux (en particulier ceux qui ne figurent pas dans la liste des mots vides utilisée pour filtrer le corpus), apparaissent dans beaucoup de contextes donc auront des vecteurs de contexte très importants. L'approximation $p = q$ n'est donc pas déraisonnable.

En résumé, la complexité en temps de la construction des vecteurs de contexte est majorée par $K.n^a.K.n^a$ ou bien $f.n$ (taille de la fenêtre fois taille de l'entrée) soit en $O(max(f.n, n^{2a}))$. La complexité en espace est en $O(n^{2a})$ car l'espace est utilisé pour stocker la matrice des vecteurs de contexte de dimension au pire $(K.n^a, K.n^a)$. La pondération des vecteurs de contexte se fait simultanément à leur construction. Les différents systèmes de pondération font intervenir les cooccurrences et les occurrences des mots. Ainsi pour un mot donné, le poids de chaque mot de contexte est fonction du nombre de cooccurrences entre lui-même et le mot en question et leurs nombres respectifs d'occurrences. Le calcul du nombre d'occurrences des mots est en $O(n)$. Le calcul du nombre des cooccurrences est simultané à la construction des vecteurs. La pondération ne change donc pas la complexité annoncée en $O(max(f.n, n^{2a}))$ en temps et en $O(n^{2a})$ en espace.

Transfert des vecteurs de contexte

Pour le transfert des vecteurs de contexte, on considère en entrée le couple constitué d'un mot et d'un de ses contextes soit $K.n^a.K.n^a$ couples au maximum. On parcourt donc l'ensemble des contextes de l'ensemble des mots distincts soit une matrice au pire carrée de dimension ($K.n^a, K.n^a$).

L'algorithme consiste à transformer chaque mot d'un vecteur de contexte à l'aide d'un lexique bilingue. L'accès au lexique étant en temps constant, la complexité en temps est donc majorée par $K.n^a.K.n^a$ soit en $O(n^{2a})$. La complexité en espace est identique en $O(n^{2a})$.

Calcul de la similarité (dans une direction de traduction)

Nous considérons que l'entrée du calcul de la similarité entre vecteurs de contexte est là aussi l'ensemble des occurrences des corpus. Les formules des mesures de similarité (cosinus, Jaccard ou Dice) imposent un parcours linéaire des deux vecteurs à comparer tout en sachant que seuls les contextes communs font intervenir en même temps les deux vecteurs de contexte. Il y a au maximum $K.n^a.K.n^a$ couples mot/contexte pour les deux corpus. Calculer une similarité revient à comparer chaque couple mot/contexte d'un corpus avec au plus chaque couple mot/contexte de l'autre corpus. Pour chaque mot du corpus source on repère les contextes communs avec chaque mot du corpus cible. Pour un contexte donné, l'accès au contexte commun avec le mot cible est constant en temps. Il y a donc au plus $K.n^a.K.n^a.K.n^a$ opérations. La complexité en temps est donc majorée par $(K.n^a)^3$ soit en $O(n^{3a})$.

Les résultats sont ensuite ordonnés pour chaque mot par similarité décroissante. Ainsi pour chaque mot parmi au maximum $K.n^a$ mots est appliqué un algorithme de tri des $K.n^a$ maximum similarités dont la complexité est en $O(N.log(N))$. Par conséquent, la complexité du tri par similarité décroissante est en $O(n^a.n^a.log(n))$ soit $O(n^{2a}.log(n^a))$.

La complexité totale en temps de ces deux dernières opérations est le maximum des complexités soit $O(n^{3a})$. En espace, l'algorithme manipule une matrice de contextes au pire carrée de dimension ($K.n^a, K.n^a$) ainsi qu'une matrice ($K.n^a, K.n^a$) de similarités. La complexité en espace est en $O(n^{2a})$.

Calcul de la similarité croisée

Le calcul de la similarité croisée est effectué à partir de deux similarités simples correspondant à chaque direction de traduction des langues. Il consiste à comparer ces similarités simples. Nous introduisons ici un nouveau paramètre s. Il s'agit pour une similarité dans une direction donnée de langues du nombre de mots candidats retenus de la langue cible, proches d'un mot de la langue source. Dans nos expériences, s vaut 30.

Pour chaque similarité dans une direction associant deux mots on examine les similarités de l'autre direction. Ainsi pour chaque paire de mots associés dans une direction ($(s.K.n^a)$ paires au plus) on scrute chaque paire associée dans l'autre direction ($(s.K.n^a)$ paires au plus). L'implémentation permet de procéder en temps constant à cette étape pour chaque paire de mots associés. La complexité est donc en $O(s.n^a)$.

En espace, on manipule une matrice ($K.n^a, K.n^a$) de mots associés par la similarité croisée. La complexité en espace est donc en $O(n^{2a})$.

Complexité théorique de l'algorithme complet

Les algorithmes s'enchaînant, la complexité de l'algorithme complet est la complexité maximum rencontrée. L'entrée de l'algorithme complet est constitué par les occurrences des corpus soit n occurrences, la taille de la fenêtre de mots f et le nombre de candidats retenus s.

Segmentation : en temps $O(n)$, en espace $O(n)$.

Construction de vecteurs : en temps $O(max(f.n, n^{2a}))$, en espace $O(n^{2a})$.

Transfert de vecteurs : en temps $O(n^{2a})$, en espace $O(n^{2a})$.

Calcul de la similarité simple : en temps $O(n^{3a})$, en espace $O(n^{2a})$.

Calcul de la similarité croisée : en temps $O(s.n^a)$, en espace $O(n^{2a})$.

La complexité maximale en temps est soit celle de la construction de vecteurs, soit celle du calcul de la similarité simple. Dans le cas où $a < 1/3$, la complexité maximale est en $O(f.n)$ et sinon en $O(n^{3a})$. En résumé, la complexité en temps de l'algorithme complet est donc en $O(max(f.n, n^{3a})$ avec a dans l'intervalle [0;1] et f >1. La complexité en espace est soit en $O(n^{2a})$, soit en $O(n)$. Dans le cas où $a < 0.5$, elle est en $O(n)$ et dans le cas inverse elle est en $O(n^{2a})$. La complexité en espace est donc en $O(n^{max(1,2a)})$.

Pour généraliser ce résultat, appelons g la fonction qui lie pour un corpus donné le nombre d'occurrences au nombre de mots distincts, soit $p = g(n)$ (à une constante multiplicative ou additive près). Les raisonnements précédents restent valides avec la fonction g. On obtient donc le tableau des complexités pour les traitements. La complexité en temps est donc $O(max(f.n, g^3(n)))$ et celle en espace $O(max(n, g^2(n)))$.

Revenons à la nature de la fonction g. Heaps ne donne pas d'indication sur la validité de sa fonction sur les grandes tailles de corpus. Il s'avère que sur les grands corpus le nombre de mots distincts est surestimé (Yang *et al.*, 2000). Ce même auteur estime et vérifie expérimentalement que la fonction g ne peut être réduite à une fonction simple mais est composée de plusieurs fonctions définies sur des intervalles différents.

Sur un premier intervalle de taille de corpus, la formule de Heaps est applicable. Sur un deuxième intervalle, la formule de Heaps est encore applicable mais les paramètres changent. En particulier, pour une taille suffisamment importante le paramètre

Traitement	**complexité en temps**	**complexité en espace**
Segmentation	$O(n)$	$O(n)$
Construction de vecteurs	$O(max(f.n, g^2(n)))$	$O(g^2(n))$
Transfert de vecteurs	$O(g^2(n))$	$O(g^2(n))$
Calcul de la similarité classique	$O(g^3(n))$	$O(g^2(n))$
Calcul de la similarité croisée	$O(s.g^(n))$	$O(g^2(n))$

TAB. 3.10 – *Table de complexité.*

a est proche de 0.3. Cela a une conséquence sur la complexité. Pour n suffisamment grand, $max(f.n, n^{3a})$ vaut $f.n$. La complexité en espace et en temps est donc proche de $O(n)$ pour n suffisamment grand.

Pour le plus grand des corpus utilisés dans nos expériences, les différents paramètres précisés au chapitre 4 sont les suivants :

n = 29 536 583 occurrences (deux langues cumulées).

p = 359766 mots distincts (deux langues cumulées).

f = 5 ou 7 mots.

s = 30 candidats.

On remarque que p^3 est largement supérieur à $f.n$ et que p^2 est lui aussi supérieur à n. La complexité en temps est donc plutôt en $O(p^3)$ et la complexité en espace est en $O(p^2)$. Il faut néanmoins prendre en considération qu'il s'agit d'une complexité théorique donc calculée dans le pire de cas. En pratique, la matrice constituée des vecteurs de contexte est creuse (remplie de 0). Par conséquent pour la construction des vecteurs on ne considère qu'un nombre réduit de mots de contextes et non l'ensemble. Un des facteurs multiplicatifs qui entre en jeu dans le calcul de la complexité théorique est donc largement surestimé.

Chapitre 4

Expériences d'acquisition lexicale bilingue dans le domaine médical

4.1 Introduction

Dans les chapitres précédents nous avons exposé d'une part les fondements théoriques du modèle que nous proposons et d'autre part les différents traitements qui le composent. Ce chapitre relate les expériences menées pour le valider dans le cadre de l'acquisition lexicale bilingue dans le domaine médical.

Ce modèle exploite des corpus des deux langues pour construire les distributions des mots ainsi que des ressources lexicales bilingues pour leur comparaison. Nous présentons dans un premier temps la constitution des ressources propres au domaine médical destinées à ces expériences. Dans beaucoup d'étapes de l'approche générale que nous adoptons plusieurs solutions sont possibles : différentes tailles de fenêtre de contexte, mesures de similarité, systèmes de pondération, etc. Plusieurs expériences portant sur ces différentes alternatives ont été menées et décrites dans un deuxième temps. Elles visent à mettre en évidence de façon expérimentale les limites ou avantages des différentes possibilités. Enfin, l'expérience centrale tente de mettre en évidence l'apport de la similarité croisée sur les méthodes classiques.

4.2 Constitution de ressources bilingues

4.2.1 Constitution de corpus comparables du domaine médical : exploitation des sites-catalogues spécialisés

Par corpus comparables, nous désignons des ensembles de textes pour chaque langue qui portent sur le même domaine de spécialité, en l'occurrence le domaine médical. Comme mentionné précédemment, il est difficile de recenser des textes représentatifs d'un domaine vaste, dans lequel se trouvent des sous-domaines, voire des

micro-domaines spécialisés (chapitre 2). C'est notamment le cas du domaine médical.

L'Internet est une source majeure de données textuelles numérisées. Il présente en même temps une grande facilité d'accès à l'information mais également le problème de sa sélection dans l'avalanche d'informations qu'il propose. Pour ne pas courir le risque d'engendrer trop de bruit, ni celui d'être trop restreint à certaines disciplines, nous avons décidé d'exploiter des sites-catalogues spécialisés dans le domaine médical plutôt que d'utiliser les moteurs de recherche généralistes. Pour recenser les documents publiés dans des sites médicaux accessibles sur l'Internet, nous avons choisi CISMeF (Darmoni *et al.*, 2000) (www.chu-rouen.fr/cismef) pour construire le corpus français, et CliniWeb (Hersh *et al.*, 1999) (www.ohsu.edu/cliniweb) pour le corpus anglais.

Le choix de ces sites-catalogues spécialisés nous permet d'une part, de procéder au filtrage et à la sélection des textes de qualité propres au domaine et d'autre part, d'utiliser l'indexation des documents par les mots clés du thésaurus MeSH (Medical Subject heading). Cette indexation présente une classification organisée et hiérarchique des informations. Cela permet de construire des corpus comparables traitant les mêmes sujets dans les deux langues, en rapatriant par exemple des documents indexés avec des termes MeSH appartenant au même concept. Par ailleurs, la mise à jour régulière des informations sur ces sites est un atout important pour l'acquisition de lexique (section 2.1) et notamment pour leur actualisation.

Avant de procéder à la collecte de tous les documents indexés par les deux catalogues CISMeF et CliniWeb, nous avons choisi un thème plus restreint pour constituer dans un premier temps un sous-corpus que nous appelons *corpus C23*. Le code "C23" correspond à l'un des concepts du thésaurus MeSH : signes et symptômes, états pathologiques (*Pathological Conditions, Signs and Symptoms*). La première étape consiste alors à établir, pour chacun des deux catalogues, une liste d'urls indexés par les mots-clés MeSH appartenant à ce concept. Le nombre d'urls ainsi recensés est de 2 338 pour CISMeF et de 921 pour CliniWeb. A partir de cette liste, nous avons téléchargé, à l'aide de l'utilitaire *wget* de GNU, les pages correspondantes et celles pointées par un lien directe dans ces pages. En principe, les pages téléchargées ne contiennent pas d'images[1]. Le résultat comprend 10 539 fichiers pour CISMeF et 2 036 pour CliniWeb. Notons que les chiffres annoncés ici proviennent d'expériences réalisées fin 2001. Actualisés, ils peuvent être différents pour plusieurs raisons : la mise à jour des index, le fonctionnement et la disponibilité d'accès des sites indexés, etc.

La dernière collecte des urls pour ces expériences a été effectuée en Octobre 2003. Il s'agit de rapatrier cette fois toutes les pages indexées pour les deux sites-catalogues. Le résultat donne 56 635 fichiers téléchargés de CISMeF et 14 826 fichiers téléchargés de CliniWeb. Ces fichiers sont la base du corpus complet que nous notons *corpus TOUT* par la suite.

Malgré le filtrage d'images avec *wget*, les documents provenant des différents

1. l'option –reject de *wget* permet de préciser les types de fichiers à ne pas prendre en considération pendant le téléchargement, http://www.gnu.org/software/wget/manual/wget-1.8.1/html_chapter/wget_2.html\#SEC12.

sites médicaux ainsi collectés présentent une hétérogénéité de formats (.html, .pdf, .txt, .doc, .css, etc) et nécessitent certains traitements avant d'être exploités pour l'analyse de corpus.

Mise au format texte

La majorité des fichiers téléchargés sont des fichiers HTML (extension .htm ou .html) : 87,4% pour CISMeF et 92,1% pour CliniWeb. Nous utilisons le module HTML::FormatText du Perl pour la conversion en format texte (.txt)[2]. Pour les fichiers PDF, nous employons *pdftotext*[3] pour les convertir en texte. Pour les fichiers sans extension, le test avec l'utilitaire *file* est appliqué pour définir leur type afin de choisir le programme de conversion le plus approprié. Certains fichiers sont vides après la conversion, cela peut être expliqué par le fait que les pages téléchargées ne contiennent que des liens ou des *frames*.

Filtrage de langue

Lors des premières expériences sur le sous-corpus C23, nous avons pu constater la présence d'autres langues que le français (*e.g.*, l'anglais, l'espagnol, etc) dans les pages indexées par CISMeF. Cela nous a conduit à effectuer un filtrage de langue sur tous les fichiers convertis en texte. En reprenant les données utilisées dans l'expérience de (Grefenstette & Nioche, 2000), un programme Perl sélectionne les lignes respectant les critères d'identification[4] de l'anglais pour les textes provenant de CliniWeb et du français pour les textes issus de CISMeF.

Après la conversion au format texte et le filtrage de langue, il reste 32 951 fichiers de type texte pour la partie CISMeF et 11 755 fichiers texte pour la partie CliniWeb. Le tableau 4.1 illustre quelques statistiques (nombre de fichiers, de mots et d'octets) sur la composition des deux parties des corpus comparables. Le calcul est effectué avec l'utilitaire *wc* de GNU. Il faut mentionner que pour le sous-corpus C23, aucun filtrage de langue n'a été appliqué.

Remarques sur les corpus résultants

En examinant les corpus résultant des différents traitements, deux remarques s'imposent. Ils sont d'une part bruités et d'autre part de tailles différentes.

Malgré l'application des traitements cités ci-dessus, les textes provenant de CISMeF et de CliniWeb présentent beaucoup de bruit : fautes d'orthographe, formats dif-

2. Dans les premières expériences (Chiao & Zweigenbaum, 2003), l'utilitaire lynx a été utilisé avec les options -dump -nolist pour convertir les fichiers HTML en texte brut. Nous avons constaté que *lynx* garde les noms des fichiers image lors de la conversion. Cela introduit beaucoup de bruit dans le corpus.

3. Les options -raw et -layout ont été appliquées en fonction de la mise en page des fichiers PDF.

4. Il s'agit des critères statistiques basés sur l'estimation de fréquences des mots les plus utilisés dans une langue donnée.

Corpus	Nb de fichiers	Nb de mots	Nb d'octets
Corpus **TOUT**, converti avec *Perl, pdftotxt* avec filtrage de langues			
CISMeF	32 951	54 464 201	391 490 352
CliniWeb	11 755	7 619 493	54 020 048
Sous-corpus **C23**, converti avec *lynx, pdftotxt* sans filtrage de langues			
CISMeF	10 539	16 741 151	136 423 462
CliniWeb	2 036	1 074 523	9 225 777

TAB. 4.1 – *Tailles des corpus provenant de CISMeF et de CliniWeb en nombre de fichiers, de mots et d'octets.*

férents à la sortie des convertisseurs utilisés (*Lynx*, *Perl*, *pdftotxt*), présence des mots en langue étrangère, etc. (*dysfonctionnement*/*disfonctionnement* ; *zugänglich*, *lõabsentzisme*, *líutilisation*, *accueilretour*, *virusvih*, etc.). Cela rend plus difficile leur exploitation et a un impact sur les résultats des prétraitements (chapitre 3) nécessaires pour l'extraction de traduction. C'est selon nous la cause principale des erreurs de segmentation et de l'analyse de cooccurrences.

La différence de taille entre les deux corpus CISMeF et CliniWeb peut avoir un impact sur la performance du modèle proposé. En particulier des distributions de fréquences des mots trop éloignées risquent de biaiser les comparaisons des mots à traduire et des mots candidats à la traduction. Si un mot dans une langue présente moins de contextes que sa traduction dans l'autre langue, l'appariement sera moins performant.

4.2.2 Construction d'un lexique bilingue de base

Nous décrivons dans cette section les ressources bilingues de différentes natures qui ont servi à constituer un lexique bilingue de base. Il est le résultat d'une combinaison de deux dictionnaires bilingues : l'un médical et l'autre général. Il est composé de 22 036 entrées dont 20 803 sont des mots simples, et le reste des mots composés liés par le trait d'union.

Ce lexique résultant est destiné à jouer le rôle de passerelle entre les deux langues. Il est en particulier utilisé dans le processus que nous avons appelé le transfert des vecteurs de contextes qui consiste à normaliser ces vecteurs. Le transfert vise entre autres à traduire les vecteurs français vers l'anglais pour qu'ils puissent être comparés avec les vecteurs anglais. Dans le cas où une entrée possède plusieurs traductions possibles (c'est souvent le cas des entrées polysémiques), plusieurs méthodes sont possibles pour le choix de la traduction. Une méthode consiste à utiliser toutes les traductions du lexique en leur attribuant le même poids. Néanmoins le bruit introduit par cette méthode nous a incité à ne prendre en compte que la première traduction proposée dans le lexique sachant qu'il s'agit de la traduction la plus fréquente. Cette heuristique permet de minimiser le risque d'un mauvais choix sans pour autant fournir la bonne traduction

de façon systématique.

Ressource lexicale spécialisée

Plusieurs ressources bilingues spécialisées en médecine ont été utilisées afin de construire le lexique bilingue de base adapté aux besoins de notre modèle. Le dictionnaire de base utilisé est un dictionnaire médical français en ligne *Dictionnaire Médical Masson*[5] qui regroupe 30 000 termes utilisés dans tous les domaines de la médecine actuelle. La plupart des entrées ont une ou plusieurs traductions en anglais.

Pour chaque entrée sont extraites des sous-entrées, des synonymes et des traductions s'ils existent, dans le but d'établir une liste bilingue contenant les entrées des mots simples et des mots composés de plusieurs mots liés par le trait d'union. Le nombre d'entrées de cette nouvelle liste est de 17 732. Les entrées des mots composés avec la présence d'un blanc ou d'un espace ne sont pas extraites car elles posent le problème de leur reconnaissance en corpus.

Nous utilisons aussi le métathésaurus UMLS pour enrichir ce lexique de base ainsi constitué. La version UMLS2000 comprend environ 730 000 concepts et 1,5 millions de termes différents, issus de plus de 60 sources de vocabulaires différentes (MeSH, SNOMED, CIM, DSM...) (NLM, 2000). Nous avons extrait les thésaurus bilingues français/anglais : ICPCFRE (723 termes), INS2000 (28 775 termes MeSH traduits en français par l'INSERM[6]), SNMI98 (164 180 termes) et WHOFRE (3 717 termes). Dans un premier temps, les entrées du lexique de mots simples en construction et celles des thésaurus issus de l'UMLS ont été alignées. Le nombre d'entrées alignées est de 8 531 mots simples : 79 pour (ICPXFRE), 800 pour (WHOFRE), 3 245 pour (INS2000) et 4 407 pour (SNMI98). A l'aide de ces thésaurus, 705 nouvelles entrées de mots simples ont été ajoutées au lexique de base spécialisé en construction. Le lexique ainsi construit contient finalement 18 437 entrées de mots français avec leurs traductions en anglais.

Ressource lexicale générale

Pour constituer un lexique bilingue de mots généraux, nous avons extrait une liste bilingue de 4 272 mots simples à partir du dictionnaire bilingue fourni dans la distribution des logiciels Linux[7]. Notons que cette liste est complémentaire du lexique spécialisé construit précédemment. Il n'y a pas de doublon dans les entrées des deux listes. L'intérêt de disposer de telles ressources sera montré en évaluant l'effet positif de la prise en considération des mots généraux dans les vecteurs de contextes (section 4.3.6) sur l'extraction de lexique spécialisé bilingue.

5. http://www.atmedica.com

6. Institut National de la Santé Et de la Recherche Médicale (INSERM). http://disc.vif.inserm.fr:2010/basismesh/mesh.html

7. Dans le paquet de *dictd-dictionaries* i586-linux.

4.3 Expériences d'extraction de lexique bilingue spécialisé

L'objectif de ces expériences est d'évaluer dans quelle mesure le modèle proposé trouve la bonne traduction d'un mot donné parmi les candidats qu'il propose. Pour permettre cette évaluation il est nécessaire de disposer d'un ou plusieurs lexiques de test bilingues. Un lexique de test est ici une liste de mots (et de leurs traductions) pour lesquels le modèle à évaluer va proposer des candidats à la traduction qui seront comparés aux véritables traductions. Notre modèle propose ainsi pour un mot donné de ce lexique, une liste de candidats ordonnée selon les valeurs d'une mesure de similarité. La méthode d'évaluation consiste donc à repérer, pour un mot dans le lexique de test, sa traduction attendue dans une liste de n candidats proposés.

La valeur de n a été fixée à 30 essentiellement pour la raison que cette liste de candidats est destinée à une validation manuelle ultérieure dans la cadre de l'actualisation de lexique bilingue. Pour que la validation reste une opération raisonnable en temps il faut limiter le nombre de candidats proposés à au plus quelques dizaines. La mesure de la performance est ainsi calculée en tenant compte du rang de la traduction attendue parmi les candidats proposés. Le principe de l'évaluation repose sur l'hypothèse que plus les traductions attendues sont classées parmi les premières traductions proposées, meilleure est la performance du modèle.

Dans un premier temps, nous décrivons les différents lexiques de test utilisés comme listes de mots à traduire. Puis nous précisons comment sont construites les listes des candidats à la traduction. Une section est consacrée à la façon de présenter les résultats. Nous explicitons ensuite les paramètres étudiés et commentons les résultats de leurs évaluations. Enfin, avant de conclure nous présentons les résultats de l'évaluation de l'apport de la similarité croisée par rapport aux approches classique.

4.3.1 Lexiques de test

Liste des mots pivots P

Le premier des lexiques de test est appelé liste des mots pivots P. Deux listes bilingues P ont été constituées à partir des corpus respectifs C23 et TOUT. Une liste P est constituée de couples bilingues de mots simples (composés d'une seule unité lexicale) et pleins qui font partie du lexique bilingue de base et dont chaque mot d'une langue donnée est présent dans la partie monolingue correspondante du corpus traité. Ainsi la liste des mots pivots P du corpus C23 est composée d'un nombre total de 6 210 couples de mots simples français-anglais qui sont traductions mutuelles et sont inclus dans le lexique bilingue de base. Chaque mot de P apparaît dans la partie source (CISMeF) ou la partie cible (CliniWeb) du corpus C23. Pour le corpus TOUT, la liste des mots pivots comprend 8946 couples de mots.

Les mots pivots ont un statut particulier. Ils constituent en effet les mots des

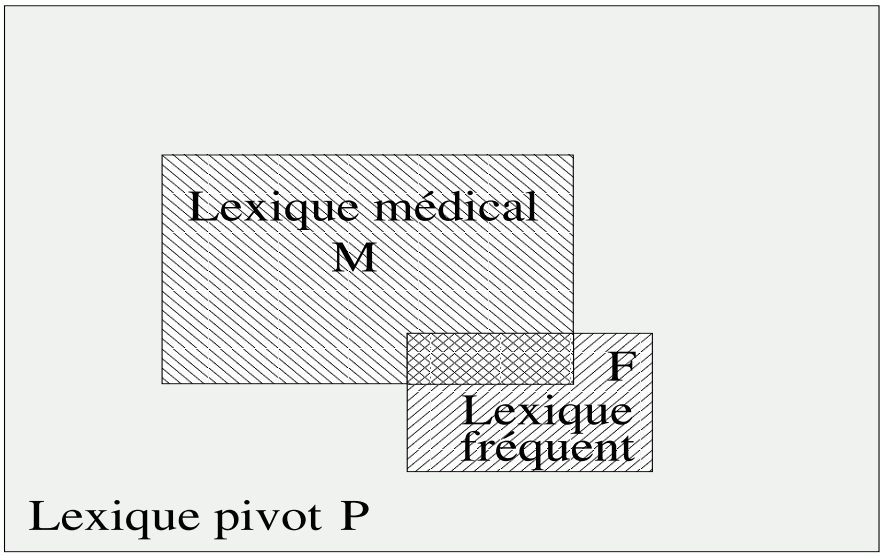


FIG. 4.1 –

contextes qui peuvent être comparés d'une langue à l'autre. A ce titre, chaque mot pivot est une dimension de l'espace vectoriel des contextes sur lequel est constitué le vecteur de contextes de chaque mot à traduire et de chaque mot candidat. Nous sommes donc en présence d'un espace vectoriel de dimension 6 210 et 8 946 pour les corpus respectifs C23 et TOUT.

Pour le corpus C23, de la liste des mots pivots a été extraite une liste de mots spécialisés, c'est-à-dire appartenant aux ressources lexicales spécialisées décrites dans la section précédente. Cette nouvelle liste notée P' contenant 4 963 couples de mots est utilisée dans l'expérience, qui étudie l'impact de la nature des contextes sur les performances.

A partir d'une liste de mots pivots, deux sous-listes ont été créées selon deux critères différents : celui de la spécialisation des mots et celui de leur fréquence. Par la suite, la liste des mots spécialisés est notée liste M et la liste des mots fréquents est notée liste F.

Liste Snomed M

La liste M est établie selon le critère de spécialisation des mots. Elle est composée des mots issus du thésaurus médical SNOMED (Côté, 1996) qui font partie de la liste des mots pivots. Pour le corpus C23, la liste M comprend 886 couples de mots. Pour le corpus TOUT, elle en comprend 1086. Contrairement à ceux de la liste F décrite après, les mots de la liste M peuvent être peu fréquents dans le corpus C23 : f(pellagre) = 9 ; f(thrombocyte) = 3, etc. Certains mots peuvent être fréquents comme f(bactérie) = 720 ; f(sodium) = 500.

Liste des mots fréquents F

Certains paramètres étudiés (sections 4.3.6 et 3.2.2) dans ces expériences nécessitent une présence minimum des cooccurrences pour optimiser le calcul statistique. C'est la raison pour laquelle nous avons établi une liste de mots en fonction de leur fréquence élevée dans les deux parties monolingues du corpus. Une liste bilingue de 96 mots fréquents a été ainsi extraite à partir du corpus C23 et une autre de 90 mots à partir du corpus TOUT.

Les mots présents dans cette liste peuvent être des mots généraux tels que *analyse, sang, eau, mortalité*, ou des mots spécialisés comme *chirurgie, métastatique, prostate*, etc.

4.3.2 Candidats à la traduction

Les candidats à la traduction sont extraits de la partie cible (partie anglaise) des corpus (CliniWeb). Deux listes de candidats à la traduction ont été constituées. Ces deux listes ont des finalités distinctes.

La liste T est composée de tous les mots du corpus CliniWeb, qu'ils soient inconnus ou connus dans le lexique bilingue de base. La liste T correspond à 28 633 mots pour le corpus C23 et à 70 909 mots pour le corpus complet TOUT. Cette liste complète est destinée à associer à un mot d'une langue donnée, non seulement sa traduction mais aussi des mots proches dans l'autre langue. Le fait de conserver dans cette liste les mots connus du lexique permet en effet de proposer des mots connus comme candidats. Mais en dehors de la traduction attendue, s'ils sont connus ils ne sont a priori pas les bonnes traductions mais peuvent se révéler sémantiquement proches. Une des façons d'exploiter ces mots proches est de les utiliser dans l'extension de requêtes en recherche d'information translangue (chapitre 5 et 6).

La liste U contient tous les mots inconnus de la partie cible de CliniWeb. La liste U correspond à 22 424 mots pour le corpus C23 et à 61 854 mots pour le corpus complet TOUT. Afin de pouvoir calculer le rang de la traduction attendue à des fins d'évaluation, pour un mot donné, la traduction attendue (qui est un mot connu) de chaque mot a été ajoutée à cette liste. Cet ajout est en pratique réalisé dynamiquement pour un mot donné à chaque calcul des mots candidats à la traduction. Ainsi dans la liste U ne sont présents que des mots inconnus plus la traduction attendue du mot en cours de traduction. La finalité de la liste U est plutôt d'actualiser des lexiques bilingues, *i.e.*, d'acquérir la traduction de nouveaux mots qui ne sont pas dans le lexique.

Pour optimiser les temp de calcul et uniquement pour les premières expériences portant sur l'évaluation de certains paramètres sur le corpus C23, un sous-ensemble de mots fréquents de la liste U a été constituée. Il contient 6 018 mots inconnus dont la fréquence est supérieure à 5. Nous le nommons liste $U\prime$. Pour les mêmes raisons, ces premières expériences ont été aussi menées sur une liste réduite de mots connus, en l'occurrence la liste $P\prime$ des mots pivots connus spécialisés décrite dans la section

précédente, soit 4 963 mots.

A ce point de la présentation des expériences, il est intéressant de s'arrêter sur leur niveau de difficulté. L'objectif de ces expériences est, pour un mot donné de trouver sa bonne traduction parmi plus de 60 000 candidats! L'espace de recherche est très important car la comparabilité des corpus ne permet pas de le réduire. A titre de comparaison, dans des corpus non plus comparables mais alignés au niveau des phrases, chercher la traduction d'un mot apparaissant 10 fois revient à la chercher parmi au plus 100 mots (en faisant l'hypothèse d'une taille moyenne de phrase de 10 mots pleins). Même si cette comparaison est approximative, elle montre que la difficulté à traiter les corpus comparables est supérieure de plusieurs ordres de grandeurs à celle à traiter des corpus alignés.

4.3.3 Choix de la présentation des évaluations

Nous avons fait varier au cours de ces expériences trois types de paramètres et avons choisi plusieurs façons de les présenter. Les types de paramètres évalués, présentés dans les sections qui suivent, sont les suivants :

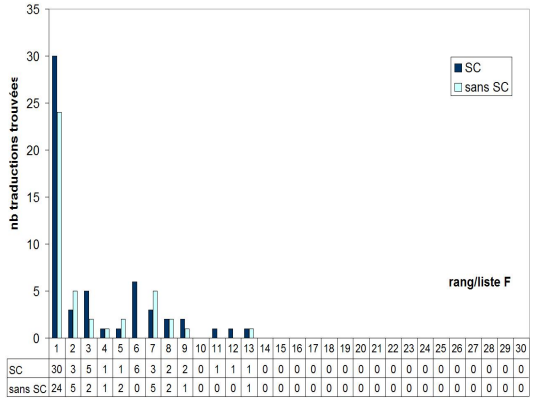
– La taille de corpus.

– La construction des vecteurs de contexte.

– La comparaison des vecteurs de contexte.

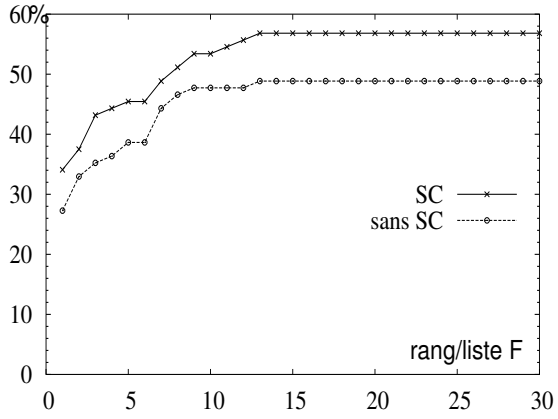
La première façon de présenter les résultats et la plus utilisée ici est une présentation du pourcentage cumulé de mots bien traduits par rapport à l'ensemble des mots à traduire, selon le rang des mots candidats proposés. Le rang provient de la liste des mots candidats ordonnée par similarité décroissante. Deux raisons justifient ce choix. Tout d'abord, la présentation en pourcentage permet de comparer les performances selon différents critères d'évaluation. Tandis que la présentation en valeur absolue (nombre de mots bien traduits) rend la comparaison difficile surtout lorsqu'il s'agit de comparer l'effet de différentes tailles de corpus. De plus, la courbe cumulative permet d'observer l'évolution globale de la performance. Ce que l'on cherche, c'est bien à mesurer le pourcentage de mots bien traduits parmi les n premières propositions et non pas celui à un rang donné.

La figure 4.2 illustre bien la différence entre ces deux types de présentation. Même si la présentation des résultats par rang des candidats proposés est privilégiée ici, une autre façon de présenter les choses a été mise en œuvre. Nous avons mentionné dans le chapitre 3 l'intérêt d'étudier le comportement de la distribution des fréquences dans des corpus comparables. L'effet de la fréquence des mots sur la performance des méthodes d'extraction de lexiques à partir de corpus comparables est une question récurrente (Chiao & Zweigenbaum, 2004). En général, l'évaluation de ces méthodes est limitée aux mots les plus fréquents (Rapp, 1999), parce que les mots fréquents sont



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
SC	30	3	5	1	1	6	3	2	2	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sans SC	24	5	2	1	2	0	5	2	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

a



b

FIG. 4.2 – *Deux présentations des résultats : histogramme en valeur absolue et courbe cumulative en pourcentage. L'axe des X correspond au rang des candidats. L'axe des Y correspond au nombre absolu (a) ou cumulé en pourcentage (b) des traductions trouvées.* ***SC*** *= application de la similarité croisée.*

censés avoir plus de contextes, et ainsi plus de cooccurrences qui favorisent les mesures statistiques sur lesquelles reposent toutes les méthodes d'extraction. Les mots inconnus, *i.e.*, mots qui ne sont pas dans les dictionnaires existants, cas des néologismes, sont considérés comme peu fréquents dans le corpus et donc peu aptes à être traduits par ces méthodes.

Une des idées avancée ici est que la prise en compte de quelques contextes peut tout de même suffir pour trouver la bonne traduction d'un mot. Pour confirmer expérimentalement cette idée, une deuxième façon de présenter les résultats est proposée. Elle consiste à afficher le pourcentage de mots bien traduits par tranches de fréquences homogènes des mots à traduire. L'intérêt d'une telle présentation est de pouvoir mesurer sur quelles tranches de fréquences une méthode fonctionne le mieux.

Un des problèmes de la présentation par fréquence est que pour deux corpus de tailles différentes, les fréquences ne sont pas comparables. Un mot fréquent pour l'un des corpus le sera moins pour l'autre. Pour remédier à ce problème, les tranches de fréquences sont construites de telle façon qu'elles restent comparables d'un corpus à l'autre. L'algorithme qui construit ces tranches consiste à parcourir pour un corpus donné les fréquences des mots triées par ordre décroissant. Chaque mot ainsi parcouru est affecté à la tranche de fréquences en construction. Le passage à la construction d'une nouvelle tranche s'effectue quand la fréquence du mot traité est inférieure de plus de dix pourcents à celle du mot traité précédemment.

4.3.4 Effet de la fréquence des mots

Afin de mettre en évidence l'effet de la fréquence des mots sur les performances, nous présentons dans un premier temps les résultats de l'extraction de lexique bilingue en fonction de la fréquence des mots sur les deux corpus de tailles différentes : TOUT et C23.

Les figures 4.3 (TOUT) et 4.3 (C23) illustrent la performance de l'extraction pour les lexiques de test P et M en considérant les 10 premiers rangs de la liste des candidats inconnus U. Les résultats sont présentés par tranches de fréquences décroissantes ($R_{distribution}$). La première couvre les mots les plus fréquents et la dernière correspond au cas des hapax. Par exemple dans le corpus TOUT, le mot français le plus fréquent est *santé* avec 132 564 occurrences ($R_{distribution} = 1$). Dans le corpus C23, le mot le plus fréquent est *ca* (pour calcaire) avec 36 266 occurrences. Il est vrai que

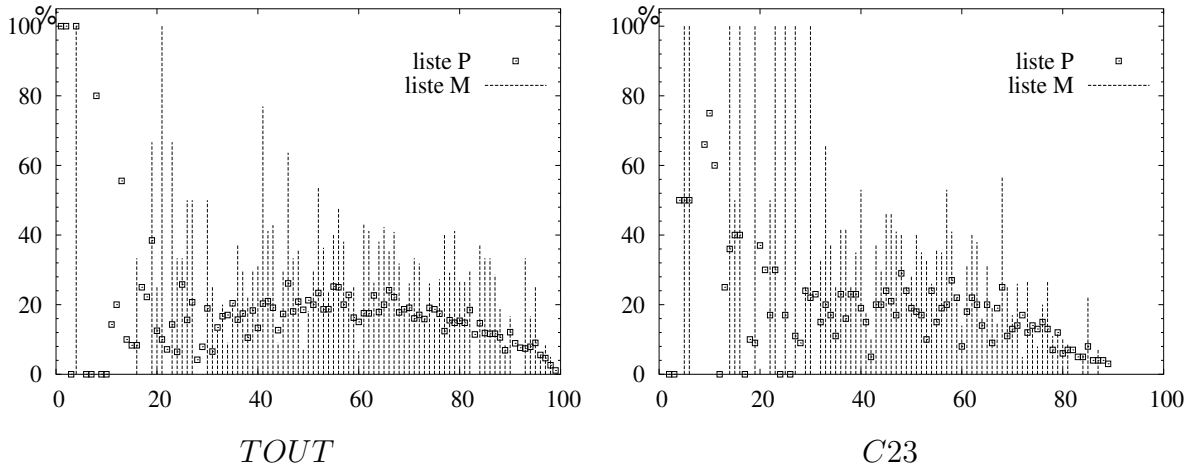


FIG. 4.3 – *Pourcentages de mots bien traduits parmi les 10 premiers candidats pour les deux corpus C23 et TOUT en fonction de la distribution des fréquences des mots.*

notre modèle obtient de meilleurs résultats en traduisant les mots les plus fréquents, qu'ils soient généraux ou médicaux (liste P), ($R_{distribution} <= 15$, *i.e.*, fréquence >= 19 890 dans le corpus TOUT et >= 7 689 dans le corpus C23). Ainsi plus de 50% des mots fréquents de la partie française du corpus ont été correctement traduits parmi les 10 premiers candidats proposés. On remarque aussi que la performance décroît pour les très basses fréquences. Néanmoins il est notable qu'en dehors des tranches de très hautes ou très basses fréquences, notre modèle affiche une performance uniforme. On peut en conclure que la portée de cette méthode ne concerne pas uniquement les mots très fréquents mais également les mots de moyenne ou faible fréquence.

L'examen de la figure précédente nous révèle aussi que plus que la fréquence, c'est la nature des mots à traduire qui a un impact sur la performance de leur traduction. En effet, les résultats de la traduction des mots médicaux (liste M) montrent que

le modèle améliore nettement la performance par rapport à celle des mots plus généraux (liste P) quelque soit la fréquence des mots. Par exemple, parmi les 10 premiers candidats proposés, pour les mots médicaux ayant une fréquence entre 41 et 45, (correspondant à la tranche 68 dans le corpus C23), plus de la moitié (57,14%) ont été bien traduits, alors que qu'ils ne sont que 25,19% lorsque l'on cherche à traduire aussi les mots généraux. De même pour le corpus TOUT, 40% des mots médicaux ayant une fréquence entre 49 et 53 ($R_{distribution} = 77$) ont été correctement traduits alors qu'ils ne sont que 12,4% si les mots du lexique général sont également pris en compte.

Cette meilleure performance pour la traduction des mots médicaux peut être expliquée par le fait que les mots de spécialité sont en général moins ambigüs dans un corpus d'un même domaine. Une autre explication possible est que le modèle prend en compte la dépendance collocationnelle mentionnée en 2.3.1 et 3.2 en considérant des fenêtres réduites de contextes. Ce phénomène est plus marqué pour les mots de spécialité (McKeown & Radev, 2000). Par exemple, le mot *moignon/stump* cooccurrent souvent avec *amputation/amputation* dans les deux parties monolingues du corpus ce qui favorise leur appariement. La prise en compte des dépendances collocationnelles est peut-être une explication également de la bonne tenue de notre modèle sur les moyennes et basses fréquences.

4.3.5 Effet de la taille des corpus

Dans les modèles d'extraction de lexique fondés sur l'analyse statistique du corpus, la taille de corpus est sans doute un paramètre important qui conditionne les résultats. En effet, on peut supposer que plus un corpus est d'un volume important, mieux il peut faire ressortir les régularités entre un mot et sa traduction. L'objectif de cette section est l'étude de l'effet de la taille du corpus sur la performance de notre modèle d'acquisition lexicale bilingue.

Pour cette étude, nous avons utilisé les deux lexiques de test extraits des corpus C23 et TOUT selon les critères décrits dans la section 4.3.1: soit M pour le lexique spécialisé et P pour le lexique des mots pivots. Pour chaque mot des lexiques, les 30 premiers candidats traductions ont été calculés sur la base de la mesure de similarité *Jaccard* et du système de pondération *tf.idf*. Les traductions ont ainsi été extraites parmi 28 633 (respectivement 22 424) mots anglais issus du corpus C23 et 70 979 (respectivement 62 034) mots du corpus TOUT pour la liste de candidats T (respectivement U).

Le tableau 4.2 illustre la différence de taille entre les deux corpus, ainsi que les paramètres caractérisant les corpus.

Dans un premier temps, nous présentons dans le tableau 4.3, pour chacune des deux listes de test issues des deux corpus, le nombre total des traductions trouvées en première position (Nb traductions en $R1$) et le nombre de celles trouvées parmi les 30 premières traductions proposées (Nb traductions en $R30$), par rapport au nombre des traductions attendues. L'ensemble des candidats est l'ensemble U.

	Corpus C23	**Corpus TOUT**
Nombre d'occurrences FR	7 604 381	25 633 353
Nombre d'occurrences EN	639 662	3 903 230
Mots FR différents	130 390	280 857
Mots EN différents	28 633	70 979
Lexique de test P	6 210	8 946
Lexique de test M	886	1 086
Liste de candidats T	28 633	70 979
Liste de candidats U	22 424	62 034

TAB. 4.2 – *Tailles des corpus C23 et TOUT en nombre d'occurrences de mots et nombre des mots différents ainsi que tailles des lexiques de test et des listes de candidats.*

	Nb traductions en R_1 (%)		**Nb traductions en R_{30} (%)**	
Corpus	Liste P	Liste M	Liste P	Liste M
C23	309 (4,98%)	93 (10,61%)	1 028 (16,55%)	246 (27,76%)
TOUT	679 (7,59%)	181 (16,67%)	1 854 (20,72%)	402 (37,02%)

TAB. 4.3 – *Performance de l'extraction de lexique bilingue à partir des corpus C23 et TOUT au rangs 1 et 30.*

Ces premiers résultats montrent l'effet de la taille des corpus. Que ce soit au rang 1 ou au rang 30, les résultats sont meilleurs pour le plus grand des corpus.

L'ensemble des performances comparées des deux corpus sont présentées dans les figures 4.4. Ces résultats ont été produits sans appliquer la similarité croisée. Les résultats sont présentés pour les 30 premiers rangs des traductions proposées en pourcentage cumulé de mots bien traduits. Nous pouvons constater une amélioration globale des résultats sur le corpus entier TOUT par rapport au sous-corpus C23 dans les figures 4.4.

Pour la liste des mots fréquents F, au rang 1, 19,32% et 27,27% sont correctement traduits avec le corpus TOUT contre 7,29% et 17,71% avec le corpus C23, en utilisant respectivement les deux listes des candidats T et U. Cet écart en performance se réduit lorsqu'on prend en considération un plus grand nombre de candidats proposés. Pour cette même liste, parmi les 15 premiers candidats, 42,05% et 48,86% des mots trouvent leur bonne traduction avec le corpus TOUT, contre 39,58% et 47,92% avec le corpus C23. Par exemple, *hepatitis*, traduction du mot *hépatite* est proposé comme premier candidat à la traduction avec le corpus TOUT tandis qu'il est classé en 7ème position avec le corpus C23. De même le mot *tumeur* dont la traduction *tumor* est classée en premier sur la liste des candidats avec le corpus TOUT est en 7ème position avec le corpus C23.

Dans le cas des mots médicaux (liste M) ou des mots pivots (liste P), la meilleure

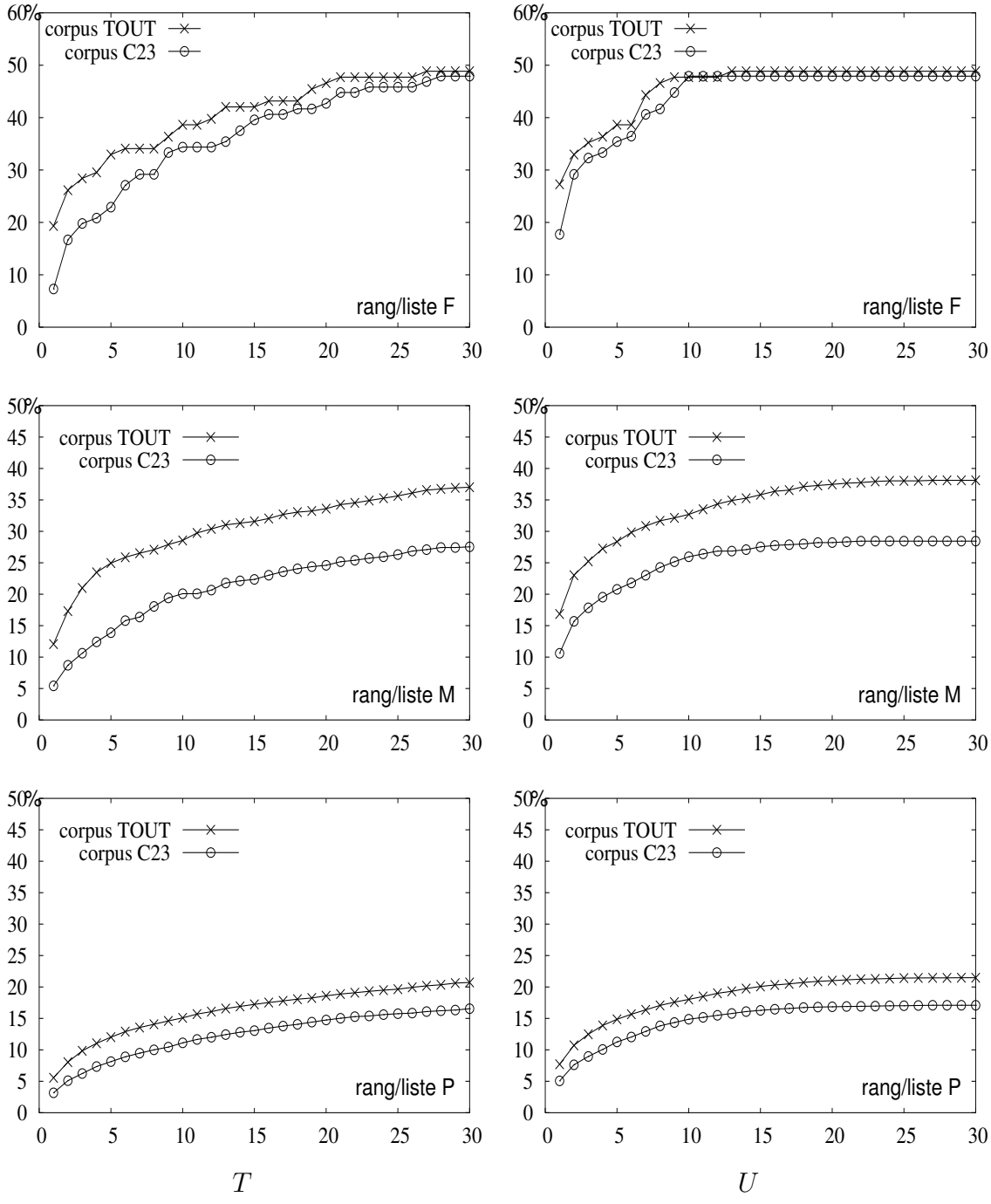


FIG. 4.4 – *Comparaison des performances pour les deux corpus TOUT et C23 sur les listes de test F, M et P à partir des listes de candidats à la traduction de tous les mots du corpus (liste T) et des mots inconnus (liste U).*

performance du corpus TOUT sur le corpus C23 est confirmée d'une façon plus homogène. Cet écart en performance entre les deux corpus est plus important lorsqu'il s'agit d'extraire la traduction des mots spécialisés (rang/liste M). La courbe pour le corpus TOUT domine en moyenne de 9 points celle de C23. Avec la liste P constituée des mots généraux et médicaux, l'écart entre C23 et TOUT est de seulement en moyenne de 3 points.

Le tableau 4.4 montre les mots qui apparaissent le plus souvent dans les vecteurs de contexte de certains mots construits à partir du corpus TOUT et à partir du corpus C23. La tête de vecteur est marquée en gras et suivie des mots figurant dans son contexte, classés par ordre décroissant de cooccurrence. En comparant les vecteurs extraits des deux corpus, on s'aperçoit que les mots qui sont traductions mutuelles partagent plus de contextes en commun dans le corpus TOUT que dans le corpus C23. Cette expérience confirme donc qu'un corpus plus conséquent permet de construire des vecteurs de contextes comprenant plus de mots ce qui permet d'augmenter ainsi les chances d'appariement (Grefenstette, 1996). Notons également que la performance est améliorée avec un corpus plus important alors que le niveau de difficulté est bien supérieur. En effet pour la liste T de candidats, dans le cas du corpus TOUT le modèle doit trouver la bonne traduction d'un mot parmi plus de 70 000 candidats alors que dans le cas du corpus C23 ce nombre est ramené à un peu plus de 28 000.

4.3.6 Paramètres de la construction des vecteurs de contexte

Variation de la taille de la fenêtre de contexte

Rappelons que le contexte d'un mot dans notre modèle est défini au sein d'une même phrase (section 3.3.2) et par une taille de fenêtre de mots pleins. Dans cette expérience nous faisons varier la taille de la fenêtre de mots afin d'observer son effet sur la performance. Pour chaque mot du corpus C23, deux vecteurs de contextes ont été calculés en fonction de deux longueurs de fenêtre de mots : 5 mots et 7 mots. Pour un mot donné, le contexte défini par la fenêtre de 5 mots (respectivement 7 mots) correspond à deux (respectivement trois) mots à gauche et à deux (respectivement trois) mots à droite.

Résultats de l'évaluation de l'effet de la taille de fenêtre

La figure 4.5 présente les résultats d'extraction en utilisant ces deux fenêtres de longueur différente. On constate que la courbe *5 mots* et la courbe *7 mots* sont confondues sauf dans le cas de l'extraction du lexique des mots fréquents (rang/liste F) où la fenêtre de 5 mots produit de meilleurs résultats quelque soit la liste des candidats utilisée (figures 4.5T et 4.5U).

Cette observation indique que la taille de la fenêtre de contexte a un effet lié à la fréquence des mots à traduire. Pour vérifier cette idée, les résultats d'extraction sont présentés en fonction de la distribution des fréquences des mots dans le corpus.

Corpus TOUT				Corpus C23			
Vecteur FR traduit		**Vecteur EN**		**Vecteur FR traduit**		**Vecteur EN**	
vessie		**bladder**		**vessie**		**bladder**	
	cancer		*cancer*		cancer		dysfuction
	tumor		*urinary*		*urinary*		loss
	kidney		*kidney*		prostate		*kidney*
	urinary		*urine*		*kidney*		*urinary*
	urine		*ureter*		lung		normal
	ureter		wall		organ		*ureter*
	neck		*urethra*		*tumor*		feedback
	urethra		prostate		*ureter*		stone
ostéomyélite		**osteomyelitis**		**ostéomyélite**		**osteomyelitis**	
	acute		*chronic*		acute		bone
	infection		*infection*		*infection*		septic
	chronic		*bone*		chronic		*infection*
	arthritis		*acute*		form		arthritis
	septic		*arthritis*		osteitis		spinal
	form		patient		osseous		sickle
	fracture		*septic*		tumor		osteoma
	bone		cellulitis		staphylococcal		meningitis
abeille		**bee**		**abeille**		**bee**	
	wasp		*wasp*		*wasp*		*wasp*
	venom		yellow		gluten		african
	effect		jacket		*venom*		*venom*
	behavior		*venom*		disease		spider
	wax		reaction		fly		reader
	structure		precipitate		ant		mexico
	disease		*effect*		image		effect

TAB. 4.4 – *Mots ayant une forte cooccurrence dans les vecteurs de contexte des mots* ***vessie/bladder, ostéomyélite/osteomyelitis, abeille/bee****, construits à partir des deux corpus de tailles différentes. Les mots soulignés sont des contextes communs.*

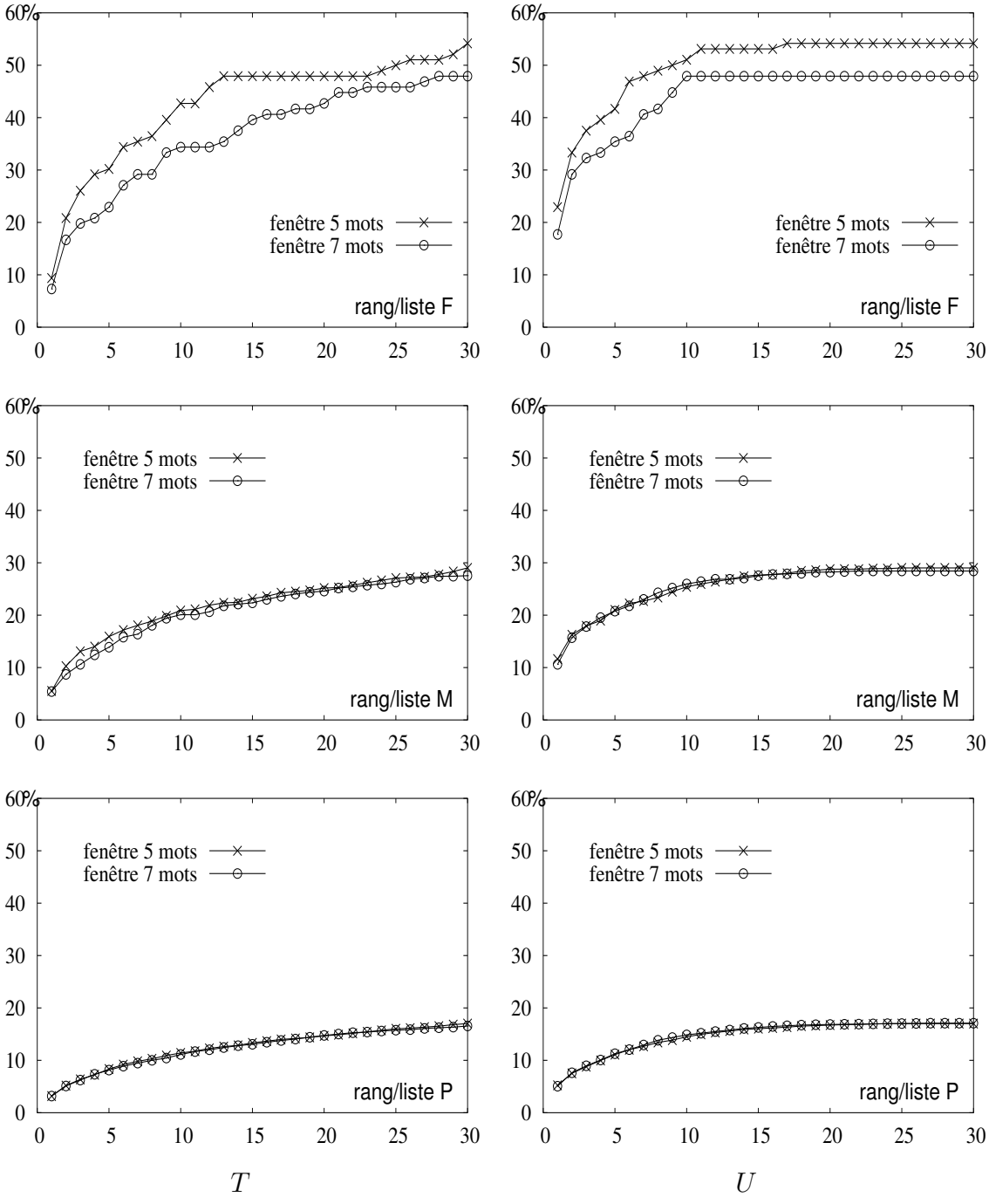
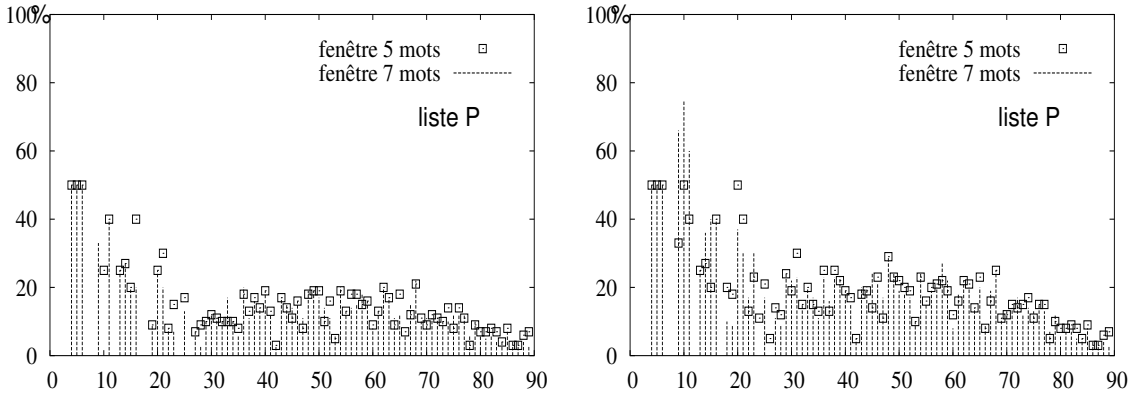


FIG. 4.5 – *Comparaison des performances entre deux tailles de fenêtre de contexte (5 et 7 mots) pour les trois lexiques de test : les mots fréquents F, les mots médicaux M et les mots pivots P.*

Les figures 4.6T et 4.6U illustrent les performances sur les 10 meilleurs candidats par tranche de fréquences homogènes (section 4.3.3). Nous pouvons constater que la diffé-



T U

FIG. 4.6 – *Comparaison des performances sur les 10 premiers candidats pour deux tailles de fenêtre de contexte, en fonction de la distribution des fréquences des mots.*

rence de performance n'est pas localisée de façon homogène sur les premières tranches de fréquence. La performance sur les 10 meilleurs candidats à la traduction, pour les mots les plus fréquents (les 10 premières tranches de la distribution), est similaire pour les deux longueurs de fenêtre examinées ici : 5 et 7 mots. Dans le cas où l'on cherche la traduction parmi les mots inconnus (figure 4.6U), la performance sur certains mots fréquents (tranches 9, 10, 11) obtenue avec la fenêtre de 7 mots dépasse celle obtenue avec la fenêtre de 5 mots. Cette différence s'estompe lorsque l'on considère les résultats de tous les mots du lexique (liste P).

En examinant de près les vecteurs de contextes des mots fréquents de la liste F, on s'aperçoit que dans certains cas, la fenêtre de 5 mots permet de mettre en valeur des relations de type collocationnel, par rapport à la fenêtre de 7 mots qui introduit d'autres types de relations. Prenons l'exemple des vecteurs traduits du mot *métastatique* et des vecteurs originaux de sa traduction *metastatic*, qui sont tous les deux fréquents dans le corpus C23 ($freq(métastatique) = 403$, $R_{distribution} = 40$; $freq(metastatic) = 117$, $R_{distribution} = 35$). Pour la fenêtre de 5 mots, nous remarquons que davantage de cooccurrences significatives (*cancer, tumor, melanoma, bone, neuroblastoma*, etc.) font partie des entrées du métathésaurus UMLS (respectivement *metastatic cancer, metastatic tumor, metastatic melanoma, metastatic tumor of bone, metastatic neuroblastoma*). Pour la fenêtre de 7 mots, les cooccurrences parmi celles ayant plus de poids sont *pain, cancer, tumor, patient, chemotherapy*, etc.

De même, pour les vecteurs du mot *sommeil* et des vecteurs originaux de sa traduction *sleep* ($freq(sommeil) = 590$, $R_{distribution} = 36$; $freq(sleep) = 190$,

$R_{distribution} = 30$), nous trouvons dans la fenêtre de 5 mots, des mots ayant un poids plus important *disorder, apnea, syndrome, obstructive*, etc. et qui figurent dans les entrées UMLS *sleep disorder, sleep apnea syndrome, obstructive sleep apnea*. Dans la fenêtre de 7 mots, des mots comme *society, activity*, etc., ont un poids plus important que les mots cités précédemment.

Variation du type des mots de contexte : spécialisé ou spécialisé+général

Cette expérience a pour objectif de mesurer l'impact des ressources bilingues partielles utilisées pour comparer les vecteurs de contexte. Il s'agit en particulier de déterminer si les mots généraux sont des mots de contexte qui permette de mieux rapprocher un mot et sa traduction. Pour cela, à partir du corps C23 deux listes de mots ont été établies pour la traduction des vecteurs : une liste spécialisée ($P\prime$), composée de 4963 entrées provenant du lexique médical (section 4.2.2) et une liste complète (P) des mots pivots, composée des 6 210 entrées issues de la combinaison du lexique médical et général.

Les expériences réalisées portent sur la liste réduite des mots fréquents du corpus C23 (liste F). Rappelons que les mots présents dans cette liste peuvent être des mots généraux ou des mots spécialisés. Il n'est en effet pas envisageable de mener ces expériences sur des mots peu fréquents car ils ont un vecteur de contexte généralement pauvre. Or le fait d'enlever les mots généraux appauvrit encore plus les contextes à étudier.

Pour chaque mot de la liste F, deux vecteurs de contexte ont été construits et traduits à l'aide des deux listes décrites ci-dessus. Pour comparer les vecteurs, la méthode classique a été utilisée.

Résultats de l'évaluation de l'effet du type des mots de contexte

Les résultats sont illustrés par les figures 4.7P$\prime$ et 4.7U$\prime$. Pour des raisons de visibilité, nous présentons les résultats obtenus en utilisant *Jaccard* et *tf.idf* comme mesure de similarité et pondération[8].

Les listes de candidats utilisées dans ces expériences sont des listes réduites[9] : liste $P\prime$ pour les mots spécialisés connus et liste $U\prime$ pour les mots inconnus fréquents. De plus, seuls les 20 premiers rangs sont affichés.

Dans le cas de l'extraction des traductions parmi tous les mots fréquents spécialisés du corpus anglais (figure 4.7P'), l'écart entre les deux types de contexte est net et parle en faveur de la prise en compte des mots généraux. Lorsque les mots généraux sont inclus dans les vecteurs de contexte (courbe Contexte P), toutes les traductions

8. D'autres mesures de similarité et pondérations ont été évaluées dans (Chiao & Zweigenbaum, 2003). Les mauvais résultats de ces mesures nous ont incités à ne pas les présenter ici.

9. Cela explique les très bons résultats obtenus par rapport aux résultats moins bons obtenus quand les listes des candidats sont des listes complètes.

sont correctement extraites parmi les 19 meilleurs candidats. Seuls 59,4% des traductions sont extraites parmi les 20 meilleurs candidats lorsque l'on n'utilise que les mots spécialisés pour le transfert de vecteurs de contexte.

Observons maintenant la tendance des courbes concernant l'extraction des traductions parmi les candidats inconnus fréquents (figure 4.7U′). Nous remarquons que l'écart de performance entre les deux types de contextes est moins important, surtout si on considère les résultats uniquement sur les 5 premiers rangs. Environ 45% des mots trouvent ainsi leur traduction avec un vecteur de contexte plus large (6 210 mots), et 42% avec un vecteur de contexte plus petit (4 963 mots) aux 5 premiers rangs. 97% des traductions correctes sont proposées dans les 20 premières positions avec le contexte général P alors que 93% le sont avec le contexte spécialisé P'.

La prise en compte des mots généraux dans les contextes d'un mot améliore globalement la performance, quelque soit la nature des candidats à la traduction (les mots de spécialité ou les mots inconnus fréquents). La progression nette des scores concernant la liste des candidats spécialisés connus (figure 4.7M′) révèle l'effet positif d'ajouter les mots généraux dans le vecteur de contexte. Dans ce cas où l'on cherche la bonne traduction parmi les mots du lexique, l'utilisation d'un vecteur de taille plus grande permet de mieux représenter les mots à traduire ainsi que leurs traductions, ce qui améliore les résultats.

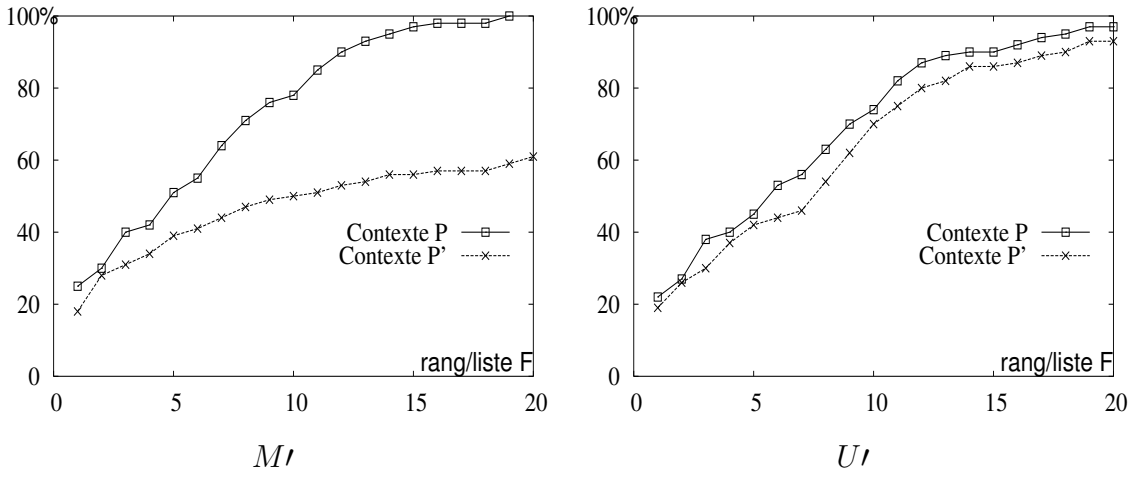


FIG. 4.7 – *Comparaison des résultats d'extraction entre types de contexte spécialisé P' et spécialisé+général P, avec la liste des candidats à la traduction des mots du lexique médical (liste M') et celle des mots inconnus fréquents (liste U').*

4.3.7 Paramètres de la comparaison des vecteurs de contexte

Les expériences menées ici se proposent d'évaluer l'impact des paramètres qui interviennent dans la comparaison des vecteurs de contexte. Il s'agit en particulier de comparer différents systèmes de pondération et différentes mesures de similarité.

Résultats de l'évaluation des différents systèmes de pondération

Dans un premier temps, nous nous sommes intéressées aux différents systèmes de pondération que nous avons présentés précédemment (section 3.2.2) : information mutuelle (MI), *tf.idf*, *loglike*. Les résultats obtenus sans pondération, c'est-à-dire calculés à partir des cooccurrences simples sont désignés par *cooc*.

Afin d'optimiser le calcul des différentes pondérations, la liste des 96 mots fréquents décrite dans la section 4.3.6 a été utilisée comme lexique de test. Les différentes mesures de pondération ont été appliquées au sous-corpus C23. Pour chaque mot français du lexique de test, son vecteur de contexte pondéré par différentes mesures a été ensuite comparé à l'aide d'une similarité avec les vecteurs des candidats à la traduction provenant des deux listes ($P\prime$ et $U\prime$) décrites dans la section 4.3.2.

Rappelons que le nombre des candidats à apparier correspond à 4 963 mots pour la liste $P\prime$ contenant les mots spécialisés connus et à 6 018 mots pour la liste $U\prime$ constituée des mots inconnus fréquents plus la traduction attendue. Les résultats sont calculés sur la base de deux mesures de la similarité : cosinus et *Jaccard*, en utilisant une fenêtre de 7 mots pour calculer les cooccurrences.

La présentation des résultats est en fonction du rang de la traduction attendue parmi les 20 premiers candidats proposés pour chaque mot de la liste de test. Les figures 4.8P$\prime$ correspondent aux résultats de la liste $P\prime$ et les figures 4.8U$\prime$ à ceux de la liste $U\prime$.

Nous observons que la courbe de *tf.idf* domine légèrement lorsqu'il s'agit d'extraire la traduction parmi les mots du lexique médical (liste $P\prime$) quelque soit la mesure de similarité utilisée (rang/cosinus, rang/jaccard). Avec le cosinus par exemple, 59,4% des mots fréquents ont leur traduction attendue classée parmi les 20 premières positions. On remarque qu'avec le cosinus, les courbes *MI* et *cooc* sont confondues et la courbe *loglike* présente une évolution parallèle aux trois autres pondérations, en restant nettement en deçà. Avec la mesure *Jaccard*, cet écart est moindre mais reste significatif. Si on cherche la traduction parmi les 10 premiers candidats, 50% des mots sont traduits avec *tf.idf* contre 42,7% avec *loglike*.

Pour les trois premiers rangs, les résultats obtenus avec la liste des candidats inconnus fréquents $U\prime$ n'affichent pas de différence entre les différentes pondérations que ce soit avec le cosinus ou la mesure *Jaccard* sauf pour *cooc* qui est meilleur. A partir du quatrième rang, nous constatons ici aussi un écart important des résultats obtenus avec *loglike* et les autres pondérations. Avec le cosinus, les courbes *MI* et *cooc* atteignent 63,8% de performance parmi les 20 premiers candidats proposés, la courbe *tf.idf* atteint 60,6%, tandis que la courbe *loglike* reste à 45,7%. Avec la mesure

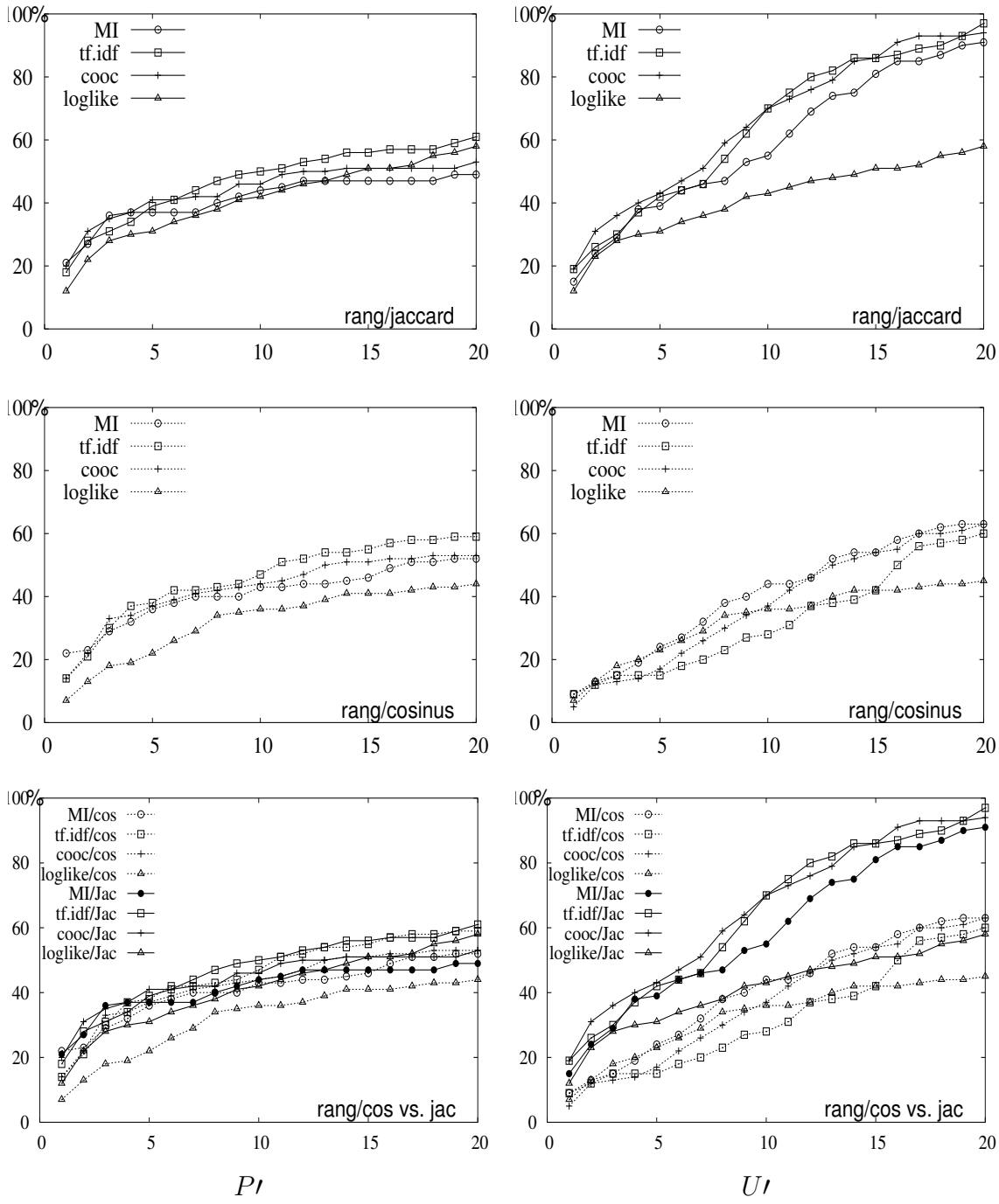


FIG. 4.8 – *Comparaison des performances entre les différentes combinaisons de mesures de similarité et de systèmes de pondération, avec la liste des candidats à la traduction des mots du lexique médical (liste P') et celle des mots inconnus fréquents (listeU').*

Jaccard, presque tous les traductions ont été extraites parmi les 20 premiers candidats en utilisant *tf.idf*, *MI* ou sans pondération *cooc*, contrairement à la pondération *loglike* qui donne un taux de 58,5%.

En résumé, parmi les trois systèmes de pondération traités, *tf.idf* se détache plus ou moins nettement que ce soit en faisant varier les mesures de similarités ou le type de candidats visés. Notons tout de même que les résultats sans pondération *cooc* sont très proches (voire meilleurs sur certains premiers rangs). Ces résultats tendent à montrer que finalement la présence conjointe et fréquente de deux mots, c'est-à-dire la cooccurrence est le paramètre le plus important à considérer. Les autres paramètres notamment présents dans les autres systèmes de pondération sont secondaires.

Résultats de l'évaluation des différentes mesures de similarité

Nous comparons maintenant les résultats des mesures de similarité cosinus et *Jaccard* en superposant les courbes correspondantes (rang/cos vs. jac) dans les figures 4.8P/ et 4.8U/.

Lorsqu'on cherche la traduction des mots du lexique spécialisé (figure 4.8P/, rang/cos vs. jac), nous remarquons à première vue que les résultats des deux mesures de similarité ne diffèrent pas beaucoup. Seule la courbe loglike/cos est nettement en dessous des autres courbes. Si nous examinons de près chaque combinaison de pondération avec les mesures de similarité, c'est *Jaccard* qui l'emporte sur le cosinus notamment pour les quatre premiers rangs. Avec la combinaison tf.idf/Jac par exemple, 18,8% des mots sont correctement traduits par rapport au 14,6% avec la combinaison tf.idf/cos. De même pour cooc/Jac, loglike/Jac qui atteignent 12,5% et 20,8% respectivement tandis que les courbes correspondantes du cosinus restent à 7,29% et 14,6% respectivement.

Dans le cas où l'on cherche à extraire les traductions parmi les mots inconnus fréquents, la performance de *Jaccard* l'emporte lorsqu'il est combiné avec n'importe laquelle des pondérations (*MI*, *tf.idf* et *loglike*) ainsi que pour la combinaison avec les seules cooccurrences *cooc*. C'est particulièrement net avec *tf.idf*. *Jaccard* produit un score de 94% en considérant les 20 premiers candidats

En résumé, la mesure de similarité *Jaccard* l'emporte sur le cosinus que ce soit en faisant varier les pondérations ou les types de mots visés.

4.3.8 Similarité classique vs similarité croisée

Dans toutes les méthodes classiques d'extraction de lexique bilingue à partir de corpus comparables, les processus d'extraction s'appliquent dans une seule direction de langues, c'est à dire de la langue source vers la langue cible (langue source → langue cible).

Notre modèle repose sur une mesure de similarité appliquée aux deux directions de langues (langue source ↔ langue cible). Il s'agit de la similarité croisée décrite

au chapitre 3. L'objectif est ici de déterminer l'impact de la prise en compte de la similarité croisée sur les résultats d'extraction des traductions. Nous évaluons donc l'approche classique, c'est à dire utilisant une mesure de similarité dans une seule direction de langues, et notre approche et comparons les résultats sur les deux corpus comparables (C23 et TOUT).

Différents paramètres ont été fixés en fonctions des résultats précédents. Les mots à traduire sont issus de trois lexiques de test (F pour les mots fréquents, M pour les mots spécialisés et P pour les mots généraux/spécialisés). Les mots candidats à la traduction sont issus de deux listes : la liste des mots inconnus U et la liste complète T. La taille de la fenêtre de cooccurrence est fixée à 7 mots. Le système de pondération utilisé est *tf.idf*. La mesure de similarité *Jaccard* est utilisée dans le modèle dit classique ainsi que dans notre modèle dans les deux directions de langues.

Des exemples de résultats d'extraction sont donnés au tableau 4.5 pour la méthode classique et au tableau 4.6 pour la similarité croisée. On y remarque par exemple

Mot Fr	**Traduction**	**R**	**5** meilleurs candidats à la traduction selon Jaccard
nécrose	necrosis	1	**necrosis .181**, chronic .148, renal .142, inflammation .135, infarction .123
gène	gene	1	**gene .247**, mutation .243, recessive .197, protein .194, chromosome .145
sclérose	sclerosis	3	sep .263, lateral .187, **sclerosis .178**, passe .177, poliovirus .158
abcès	abscess	9	perforation .227, rupture .192, visible .156, invasive .155, impose .149

TAB. 4.5 – *Exemple d'extraction avant application de la similarité croisée ; R = rang de la traduction attendue dans la liste des candidats proposés.*

Mot Fr	**Traduction**	**R**	**5** meilleurs candidats à la traduction proposés selon la similarité croisée
nécrose	necrosis	1	**necrosis 1**, chronic 3.74, renal 4.2, inflammation 5.53, infraction 10
gène	gene	1	**gene 1**, mutation 1.33, protein 2.66, chromosome 2.85, recessive 3
sclérose	sclerosis	1	**sclerosis 1.5**, sep 2, lateral 2.66, passe 8, poliovirus 10
abcès	abscess	1	**abscess 1.8**, perforation 2, rupture 4, visible 6, invasive 8, impose 10

TAB. 4.6 – *Exemple d'extraction après application de la similarité croisée ; R = nouveau rang de la traduction attendue dans la liste des candidats proposés.*

que pour le mot *sclérose*, la méthode classique propose la traduction attendue *sclerosis* comme 3ème meilleur candidat (R = 3) derrière *sep* et *lateral*, tandis que la méthode basée sur la similarité croisée propose la bonne traduction en première position (R = 1). De même, le mot *abcès* dont la traduction attendue *abscess* est classée au 9ème rang avec la méthode classique et avancé au premier rang après l'application de la similarité croisée.

Le tableau 4.7 illustre la comparaison en nombre de traductions correctes extraites par la méthode basée sur la similarité croisée (notée par la suite *SC*) et la mé-

thode classique (notée *sans SC*). Nous présentons pour chaque liste de test (P, M, et F) provenant des deux corpus TOUT et C23, le nombre de bonnes traductions trouvées par les deux méthodes, en première position (R_1) et parmi les 30 premières positions (R_{30}) sur la liste des candidats inconnus (U).

Méthode	**Nb traductions en R_1**			**Nb traductions en R_{30}**		
Corpus TOUT	liste P	liste M	liste F	liste P	liste M	liste F
Méthode classique	679	181	24	1 854	402	43
Similarité croisée	**804**	**226**	**30**	**1 978**	**335**	**50**
Corpus C23						
Méthode classique	309	93	17	1 028	244	46
Similarité croisée	**367**	**112**	**20**	**1 065**	**253**	**46**

TAB. 4.7 – *Résultats de l'extraction de lexique bilingue aux rangs 1 et 30 par la méthode classique et la similarité croisée.*

Par rapport à la méthode classique, l'approche fondée sur la similarité croisée permet de proposer plus de traductions correctes au rang 1 (candidats classés en R_1) quelque soit le lexique de test utilisé (liste F pour les mots fréquents, M pour les mots spécialisés, ou P pour tous les types). Nous avons mentionné précédemment (section 4.3.2) que l'avantage d'employer la liste U comme candidats à la traduction est de favoriser la traduction de nouveaux mots, ***i.e.***, des mots inconnus. Dans le cadre de l'actualisation de lexique, l'amélioration de la performance d'extraction apportée par l'application de la similarité croisée permet d'ajouter de nouvelles entrées dans le lexique existant.

Examinons maintenant les résultats complets concernant les 30 meilleurs candidats proposés par les deux méthodes à partir des deux listes de candidats T et U pour les deux corpus : figures 4.9T et 4.9U pour le corpus TOUT, figures 4.10T et 4.10U pour le corpus C23.

On constate que l'effet de la similarité croisée se manifeste surtout au premier rang de la liste ordonnée des candidats à la traduction. Elle permet d'extraire plus de traductions correctes par rapport à la méthode classique, quelque soit la taille du corpus traité. L'évolution globale des courbes SC sur les 30 premiers rangs, reste meilleure pour les trois listes des mots (F pour les mots fréquents, M pour les mots médicaux et P pour tous les mots pivots)[10]. Cet écart est plus important lorsque la taille de corpus augmente (figures 4.9).

On voit ici l'effet de la similarité croisée. Si la traduction attendue pour un mot donné est mal classée par la similarité classique orientée dans une direction de langues ($A \rightarrow B$), la prise en considération des rangs issus de la similarité dans l'autre direction ($B \rightarrow A$) donne une possibilité de plus pour l'appariement. La méthode met en valeur

10. La seule exception concerne le résultat obtenu avec la liste de candidats T sur la liste F des mots fréquents pour le corpus C23 et uniquement à partir du rang 21.

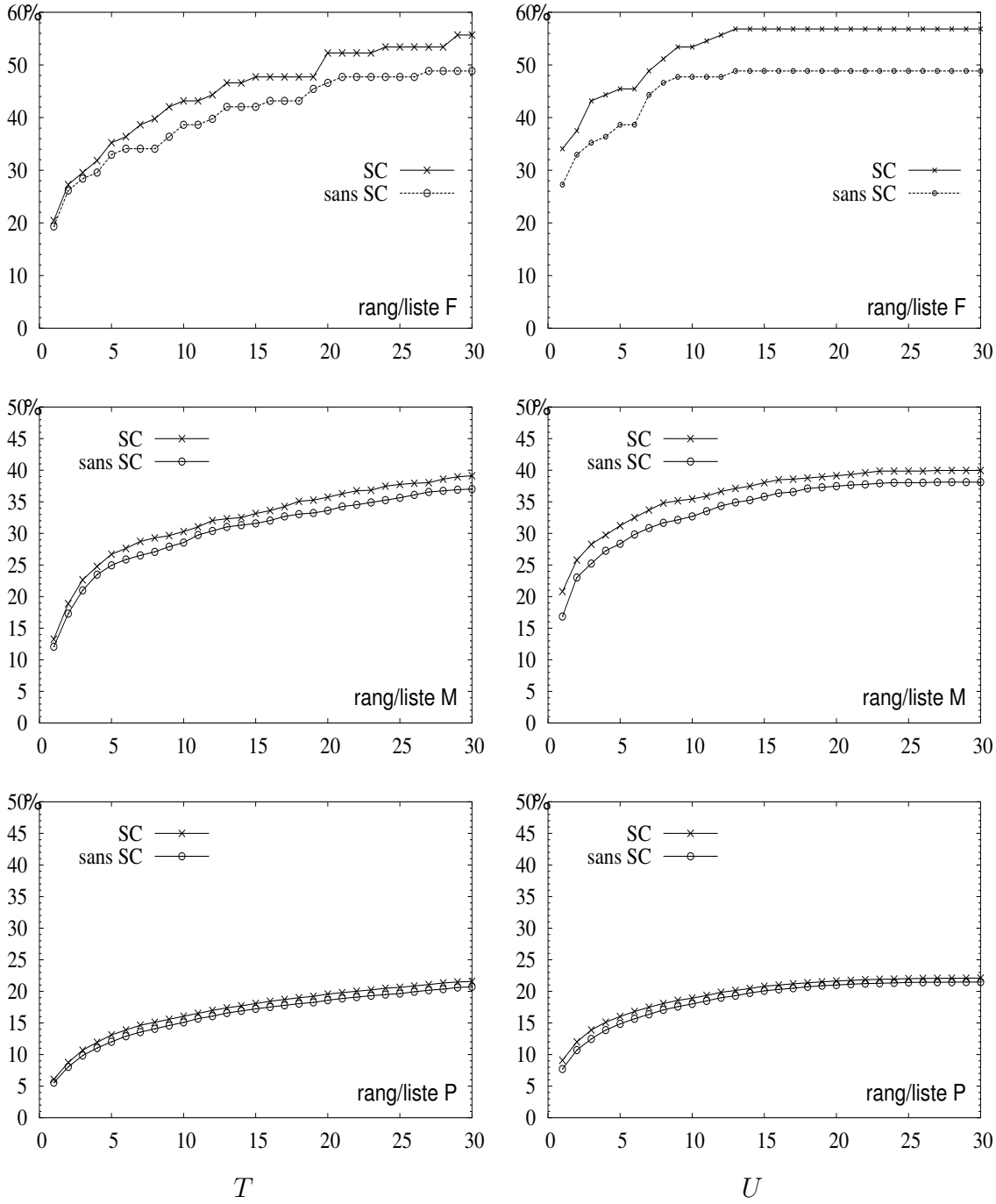


FIG. 4.9 – *Comparaison des performances avec et sans application de la similarité croisée (SC et sans SC) pour le corpus TOUT.*

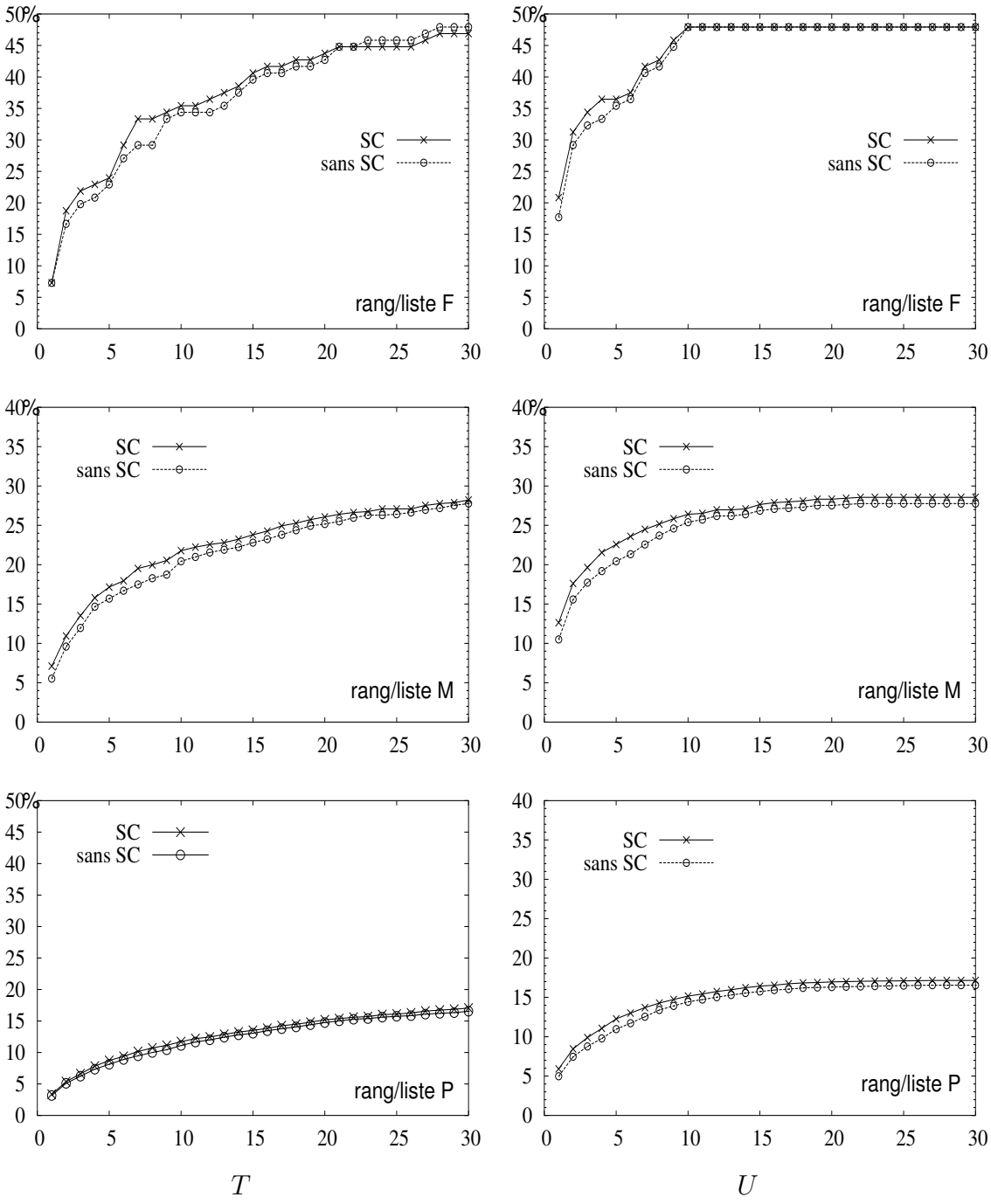


FIG. 4.10 – *Comparaison des performances avec et sans application de la similarité croisée (SC et sans SC) pour le corpus C23.*

les couples de mots ayant une similarité (relative) élevée dans au moins une direction de langues.

Lorsque l'on compare les résultats de l'application de la similarité classique dans chacune des directions de langues (français → anglais et anglais → français), les meilleures performances sont observées pour la direction du français vers l'anglais. Dans le cas du lexique des mots pivots P, l'application de la similarité de l'anglais vers le français propose dès le premier candidat la traduction correcte pour 461 mots. Parmi les 30 premiers candidats, 1 259 mots ont été bien traduits. Du français ver l'anglais, 679 mots français trouvent leur bonne traduction au rang 1, alors que le nombre de mots atteint 1 854 si les 30 premiers candidats sont retenus. La situation est identique pour le lexique médical M, 146 mots anglais contre 181 mots français trouvent leur traduction correcte au rang 1 et 339 mots anglais contre 402 mots français au rang 30. Cette meilleur performance est probablement liée à la différence de taille entre le corpus CISMeF et le corpus CliniWeb. Le fait que le volume du corpus CISMeF est presque 3 fois plus important que celui du CliniWeb élargit l'espace de recherche. Dans le cas de nos expériences, chercher la traduction d'un mot anglais revient à la chercher parmi les 280 857 mots du corpus CISMeF. Tandis que chercher la traduction d'un mot français revient à la chercher parmi les 70 000 mots du corpus CliniWeb.

Il est intéressant de noter que la différence des deux courbes lorsqu'il s'agit d'extraire les traductions à partir de tous les mots du corpus (figures 4.9T et 4.10T) croît en fonction du rang, notamment sur la liste des mots fréquents (rang/liste F). L'impact de la similarité croisée est visible à partir du deuxième ou troisième rang. Par exemple, avec la similarité classique, *depression*, traduction de *dépression*, est classé derrière *anxiety, stress, confusion, mental, agitation*. Avec l'application de la similarité croisée, *depression* est classé au deuxième rang, derrière *disorder*. De même pour *épilepsie* dont la traduction *epilepsy* est classée au rang 6 derrière *myoclonic, myoclonus, aplastic, epm, progressive*, tandis qu'avec le calcul de similarité croisée, la bonne traduction est classée au rang 3 derrière *myoclonique, myoclonus*.

4.4 Conclusion

Nous avons proposé une extension du modèle classique d'extraction de lexique en corpus comparables, en introduisant une nouvelle mesure : la 'similarité croisée' qui repose sur l'hypothèse de la 'symétrie distributionnelle' entre les langues (chapitre 3). Les expériences menées sur les deux corpus de tailles différentes ont montré que l'approche fondée sur la similarité croisée améliore la performance (pourcentage de mots correctement traduits) par rapport à l'approche classique. Il est intéressant de noter que cette amélioration est plus importante (figures 4.9U et 4.10U) avec l'utilisation de la liste U des candidats à la traduction inconnus dont l'objectif est de simuler la traduction des mots inconnus donc l'actualisation de lexique.

En dehors de la similarité croisée, plusieurs paramètres ont été également évalués

au cours de ces expériences. La taille de corpus est considéré comme un critère décisif de la qualité des modèles d'extraction de lexique à partir des corpus comparables (Rapp, 1999; Fung, 1998; Déjean & Gaussier, 2002). En général, un corpus de taille plus importante permet de construire des vecteurs de contexte plus représentatifs. Les résultats de nos expériences ont confirmé cet effet positif de la taille de corpus, notamment sur l'extraction d'un lexique spécialisé médical (résultats obtenus pour la liste du lexique spécialisé M).

L'ajout de mots généraux dans le vecteur de contexte a également un impact sur les performances. Il améliore l'appariement des vecteurs de contexte. Cependant, une couverture partielle et une qualité pauvre du lexique général peut être problématique dans le cas où il introduit du bruit dans les vecteurs de contexte. Prenons l'exemple du mot *trouble*, dont la traduction proposée par le lexique général est *indistinct*. Or dans le domaine médical, on s'attend plutôt au mot *disorder* comme équivalent en anglais. De même pour le mot *échelle*, la traduction proposée est *ladder* au lieu de *scale*. Même dans le cas d'une qualité optimale du lexique général, la levée des ambiguïtés de traduction reste un problème entier.

En examinant les résultats erronés, selon leur distribution de fréquence dans le corpus utilisé, nous constatons que les mots mal traduits ayant une fréquence élevée sont en général des mots généraux comme : *système, taux, zone, date, besoin, niveau, base, type, accès, image, rare*, etc. Leur contextes présentent une forte diversité et leur traduction s'avère souvent ambigue. Ainsi notre modèle basé sur la similarité des distributions rencontre ses limites à cause de contextes trop abondants. Dans le cas des mots mal traduits peu fréquents, la plupart sont des mots spécifiques, *e.g.*, *pentose, quinacrine, yttrium, squameux, polyribosome, anus, thrombocyte, thoraco-abdominal, plasma*, etc. Le manque de contextes et la différence de taille entre les corpus anglais et français peuvent expliquer les erreurs de traduction. Le fait que le vecteur anglais de ces mots ne contient que très peu de mots du lexique rend l'appariement plus difficile avec le vecteur français traduit. Certains mots forment souvent avec un terme 'productif' (Bourigault, 1994) un terme composé : *squameux/carcinome squameux, achille/tendon d'achille, anus/cancer d'anus, pentose/pentose phosphate*. Dans ce cas, le manque de contextes discriminants rend aussi difficile l'appariement des vecteurs.

Lorsqu'on examine les premiers candidats proposés par notre modèle, nous nous apercevons que beaucoup d'entre eux sont sémantiquement proches, sont des dérivations grammaticales ou entretiennent des relations du type collocationnel. Le tableau 4.8 montre quelques exemples extraits de la liste des candidats proposés.

Dans le cas des résultats concernant les dérivations adjectivales en français, nous constatons que certains d'entre eux n'ont pas de dérivation équivalente en anglais (*ovalaire, hyperalgique, pulpaire, villositaire, abdomino-lombaire, pubertaire, fibroglandulaire, cardiocirculatoire*, etc.), ou que l'adjectif anglais existe mais est absent du corpus (*plasmatic, hematic*, etc.). Notre modèle ne peut alors pas proposer ces mots comme traduction puisqu'il est impossible de construire leurs vecteurs de contexte.

L'analyse morphosyntaxique du corpus permettrait de mieux prendre en compte

Mot FR/EN	Extraits des premiers candidats à la traduction
carence/deficiency	deficiency, vitamin, folate, iron, low, zinc
acide/acid	amino, protein, acid, fatty, deficiency, folic, iron, zinc, substance
plasmatique/plasmatic	plasma, serum, high, low, increased, lithium
hépatite/hepatitis	hepatitis, infection, virus, invasive,immunodeficiency
aorte/aorta	aortic, carotid, aneurysm, aorta, artery, coronary, left
angoisse/anxiety	anxiety, depression, panic, attack, agitation, psychose, agitation

TAB. 4.8 – *Exemples des premiers candidats à la traduction proposés.*

ce phénomène en exploitant les informations de morphologie dérivationnelle pour l'extraction de lexique. Par exemple, notre modèle propose le nom *plasma* comme traduction de l'adjectif *plasmatique*. Cependant, cette proposition n'est pas considérée comme valide puisque dans le lexique de test, la traduction est l'adjectif *plasmatic*.

Les résultats obtenus par l'exploitation de la similarité croisée s'avèrent encourageants et confirment la validité de l'hypothèse de symétrie distributionnelle tout au moins pour certains mots. Notons également l'aspect générique du modèle proposé. Aucun traitement linguistique (à l'exception de la segmentation et de la mise au singulier) n'est utilisé. Les seules ressources lexicales indispensables sont une liste de mots vides et un lexique bilingue partiel spécialisé et général. Cela favorise l'implémentation et aussi la portabilité vers de nouvelles paires de langues.

Chapitre 5

Recherche d'information translangue

5.1 WWW et multilinguisme

Le développement incessant et explosif du Web fournit une masse d'informations provenant de multiples sources. Il existe néanmoins de nombreux obstacles qui entravent l'accès à l'information utile. Le multilinguisme est l'un d'entre eux.

Actuellement sur l'Internet, l'anglais est toujours la langue la plus utilisée. Pourtant elle connaît une décroissance relative assez importante en faveur des autres langues[1][2][3]. Une analyse publiée en septembre 2003[4] estime que près des deux tiers de la population internaute provient d'un pays non anglophone. Plus précisément 262,3 millions (35,6% de la population internaute) sont anglophones et 474,3 millions (64,4%) sont non anglophones[5] (tableau 5.1). L'accroissement de la diversité linguistique des utilisateurs du Web a pour conséquence la progression du multilinguisme sur le Web, tendance qui constitue un enjeu important autant commercial, technique que culturel.

C'est également une difficulté pour les internautes, les communautés spécialisées mais aussi pour les entreprises et notamment les groupes internationaux. Une meilleure connaissance de l'autre (partenaires, concurrents, réglementations locales) favorise en effet pour un groupe son implantation dans un pays étranger, la réponse à un appel d'offres, la réactivité face à la concurrence, etc. De plus, un des défis pour un groupe international est de ne pas rater la 'bonne information' même si elle est diffusée en langue étrangère. Or une grande part de l'information nécessaire à cette veille concurrentielle et économique est aujourd'hui sur le Web et souvent écrite en langue locale. L'accès à ces informations est donc entravé par les barrières linguistiques. Pour les communautés spécialisées, l'enjeu et la difficulté sont du même ordre. Le Web décloi-

1. http://www.unesco.org/webworld/points_of_views/300102_pimienta.shtml
2. http://www.nua.net/surveys/
3. http://www.worldlingo.com/resources/language_statistics.html
4. http://glreach.com/eng/ed/gre/index.php3
5. Les chiffres ont été calculés en fonction de nombre de personnes connectées sur l'Internet dans chaque zone linguistique en 2003. http://global-reach.biz/globstats/index.php3

Anglophone	262,3 M	35,6%
Non-Anglophone	474,3 M	64,4%
Langues européennes (sauf Anglais)	257,4 M	34,9%
Espagnol	58,8 M	8,0%
Allemand	51,6 M	7,0%
Français	27,2 M	3,7%
Italien	24,2 M	3,3%
Portugais	19,4 M	2,6%
Langues Scandinaves	14,3 M	1,9%
Flamand	13,5 M	1,8%
Autres	48,4 M	6,6%
Langues asiatiques	216,9 M	29,4%
Chinois	90,0 M	12,2%
Japonais	69,7 M	9,5%
Coréen	29,2 M	4,0%
Arabe	8,7 M	1,2%
Autres	19,3 M	2,5%

TAB. 5.1 – *Estimation du nombre d'internautes en 2003 (en million d'individus).*

sonne l'information et la rend disponible mais les langues la rend inaccessible.

L'émergence du Web comme principale source d'information pour les entreprises, les internautes et les communautés spécialisées, associée à l'émergence des langues locales sur le Web exacerbent la problématique du multilinguisme et lui confèrent un enjeu important. Cette problématique prend la forme d'un champ d'étude : la recherche d'information translangue (*CLIR - Cross-Language Information Retrieval*).

5.1.1 Recherche d'information translangue et barrière linguistique

La recherche d'information (*IR - Information Retrieval*), par exemple lors de l'usage d'un moteur de recherche sur Internet, a pour objectif de rapporter les documents correspondant à une requête formulée par un utilisateur. Ces documents sont en général sélectionnés par le fait qu'il contiennent les mots de la requête. Un système de recherche indexe [6] un document à partir des mots de la langue du document en question.

Du processus d'indexation résulte le fait que la recherche d'information ne pourra s'effectuer correctement sur les documents que si la requête est également formulée dans la langue des documents. Une requête en français ramène donc normalement des documents en français, et une requête en anglais des documents en anglais. La formulation initiale de la requête limite ainsi a priori le champ des réponses possibles, privant

6. L'indexation est un procédé qui consiste à associer à un document les concepts et les termes qui représentent le mieux le contenu du document dans le but de pouvoir le situer par rapport à un ensemble. Cela vise à accélérer la recherche.

l'utilisateur connaissant plusieurs langues de documents potentiellement intéressants. Cela induit une importante limitation dans l'exploitation de ressources d'information grand public ou spécialisés telles que le Web.

5.1.2 Recherche d'information translangue : un enjeu important

La recherche d'information translangue vise à sortir de ce dilemme. Son objectif est de permettre à l'utilisateur l'accès à des documents rédigés dans des langues différentes à partir de requêtes exprimées dans sa langue maternelle (par exemple rapporter des documents en français et anglais à partir d'une requête formulée en français).

En 1970, Salton (Salton, 1970) aborde en premier cette problématique et montre des résultats expérimentaux. Son expérience repose sur l'utilisation d'un thésaurus bilingue anglais-allemand et sur une liste de requêtes en anglais et leurs traductions en allemand, construite manuellement. La base documentaire de recherche est une collection de 468 documents en anglais et en allemand interrogée avec le système de recherche SMART[7].

Nous nous intéressons ici à un aspect primordial de la recherche d'information translangue, le traitement des requêtes, et notamment la traduction et l'extension de requête. Ce chapitre décrit les principales techniques utilisées dans la plupart des systèmes de recherche d'information translangue, en insistant sur la façon dont les processus de traduction et d'extension de requête contribuent à faciliter ou à améliorer les résultats de la recherche.

Dans un premier temps, les principaux modèles et fonctionnalités des systèmes de recherche d'information (section 5.2) sont présentés. Nous abordons ensuite la problématique et les méthodes utilisées en recherche d'information translangue (sections 5.3 et 5.4) . Nous nous attardons ensuite plus particulièrement sur un aspect important du processus de recherche d'information translangue : la reformulation de requête par extension (section 5.5). Il s'agit d'étendre la requête initiale avec des mots ou des termes afin de faciliter l'appariement avec les index des documents recherchés.

5.2 Systèmes de recherche d'information sur le Web

Les systèmes de recherche d'information permettent à l'utilisateur d'accéder aux documents répondant à un besoin d'information, exprimée en langue naturelle. En général, les moteurs de recherche identifient des sources d'information, les collectent, les indexent et en créent une base de données pour l'interroger. La recherche d'information sur Internet procède en quatre étapes indispensables :

– Collecte des documents.

7. SMART est un système de recherche d'information basé sur le modèle vectoriel implémenté par G. Salton.

– Indexation des documents collectés.

– Recherche des documents.

– Présentation des résultats.

5.2.1 Collecte des documents

Il existe deux modèles pour la collecte des documents : l'un manuel et l'autre automatique. Le premier apporte une valeur précise de validation des ressources mais nécessite du temps et coûte cher étant donné la diversité du contenu et des formats de documents. La méthode automatique apporte une régularité des résultats et une facilité de mise à jour des données.

Actuellement le Web présente un énorme potentiel de ressources documentaires sous format électronique, ce qui nécessite un traitement automatique. Dans le contexte multilingue, l'exploitation des sources d'informations disponibles sur le Web occupe une place de plus en plus importante dans les travaux consacrés à la recherche d'information translangue ou la traduction automatique (Chen & Nie, 2000; Resnik, 1999; Fung & Yee, 1998; Chen & Bian, 1998).

Les particularités de ces ressources provenant du Web soulèvent de nombreux problèmes à prendre en compte afin de pouvoir effectuer automatiquement une collecte 'efficace et intelligente'. Sachant qu'outre leur quantité importante, la structure des documents varie, il est nécessaire pour un outil de collecte automatique (*robots, crawler, spider...*) d'être capable d'effectuer les opérations qui suivent.

Reconnaissance automatique du codage

Il s'agit d'identifier le codage de caractères employé par différents types de ressources afin de permettre l'échange, le traitement et l'affichage des documents. La plupart de moteurs de recherche d'aujourd'hui utilisent la norme ISO-8859-1 (Latin-1) recouvrant les langues d'Europe de l'Ouest, ce qui limite l'accès aux documents en langues non latines (russe, japonais, chinois, arabe...). Afin de favoriser le multilinguisme, le consortium Unicode[8] a proposé le standard Unicode, norme de codage qui contient 95 221 caractères[9] et permet de représenter la plupart des langues vivantes.

Identification des langues

Un système de recherche d'information multilingue doit être capable d'identifier la langue des documents. En effet, les traitements utilisés en indexation impliquent souvent des techniques développées à partir des aspects linguistiques propres à une langue, par exemple la structure morphologique, syntaxique ou sémantique. Les outils

8. http://www.unicode.org
9. http://www.i18nguy.com/unicode/char-count.html

d'identification de la langue utilisent en général l'analyse morphosyntaxique (Zigler, 1991), les n-grammes (Cavnar & Trenkle, 1994), la présence de caractères spécifiques (Newman, 1987). Une démarche fréquente pour identifier une langue dans un contexte multilingue est également d'utiliser les mots vides propres à chaque langue (Grefenstette & Nioche, 2000).

Identification des formats

Le Web se caractérise par sa structure dynamique et son contenu hétérogène. Il contient essentiellement des documents HTML (HyperText Markup Language) (plus de 80% des pages Web), ASCII ('.txt') et des documents non textuels (Bray, 1996) en format PostScript ('.ps'), Acrobat ('.pdf'), images ('.gif' ou '.jpg'). De plus en plus des documents de type multimédia (audio, vidéo, image animée...) apparaissent sur Internet et les moteurs de recherche ne permettent pas de les indexer tous. Dans le cadre de notre projet, seules les données textuelles sont exploitées.

5.2.2 Indexation des documents

L'indexation est une étape primordiale dans le processus de recherche d'information car la qualité de l'index influence directement la pertinence des résultats de la recherche. Face à la masse des différentes bases documentaires numérisées, il faut pourvoir classer, gérer, comparer et consulter toutes ces données. Les approches classiques, souvent manuelles à l'aide d'un vocabulaire contrôlé (Fluhr, 2000) sont loin d'être efficaces. L'indexation automatique présente l'avantage d'être rapide et de permettre une couverture très large des ressources disponibles sur Internet. Cette approche automatique est désormais utilisée par la plupart des systèmes de recherche.

En fonction du principe du système de recherche utilisé, qu'il soit conceptuel ou lexical, l'indexation a pour but de trouver les concepts (indexation à l'aide d'un thésaurus) ou les termes qui représentent le mieux le contenu du document. Dans ce dernier cas, l'indexation se fait sur les mots simples ou sur les syntagmes car les mots simples ne donnent pas toujours une description précise. Dans le cas de la recherche d'information conceptuelle, l'indexation consiste à identifier dans le document des concepts appartenant à un thésaurus.

La structure d'index se présente sous forme d'un fichier inversé dans lequel chaque mot est mis en correspondance avec les documents dans lequel il apparaît. Le processus d'indexation est composé des traitements suivants.

Segmentation des documents en mots

Pour beaucoup de langues, la segmentation est relativement facile à réaliser parce que les mots sont séparés par un espace ou d'autres formes de séparateurs comme (',.?!', etc.). Pour certaines langues où il n'y a pas d'espace entre les mots, par exemple

le chinois ou le japonais, l'algorithme basé sur les séparateurs de mots ne peut pas fonctionner. L'utilisation de dictionnaire ou de lexique ou des algorithmes sophistiqués de segmentation en mots (Chien & Pu, 1996; Palmer & Burger, 1997) présentent des résultats assez pertinents.

Filtrage des mots vides

Afin d'augmenter la précision de la recherche et de réduire la taille de l'index, la plupart de systèmes n'indexent pas les mots vides (mots grammaticaux, mots fréquents, ou mots non discriminants). En effet, le fait de ne pas filtrer les mots vides provoque inévitablement du bruit. Les approches les plus utilisées pour filtrer les mots vides sont basées sur les occurrences des mots dans l'ensemble de la collection, ou sur une liste ou un dictionnaire de mots vides. L'indexation du texte complet considère que tous les mots sont significatifs et les indexe donc tous. Notons que le filtrage des mots vides doit se faire aussi bien au niveau de l'indexation que de celui de l'analyse de la requête.

Normalisation des mots

Les traitements les plus utilisés sont la 'lemmatisation' et la 'racinisation'. La première consiste à regrouper des mots selon les critères de morphologie flexionnelle. Généralement les formes conjuguées d'un verbe sont représentées par l'infinitif, et les variantes adjectivales par le masculin singulier. Pour obtenir de bonnes performances, il faut recourir à une analyse linguistique pour lever certaines ambiguïtés (Fluhr, 2000). La 'racinisation' consiste à traiter de plus les relations dérivationnelles entre les mots en leur associant une racine commune, ainsi *produire, production et produit* sont regroupés ensemble.

La normalisation des mots facilite les procédures de recherche, particulièrement dans le cas des langues européennes plus riches en morphologie flexionnelle (Peters & Sheridan, 2001; Savoy, 2002) mais aussi dans le cadre multilingue en exploitant les liens dérivationnels (Dal & Jacquemin, 1999). Par exemple, l'équivalent anglais de *carte routière* est *road map*. Pour le retrouver, il est nécessaire de faire le lien entre l'adjectif *routière* et le nom *road*.

Il est aussi fréquent, surtout dans le contexte multilingue, d'effectuer l'indexation sur les syntagmes (termes composés) afin de rendre la traduction plus précise et donc plus favorable à améliorer les résultats de recherche. Par exemple le terme *recherche d'information* constituera un seul index et sera traduit comme une unité plutôt que comme des mots séparés. L'identification des syntagmes se fait en général par l'utilisation d'un dictionnaire de termes composés ou d'un thésaurus, à l'aide des outils d'analyse syntaxique, ou par l'analyse statistique basée sur les cooccurrences.

Représentation des documents par des termes d'indexation

Selon les modèles de recherche (booléen, vectoriel ou probabiliste) que nous décrivons dans la section suivante, les documents et les requêtes peuvent avoir des représentations variées. Par exemple, le modèle vectoriel considère l'ensemble des documents comme un espace à n dimensions dans lequel chaque document est représenté par un vecteur à t dimensions non nulles où t est le nombre de termes d'indexation utilisés (Salton *et al.*, 1974). Dans le modèle booléen, les documents et les requêtes sont représentés par l'*ensemble* des termes d'indexation.

5.2.3 Modèles de la recherche d'information

La recherche documentaire vise à faire ressortir les documents pertinents pour une requête donnée. La notion de pertinence est très complexe et plusieurs facteurs y occupent une place importante : document, requête, mesure de pertinence, contexte et sujet de la recherche, utilisateur. D'une façon générale, un document est considéré comme pertinent lorsque l'utilisateur peut y trouver les informations dont il a besoin.

Si c'est l'indexation qui choisit les termes pour représenter le contenu du document, c'est au modèle de recherche de leur donner une interprétation. En effet, le modèle de recherche joue un rôle central au sein d'un système de recherche d'information. Il crée, pour un document ou pour une requête, une représentation interne basée sur les termes d'indexation (mots-clés) et applique ensuite une méthode d'appariement entre la représentation du document et celle de la requête afin de mesurer leur degré de correspondance.

Nous décrivons ici les trois modèles classiques de la recherche d'information (Baeza-Yates & Ribeiro-Neto, 1999).

Modèle booléen

Dans ce modèle, la pondération W_i d'un terme d'indexation T_i est définie par sa présence ou son absence dans le document, soit une valeur binaire : $W_i \in \{0, 1\}$. Un document est alors considéré comme une conjonction logique de termes d'indexation ($D = W_1 \wedge W_2 ... \wedge W_n$), et une requête est une expression booléenne conventionnelle comme par exemple $Q = W_i \vee (W_j \wedge \neg W_k)$. La correspondance entre un document et une requête est égale à 1 lorsque le document contient des termes correspondant à ceux de la requête en respectant les restrictions des différents opérateurs booléens (ET$\wedge$, OU$\vee$, NOT$\neg$). Dans le cas contraire la correspondance vaut 0. Le document est considéré comme non pertinent pour la requête.

Nous pouvons constater qu'un tel modèle ne permet pas l'appariement partiel pour lequel un document ne contient qu'une partie des termes d'indexation présents dans la requête. De plus, il n'est pas en mesure d'évaluer la pertinence d'un document par rapport à une requête car la correspondance issue du calcul booléen est évaluée par tout ou rien (1 ou 0). Il n'est pas possible d'obtenir un classement des documents selon

leur pertinence pour une requête donnée. Tous les termes du document et de la requête sont en effet pondérés de la même façon (0 ou 1). Il est donc difficile de déterminer si un terme est plus discriminant qu'un autre pour représenter le contenu du document ou de la requête.

Certaines méthodes ont été proposées comme le modèle booléen étendu (*P-norm model*) (Salton *et al.*, 1983) ou celles basées sur la théorie des ensembles flous (*Fuzzy information retrieval*) (Kraft & Buell, 1983; Ogawa *et al.*, 1991) afin de pallier à cette limitation faisant intervenir par exemple le poids des termes dans les documents ou la notion d'espace vectoriel pour représenter les documents et les requêtes.

Modèle vectoriel

Dans ce modèle, le document et la requête sont représentés par des vecteurs à n dimensions définies par l'ensemble des termes d'indexation. La pondération ici n'est plus une valeur binaire comme dans le modèle booléen. Soit un document $\overrightarrow{D_j} = (W_{1,j}, W_{2,j}...W_{n,j})$, et une requête $\overrightarrow{Q_m} = (W_{1,m}, W_{2,m}...W_{n,m})$. Chaque poids dans le vecteur désigne l'importance d'un terme d'indexation W_i dans le document ou la requête.

La pondération $W_{i,j}$ est calculée pour chacun des mots-clés T_i issus du document D_j afin de constituer une représentation vectorielle du document. Dans le cas où un terme d'indexation n'est pas présent dans le document, $W_{i,j} = 0$. Un document est ainsi représenté par un vecteur des mots-clés pondérés : $\overrightarrow{D_j} = (W_{1,j}, W_{2,j}...W_{n,j})$. Différentes pondérations peuvent être utilisées dans le modèle vectoriel (Salton & Buckley, 1988). Nous en décrivons quelques unes parmi les plus utilisées :

fréquence : nombre d'occurrences d'un terme dans un document ;

***tf.idf* :** *tf* mesure la fréquence du terme T_i dans le document D_j ; *idf* désigne l'inverse de la fréquence documentaire df (nombre de documents qui contiennent le terme T_i) et $idf = \log \frac{n}{df}$ avec *n* le nombre de documents dans la collection (Sparck Jones, 1979) (voir aussi la section 3.2.2)

Le degré de correspondance est déterminé par une mesure de similarité entre deux vecteurs $\overrightarrow{D_j}$ et $\overrightarrow{Q_m}$. Il existe plusieurs méthodes pour calculer la similarité entre deux vecteurs : cosinus (Losee, 1998), Jaccard (Romesburg, 1990), distance de Manhattan, etc. (descriptions dans la section 3.2.3).

L'avantage du modèle vectoriel par rapport au modèle booléen est de pouvoir effectuer un appariement partiel et ainsi de classer les résultats en fonction de leur degré de similarité. Il permet ainsi d'utiliser la technique 'relevance feedback' mentionnée à la section 5.5.2 avec les premiers documents trouvés. L'un des inconvénients est que les termes sont considérés comme indépendants les uns des autres.

Il existe quelques variations du modèle vectoriel plus récentes comme LSI ('Latent Semantic Indexing') (Deerwester *et al.*, 1990) qui exploite la similarité sémantique

entre les documents en tenant compte du contexte dans lequel apparaissent les termes. Le principe est de réduire les dimensions de la matrice documents-termes en rapprochant les termes selon leur sens. Tous les termes n'ont en effet pas la même importance dans la représentation d'un document. Le modèle propose de filtrer les termes ayant de faibles valeurs discriminantes. Il permet également de retrouver des documents qui n'ont aucun terme en commun avec la requête. Le modèle vectoriel généralisé (Wong *et al.*, 1985; Baeza-Yates & Ribeiro-Neto, 1999) permet de prendre en compte la dépendance entre les termes au sein des documents, manifestée par leurs cooccurrences.

Le modèle vectoriel est sans doute le modèle le plus utilisé en recherche d'information. Citons par exemple *SMART* (Salton, 1969; Rocchio, 1971), l'un des systèmes ayant eu le plus grand impact dans ce domaine.

Modèle probabiliste

Le modèle probabiliste estime, pour une requête donnée, la probabilité qu'un document appartienne à l'ensemble des documents pertinents ou à celui des documents non pertinents (Robertson & Sparck Jones, 1976; van Rijsbergen, 1979). En fait, cette approche ne propose pas de calculer exactement cette probabilité, mais plutôt d'utiliser les probabilités pour comparer et classer les documents.

Dans ce modèle, le poids des termes i, aussi bien dans un document j que dans une requête q, est représenté par des valeurs binaires : $W_{i,j} \in \{0,1\}; W_{i,q} \in \{0,1\}$. Une requête q est considérée comme un sous-ensemble des termes d'indexation. Soient R l'ensemble des documents pertinents et $\overline{R}$ l'ensemble des documents non pertinents. Soient $P(R|\overrightarrow{D_j})$ la probabilité que le document D_j soit pertinent pour la requête q ; et $P(\overline{R}|\overrightarrow{D_j})$ la probabilité que le document D_j soit non pertinent pour q. Le degré de correspondance entre D_j et q est défini par :

$$sim(D_j, q) = P(R|\overrightarrow{D_j})/P(\overline{R}|\overrightarrow{D_j})$$

En pratique, ces probabilités ne peuvent pas être calculées directement. Le théorème de Bayes est ainsi introduit pour calculer les probabilités *a posteriori* connaissant les distributions des observations *a priori*. La similarité est alors reformulée par :

$$sim(D_j, q) = P(R|\overrightarrow{D_j}) \times P(R)/P(\overline{R}|\overrightarrow{D_j}) \times P(\overline{R})$$

L'avantage du modèle probabiliste est qu'il permet de classer les documents selon leur probabilité de pertinence. Il permet ainsi de rationaliser la recherche avec une règle d'arrêt. En effet en connaissant la probabilité de pertinence d'un document, on peut estimer l'effort nécessaire pour trouver un autre document pertinent parmi les documents restants. L'inconvénient est que la fréquence des termes dans un document n'est pas prise en compte dans la pondération. De plus, les termes sont considérés comme indépendants les uns des autres. Ce modèle est à l'origine du système OKAPI (Robertson *et al.*, 1994) qui a montré des performances équivalentes au modèle vectoriel dans TREC.

5.2.4 Présentation des résultats de la recherche d'information

C'est par le biais de la présentation des résultats que l'utilisateur fait une première évaluation des documents qui lui semblent intéressants. Le jugement de la pertinence d'un document dépend tout d'abord du contenu du document, de la description du document, et du comportement de l'utilisateur (le savoir, la manière de formuler la requête ou de la reformuler, etc.). Un système de recherche doit permettre une présentation uniforme des résultats sous forme structurée en rendant compte des facteurs suivants.

Informations générales concernant la recherche

Il s'agit de donner un aperçu de l'ensemble des résultats de la recherche. La plupart des systèmes rappellent la requête originelle, les mots de la requête non traités, précisent le nombre de documents trouvés, le temps de recherche, les critères de tri des documents, et proposent une reformulation de la requête.

Informations sur la description des documents retournés

Ces informations comprennent le titre du document, l'URL ('Uniform Resource Locator') et le lien hypertextuel vers le document d'origine, l'extrait ou le résumé du document, la taille, le score de pertinence, la date de dernière mise à jour, la mise en évidence des termes de la requête avec leur contexte dans le document.

Classement des documents retournés

L'objectif est de regrouper des documents similaires en fonction de caractéristiques connues de l'utilisateur (domaine ou sujet de sa recherche, langue, etc.). L'avantage est que l'utilisateur peut avoir une idée globale des résultats et y trouver plus facilement les documents dont il a besoin. Nous donnons ici quelques critères de classement souvent utilisés : classement des documents selon leur pertinences par rapport à la requête [10], classement des documents par domaine de connaissance (informatique, science, médecine, juridique, etc.), classement par type des documents (article, catalogue, discours, etc.), classement hiérarchique des documents [11].

10. La pertinence d'un document est définie souvent en fonction de la présence et de la fréquence des termes de la requête dans le document ou du score de similarité, etc. Notons que la mesure de pertinence varie d'un système à l'autre.

11. Ce classement est souvent utilisé pour la recherche thématique dans un domaine de spécialité. Il nécessite l'utilisation d'une terminologie spécialisée, par exemple le thésaurus MeSH (http://www.nlm.nih.gov/mesh) dans le domaine médical, et multilingue dans le cas de la recherche d'information multilingue.

Possibilité de reformuler la requête

Certains systèmes de recherche permettent de reformuler la requête d'une manière explicite en interrogeant l'utilisateur ou implicitement en effectuant une nouvelle recherche à partir des premiers documents pertinents retournés. Ces nouvelles techniques ont une grande influence sur le développement de la recherche d'information (Nie, 2001).

5.3 Problématique du passage d'une langue à une autre

La recherche d'information translangue peut être considérée comme une extension de la recherche d'information monolingue. Les techniques utilisées, que ce soit pour l'indexation ou la recherche des documents, sont les mêmes. Pourtant le traitement du multilinguisme impose l'emploi de techniques supplémentaires pour traduire la requête dans la langue cible des documents recherchés ou pour traduire les documents de recherche dans des langues différentes. La recherche d'information translangue crée donc quelques problèmes particulièrement liés à la traduction automatique, qui la distinguent de la recherche d'information monolingue.

Rappelons que le principe de la recherche d'information est l'appariement entre les mots de la requête initiale et ceux des documents recherchés. La pertinence des résultats dépend en général du chevauchement entre les mots de la requête et les même mots des documents ou ceux qui sont sémantiquement proches dans les documents, en regroupant les mots appartenant à la même famille dérivationnelle[12] (Jacquemin, 1997a).

Plus un document contient les mots de la requête, plus il sera jugé pertinent à l'égard de cette requête. Or dans le cas de la recherche d'information translangue, la requête initiale est rédigée dans une langue différente de celle des documents. A l'exception de certains noms propres et cognats qui sont présentés sous la même forme dans les différentes langues, l'appariement direct sans reformulation de la requête initiale fonctionne rarement. Ainsi, l'un des principaux problèmes de la recherche d'information translangue est celui du passage d'une langue à une autre, problème plus général de la traduction (Grefenstette, 1998b).

5.4 Approches en recherche d'information translangue

5.4.1 Traduction des documents vs traduction des requêtes

Une manière de rapprocher une requête et des documents écrits dans des langues différentes est de traduire les documents dans la langue de la requête. L'avantage de

12. Ce regroupement se fait traditionnellement sur des procédures de désuffixage (stemming), qui reviennent à rapprocher des mots uniquement sur leur graphie.

cette approche est la possibilité d'exploiter le contenu des documents afin d'effectuer une traduction de qualité. En particulier, le problème posé par les termes polysémiques, c'est-à-dire l'ambiguïté de leur traduction, un des problèmes récurrents de la traduction automatique, est mieux traité par la prise en compte des informations contextuelles contenues dans les documents.

Cette approche pose cependant des problèmes insurmontables notamment dans le contexte d'une recherche sur le Web qui présente une croissance exponentielle en terme de nombre de documents disponibles. En l'occurrence le volume de stockage est multiplié par le nombre de langues que l'on souhaite mettre à la disposition de l'utilisateur. Cette façon d'opérer, irréaliste pour une recherche sur le Web, peut néanmoins être envisagée sur une base documentaire réduite ou sur un site Intranet.

L'alternative à la traduction des documents est la traduction de la requête. Cette approche a l'avantage d'être simple à mettre en œuvre. En général, les requêtes sont composées de termes simples dont la traduction est rapide. C'est la raison pour laquelle la plupart des systèmes de recherche translangue mettent en œuvre une traduction de la requête.

L'inconvénient de cette approche est de ne pas donner assez d'information sur le contexte des termes à traduire. Or la traduction de termes simples hors contexte pose des problèmes, notamment celui de la sélection de la traduction correcte dans le cas de termes ambigus. Par exemple le mot "marche" en français peut être traduit en anglais par "walking", "step", "progress", "working", etc. Plusieurs modèles de traduction décrits dans la section suivante ont été proposés dans le but de traduire les requêtes.

5.4.2 Modèles de traduction de requête

Trois principales approches de traduction des requêtes peuvent être utilisées en recherche d'information translangue.

Méthode basée sur la traduction automatique

L'utilisation d'un système de traduction automatique est l'approche la plus directe mais reste assez limitée. Le processus de traduction fournit une traduction de la requête en ne retenant qu'une seule solution parmi les résultats possibles. Les heuristiques de sélection d'une traduction parmi plusieurs sont peu efficaces (la première proposition ou la plus fréquente, etc.) parce qu'une requête contient peu de mots et donc peu de contextes qui peuvent aider l'automate à choisir. Une expérience rapportée dans TREC-6 'Text REtrieval Conference' (Oard, 1998), montre que la performance du modèle basé sur la traduction automatique dépend de la longueur de la requête. Une meilleure performance est obtenue pour des requêtes plus longues (composées de phrases). Certaines études (Hull & Grefenstette, 1996; Franz *et al.*, 1999; Braschler *et al.*, 1999; Fluhr *et al.*, 1998) ont montré que les techniques de TAL (Traitement

Automatique des Langues) comme la racinisation, l'étiquetage syntaxique, ou l'indexation sur les syntagmes et l'analyse statistique de corpus peuvent diminuer le bruit et apporter une amélioration significative des résultats.

Méthode basée sur les lexiques ou les thésaurus bilingues

Cette approche utilise les traduction de mots stockées dans un lexique bilingue pour traduire une requête. Elle présente l'avantage d'être simple. Il existe de nombreux dictionnaires en version électronique pour plusieurs langues. Les expériences de (Davis & Ogden, 1997; Ballesteros & Croft, 1996) utilisent le dictionnaire Collins anglais-espagnol et celles de (Fujii & Ishikawa, 1999) le dictionnaire anglais-japonais EDR[13].

Toutefois, l'utilisation d'un dictionnaire bilingue pour traduire une requête impose certaines limites qui réduisent la performance de la recherche. En premier lieu, les domaines spécifiques souffrent d'un déficit de vocabulaire bilingue sous forme électronique. De plus, les problèmes d'ambiguïté et donc de sélection de la bonne traduction sont ici aussi mal résolus. Les requêtes sont souvent des mots simples, leur traduction hors contexte est problématique.

Pour résoudre le problème de la sélection des traductions, le projet EMIR (European Multilingual Information Retrieval) mené par Fluhr (Fluhr *et al.*, 1998) utilise les documents recherchés pour filtrer les bonnes traductions à l'aide d'une analyse de cooccurrence. Par ailleurs, Ballesteros et Croft (Ballesteros & Croft, 1998) proposent une méthode de désambiguïsation par extension de la requête à l'aide d'analyses statistiques.

Dans la plupart des dictionnaires bilingues, il n'y a pas assez d'informations linguistiques pour pouvoir traiter le problème de la polysémie et de la synonymie ni pour décrire les relations entre les termes. Les premières expériences ont montré que l'utilisation d'un thésaurus multilingue permet de mieux traiter le problème de l'ambiguïté (Oard & Dorr, 1996). La méthode consiste à utiliser les relations conceptuelles entre les termes pour désambiguïser. Néanmoins, la constitution d'un thésaurus multilingue reste un travail laborieux et très coûteux. Le problème de la mise à jour des nouveaux concepts et de la formation de l'utilisateur à utiliser correctement les relations entre les termes sont des freins à l'usage de cette approche.

L'expérience de (Eichmann & Ruiz, 1998), utilisant le metathésaurus 'UMLS' (Unified Medical Language System) (www.nlm.nih.gov/research/umls) pour traduire les requêtes, atteint jusqu'à 70% de la performance de la recherche monolingue. Actuellement, la tendance est de compléter cette approche par une analyse statistique de corpus (similarité de vecteurs, etc.), afin d'améliorer la qualité de la traduction (Fung & McKeown, 1997; Picchi & Peters, 1998; Rapp, 1999; Chiao & Zweigenbaum, 2002a).

13. Japan Electronic Dictionary Research Institute. Technical terminology dictionary (information processing), 1995.

Méthode basée sur les corpus

Ces techniques exploitent des indicateurs statistiques portant sur les mots provenant de corpus alignés ou comparables. Lorsque les corpus sont conséquents, les techniques statistiques permettent d'obtenir des équivalents de termes dans différentes langues. A partir de corpus alignés, l'approche consiste à extraire un modèle de traduction en utilisant les cooccurrences des termes de différentes langues dans des contextes équivalents (Hiemstra *et al.*, 1997; Carbonell *et al.*, 1997; Yang *et al.*, 1998).

Dans le même ordre d'idée, la méthode 'LSI' (Latent Semantic Indexing) (Littman *et al.*, 1998; Oard & Dorr, 1996; Brown, 1998) examine la similarité des contextes dans lesquels se trouvent les termes. Pour cela, elle crée un espace sémantique multilingue à partir de textes parallèles. Les termes de langues différentes qui sont dans des contextes similaires se trouvent à proximité l'un de l'autre dans le nouvel espace ainsi créé. Lorsqu'un terme a toujours été traduit par un autre terme, leurs représentations dans cet espace sont identiques. De même, si un terme est souvent associé à un autre terme, par exemple le mot anglais *not* et le français *pas*, ils auront des représentations similaires dans l'espace sémantique. Trouver l'équivalent d'un terme dans une autre langue revient donc à trouver des termes de la langue cible ayant une distance minimale au terme source dans l'espace sémantique. La performance de cette approche dépend de la qualité et de la disponibilité de corpus alignés. La limite d'une telle approche est donc liée à l'acquisition de corpus parallèles qui reste souvent problématique et onéreuse pour les domaines de spécialité. Notons que certaines expériences tendent à exploiter les ressources disponibles sur le Web pour la constitution automatique de corpus alignés en recherche d'information translangue (Chen & Nie, 2000; Resnik, 1999).

D'autres études (Fung & McKeown, 1997; Fung & Yee, 1998; Rapp, 1999; Picchi & Peters, 1998; Sheridan *et al.*, 1998) se sont davantage intéressées à exploiter les informations provenant de corpus comparables. Dans le cadre multilingue, il s'agit d'un ensemble de documents de langues différentes rassemblés selon des critères similaires : le domaine, le genre [14], etc. L'idée repose sur l'hypothèse que les termes utilisés pour décrire un sujet particulier sont liés sémantiquement même à travers des langues différentes.

5.4.3 Désambiguïsation des traductions

Nous avons vu que chaque méthode de traduction présente des contraintes d'utilisation. Quelle que soit la méthode utilisée, la polysémie inhérente de certains mots pose le problème récurrent de la désambiguïsation de leurs traductions. Pour résoudre ce problème, plusieurs stratégies sont envisagées pour la sélection de la traduction correcte d'une requête. Soit seul le terme le plus utilisé dans les documents est retenu comme traduction de la requête (Ballesteros & Croft, 1998; Hiemstra & de Jong,

14. http://www.atala.org/je/010428/Habert/Perpignan00/footnode.html

1999), soit l'identification de termes composés, l'utilisation de corpus parallèles ou les techniques statistiques, comme information mutuelle (section 3.2.2), par exemple, sont intégrés dans le modèle de traduction (Hull, 1997; Davis, 1998; Ballesteros & Croft, 1998; Fujii & Ishikawa, 1999; Jang *et al.*, 1999).

Une autre manière de procéder à la désambiguïsation est d'utiliser toutes les traductions possibles de la requête pour effectuer la recherche en supposant que la pertinence du résultat dépend plus de la performance des approches de la recherche de documents que de celle de la désambiguïsation. Partant de cette hypothèse, (Hiemstra & de Jong, 1999) montrent qu'avec une méthode de recherche appropriée, l'utilisation de toutes les traductions possibles donne une meilleure performance que celle de l'utilisation d'une seule traduction. Il s'agit de la technique de l'extension de requête.

5.5 Reformulation d'une requête par extension

Les techniques d'extension de requête consistent à ajouter de nouveaux termes à la requête pour rendre celle-ci plus précise afin d'augmenter le taux de réussite de l'appariement avec les documents.

L'une des cause d'échec de la recherche d'information vient de la non concordance entre la requête et l'index des documents. En l'absence d'une connaissance approfondie de la collection, l'utilisateur a toutes les chances de former une requête en employant des termes qui sont sémantiquement proches mais pas identiques aux termes utilisés dans les documents. Par exemple, une requête portant sur le terme *digestion* ne va pas s'apparier avec un document dont le sujet porte sur la *salivation* alors qu'il s'agit de concepts proches.

De la même façon, une requête concernant un terme générique comme *cancérologie* doit pouvoir être précisée par une arborescence de mots clés (*antinéoplastiques, tumeurs, oncologie médicale*...) ou types de ressources (service d'oncologie médicale...).

On peut distinguer deux niveaux de reformulation de la requête. L'un concerne les relations morphosyntaxiques entre les termes (*lésion hépato-cellulaire* ↔ *lésion de la cellule hépatique* ; *cœur* ↔ *coronaire*) et l'autre les relations sémantiques (*ACTH* ↔ *adrenocorticotrophine* ↔ *corticotrophine*). C'est la reformulation au niveau des relations sémantiques comme la synonymie qui fait l'objectif de la présente étude. Ce type de liens correspond aux relations prédéfinies d'un thésaurus ou d'un réseau sémantique.

5.5.1 Problématique de l'extension de requête

Le processus d'extension est confronté à deux problématiques : d'une part le choix des termes à ajouter et d'autre part la pondération de ces termes. Le poids attribué à un terme ajouté à une requête est en général déterminé de manière empirique. On lui

donne par exemple le poids du terme initial multiplié par un coefficient réducteur (0.25 dans (Fox, 1980)).

La sélection des termes à ajouter est une opération plus complexe et a plus de conséquences sur le résultat de la recherche. En effet, une requête étendue contient plus de termes et a plus de chance de repérer des documents pertinents. Cette approche diminue donc le 'silence', *i.e.*, le nombre de documents non trouvés, mais par la même occasion elle augmente le 'bruit', *i.e.*, le nombre de documents non pertinents trouvés, car il est probable que les documents retrouvés grâce à l'extension ne soient pas toujours pertinents.

Comme pour la traduction de requête, l'extension peut difficilement se concevoir de façon entièrement automatique. Le risque est de dégrader la performance des résultats de la recherche par l'ajout de termes erronés. L'extension de requête doit être supervisée par l'utilisateur ce qui nécessite de sa part un minimum de compétences sur le domaine. Le contexte multilingue ajoute alors une difficulté car en plus de compétences sur le domaine, l'utilisateur est supposé posséder des compétences linguistiques (sur les langues concernées a priori langues étrangères pour lui) afin de décider des termes à ajouter dans ces différentes langues.

5.5.2 Différentes approches de l'extension

Les sources des termes ajoutés et les techniques de sélection des termes sont deux facteurs essentiels du processus d'extension de la requête. On peut distinguer deux sources principales en fonction des modèles de recherche (*booléen, probabiliste ou vectoriel*). D'une part les documents de la collection, où le terme ajouté est le plus souvent calculé à partir de ses cooccurrences avec les termes de la requête, et d'autre part un thésaurus, où le terme ajouté est dérivé à partir de ses relations terminologiques avec les termes de la requête.

Extension à partir des documents de la collection

Le calcul d'un terme nouveau à partir des documents repose sur l'idée que la proximité entre termes dépend de leur cooccurrence dans les documents. Plus deux mots cooccurrent dans les textes, plus ils sont considérés comme fortement liés. Les techniques utilisées calculent les liaisons entre les termes de la requête et ceux du corpus. C'est le cas du 'nearest neighbors expansion' (Willett, 1981) ou de la technique du 'maximum spanning tree' (Harper & van Rijsbergen, 1978). Il s'avère que toutes les cooccurrences ne sont pas intéressantes pour l'extension.

Pour limiter et sélectionner les sources intéressantes, la méthode 'relevance feedback' consiste à reformuler une requête à partir des documents jugés pertinents par l'utilisateur lors de la première recherche. Cette technique a été proposée par (Salton & Buckley, 1990) en utilisant l'algorithme de (Rocchio, 1971) pour calculer les poids des termes des documents afin de les classer en vue de l'extension. Cette méthode est

efficace parce que seuls les documents jugés pertinents par l'utilisateur seront retenus et traités pour y extraire des termes, mais elle demande beaucoup d'efforts de la part de l'utilisateur pour sélectionner les documents.

La variante 'local (pseudo) feedback' consiste à ne retenir que les premiers documents retournés par le système de recherche d'information lors de la première recherche, et d'ajouter à la requête initiale des termes utiles qui ont participé à la recherche, trouvés dans ces documents. L'avantage de cette approche est qu'elle s'effectue implicitement par le système et dispense l'utilisateur de la validation.

Extension à partir d'un réseau sémantique

L'autre source de termes ajoutés pour étendre une requête est un réseau sémantique tel qu'on en trouve dans un thésaurus ou une terminologie structurée. Le principe vise à exploiter les relations établies entre les termes (hiérarchie, association, équivalence, causalité...) pour déterminer les termes à ajouter. L'opération doit être effectuée en mode interactif et non uniquement automatique car cela risquerait de produire trop de bruit lors de la recherche. L'utilisateur se voit proposer une liste de termes issus d'un thésaurus, parmi lesquels il ne retient que les termes qu'il juge pertinents pour les ajouter aux termes de la requête initiale ou les remplacer.

Par ailleurs, l'expérience de (Qiu & Frei, 1993) montre que la sélection parmi les termes liés aux concepts de la requête dans son ensemble donne un meilleur résultat que celui de l'extension sur les mots isolés de la requête. Cela conforte l'idée que les termes composés doivent être traités comme des entités à part entière et milite en faveur de l'approche « conceptuelle » pour la reformulation de requête.

5.5.3 Extension de requête et recherche d'information translangue

Dans le cadre de la recherche d'information translangue, les techniques d'extension sont appliquées lors de la traduction de requête. Elles permettent de diminuer le nombre de traductions erronées et de faciliter la sélection de la traduction correcte. En fonction du modèle de traduction adopté, le processus d'extension de requête peut intervenir avant ou après la traduction et de diverses façons.

L'idée de Picchi (Picchi & Peters, 1998) est d'exploiter les contextes dans lesquels apparaît souvent le terme en question à l'aide d'un corpus spécifique du domaine. Il crée un vecteur de contexte, c'est-à-dire une liste des mots les plus proches d'un terme donné au sens de l'information mutuelle (Church & Hanks, 1990). Les mots du vecteur sont ensuite traduits à l'aide d'un dictionnaire bilingue et en résulte un vecteur de contexte dans la langue cible. A l'aide d'un corpus similaire en langue cible, les mots qui ont des vecteurs similaires sont sélectionnés pour constituer les traductions du terme en question.

Une autre approche, (Ballesteros & Croft, 1997; Ballesteros & Croft, 1998) montre que l'extension de requête améliore sensiblement la qualité de la traduction ba-

sée sur un dictionnaire. L'expérience porte sur la méthode 'local feedback' (l'analyse des cooccurrences des termes dans les premiers documents retournés). Elle montre tout l'intérêt de combiner une extension avant la traduction et une autre après la traduction de la requête.

5.6 Conclusion

Nous avons passé en revue les principes de la recherche d'information translangue à travers l'étude de la recherche d'information du point de vue général et des problèmes liés à la traduction des mots de la requête dans le cadre multilingue. Un des moyens utilisés pour lever les barrières linguistiques entre une requête et des documents de langues différentes consiste à reformuler la requête.

Ce chapitre s'est attardé sur deux traitements utilisés pour la reformulation d'une requête : la traduction qui consiste à ajouter la ou les traduction(s) de la requête initiale et l'extension qui vise à y ajouter des termes liés sémantiquement aux termes de la requête. Les deux principales sources utilisées par les méthodes de reformulation sont les documents à partir desquels les termes ajoutés (traduits ou liés) sont extraits, et les terminologies multilingues décrites où les termes ajoutés sont inférés à partir des relations terminologiques.

L'utilisation de terminologies spécialisées multilingues pour traduire ou étendre un requête pose un problème général de leur construction et de leur actualisation (section 2.1). Cela est particulièrement sensible pour les domaines en constante évolution comme la médecine. Dans le cadre de la recherche translangue, une des réponses à cette problématique est l'acquisition automatique lexicale spécialisée et multilingue à partir des textes du domaine qui vise à faciliter la traduction ou l'extension des requêtes. A travers plusieurs expériences de recherche d'information translangue, le chapitre suivant en montre l'apport notamment pour la traduction des mots 'inconnus' d'une requête, *i.e.*, qui ne sont pas dans les ressources multilingues disponibles.

Chapitre 6

Expériences de recherche d'information translangue

6.1 Introduction

Ce chapitre décrit plusieurs expériences dont le but est de mettre en évidence l'apport de notre modèle pour la recherche d'information translangue.

L'évaluation d'un système de recherche d'information est une tâche importante et délicate. La recherche d'information a pour but de trouver des documents pertinents voire utiles à partir d'une requête utilisateur. La difficulté est de savoir comment établir une correspondance entre les pertinences du point de vue de l'utilisateur et de celui du système.

En général, la pertinence d'un système de recherche doit correspondre au jugement de pertinence de l'utilisateur. Plus les réponses retournées par le système correspondent à celles que l'utilisateur attend, meilleur est le système. Dans l'idéal, le système trouve tous les documents attendus qui répondent à la requête de l'utilisateur et rien que les documents pertinents.

Cette correspondance subit des dégradations au fur et à mesure à travers les différents processus de recherche (*e.g.*, au niveau de l'utilisateur, la représentation du besoin d'information par une requête en langue naturelle ; au niveau du système de recherche, les traitements de la requête et l'appariement entre celle-ci et les documents dans la base de données, etc.). De plus, pouvoir mesurer la pertinence suppose que l'on soit capable de modéliser l'utilisateur dans ses attentes. L'évaluation tente de trouver certains comportements communs entre les utilisateurs et de les formaliser, c'est-à-dire de simuler des requêtes ou des comportements de recherche afin de mesurer l'écart de pertinence entre la machine et l'utilisateur.

Beaucoup d'éléments sont en jeu dans l'évaluation d'un système de recherche : la base des documents, la présentation des résultats, l'utilisateur, le *rappel*[1], la *pré-*

1. Le rappel mesure la proportion des documents pertinents retrouvés parmi les documents pertinents

cision[2], etc. Nous abordons dans ce chapitre les différents aspects de l'évaluation en nous focalisant sur la mesure de la paire *précision/rappel* et sur le principe de la collection de test. Avant de voir comment procéder à l'évaluation de la performance d'un système de recherche d'information (section 6.2.2), nous allons d'abord nous attarder sur la notion de *pertinence* (section 6.2.1). Nous présentons ensuite la collection de test utilisée (section 6.3) et les résultats des expériences sur la traduction de requêtes (section 6.4.3) et l'extension de requêtes (section 6.4.4).

6.2 Evaluation d'un système de recherche d'information

La notion centrale dans l'évaluation d'un système de recherche est celle de la *pertinence* d'un document à l'égard d'une requête exprimée par l'utilisateur. La performance d'un système se mesure généralement autour de cette notion. Un document est considéré comme pertinent s'il contient des informations qui répondent au besoin d'information de l'utilisateur, sinon il est considéré comme non pertinent. Le but de tous les systèmes de recherche est de pouvoir trouver le plus possible de documents pertinents et le moins possible de documents non pertinents. La problématique est alors de définir la pertinence et à partir de cette définition quels éléments prendre en compte dans sa mesure.

6.2.1 Notion de pertinence

Une des difficultés pour définir la pertinence provient de la subjectivité inhérente à cette notion. Les utilisateurs d'un système de recherche ont des besoins très variés et adoptent souvent des critères différents pour juger de la pertinence d'un document. La pertinence n'est pas une relation isolée entre un document et une requête. Elle fait appel aussi au contexte du jugement : le besoin d'information, les compétences de l'utilisateur, la représentation du document, l'impression de nouveauté, etc. (Baeza-Yates & Ribeiro-Neto, 1999).

Certains facteurs, comme l'ordre de présentation des documents, les possibilités d'expression de la requête, sont plus faciles à analyser et à modéliser. D'autres facteurs au contraire, comme l'état du savoir de l'utilisateur, le besoin d'information, sont exprimés par les utilisateurs donc varient en fonction des individus. La prise en compte de la subjectivité de l'utilisateur est particulièrement importante. Or il est difficile d'avoir un échantillon suffisamment important d'utilisateurs pour prendre en compte l'ensemble des besoins. De plus, il est possible que le même utilisateur change son jugement de pertinence au cours du temps.

---

dans la base.

2. La précision mesure la proportion des documents pertinents parmi tous les documents retrouvés par le système.

Mizzaro (Mizzaro, 1997; Mizzaro, 1998) a fait une synthèse sur la notion et les différents aspects de la pertinence dans le contexte de la recherche d'information. Draper (Draper, 1998) a réalisé une étude précise sur le modèle de Mizzaro. Dans le cadre de notre étude, nous considérons la pertinence du point de vue du système. Nous nous intéressons aux effets de la traduction de la requête sur le résultat de la recherche d'information. Nous nous contentons donc d'effectuer une évaluation basée sur les critères traditionnels de pertinence (*précision/rappel*) à l'aide d'une collection de test, approche dite 'expérimentale'[3].

Rappelons que le but de la recherche d'information est de donner l'accès aux documents répondant à un besoin d'information exprimé par une requête généralement formulée en langue naturelle. La performance d'un système de recherche doit être définie en fonction de la correspondance entre les réponses du système et les réponses que l'utilisateur espère obtenir. Beaucoup de variables interviennent dans le jugement de pertinence, *i.e.*, réponses attendues d'un utilisateur (Mizzaro, 1998; Fluhr, 2000) : le document, la description de celui-ci, l'information que l'utilisateur reçoit, l'origine de la demande d'information, le besoin d'information, la représentation du besoin d'information en langage naturel (question) et en langage machine (requête). En général on emploie des questions artificielles (présélectionnées) et on utilise des juges spécialistes du domaine pour estimer la pertinence de la réponse. C'est le principe de la *collection de test*.

Les trois composants indispensables d'une collection de test sont en général : un ensemble de documents, un ensemble de requêtes et la liste des documents pertinents pour chaque requête. Par exemple 'Cranfield II' (Cleverdon *et al.*, 1996) est une des premières collections test et contient 1 398 documents et 225 requêtes. Pour que les résultats de l'évaluation pour une collection test soient significatifs, celle-ci doit contenir un nombre assez important de documents. L'ensemble des collections[4] développé dans le cadre de TREC ('Text Retrieval Conference') à partir de 1992 contient plusieurs millions de documents[5]. Lorsque le nombre de documents devient trop important dans une collection, le jugement de pertinence devient une tâche laborieuse car son volume est proportionnel à la multiplication du nombre des documents par celui des requêtes. C'est la raison pour laquelle la technique dite 'pooling method' (Sparck Jones & van Rijsbergen, 1975) (jugement sur échantillon) a été proposée afin de permettre de construire des collections test de grande taille. L'idée consiste à fixer un seuil pour le nombre de documents jugés pertinents, pour chaque requête, par différents systèmes de recherche d'information et à n'utiliser que ce nombre de documents dans le

3. Van Rijsbergen (van Rijsbergen, 1979) distingue deux types d'expérimentations pour évaluer les résultats de la recherche d'information : les approches expérimentales et opérationnelles. La principale différence entre ces deux approches est que dans les situations expérimentales, le jugement de la pertinence du document est donné par avance par un certain nombre de personnes sélectionnées (souvent des experts du domaine) ; dans un contexte opérationnel, la pertinence du document n'est pas prédéfinie et dépend donc du point de vue individuel de l'utilisateur.

4. http://trec.nist.gov/data.html

5. Very Large Corpus Track in TREC, http://trec.nist.gov

jugement des experts.

6.2.2 Mesures de la performance des systèmes de recherche d'information

Les deux critères les plus utilisés pour évaluer un système de recherche d'information sont le taux de *précision* et celui de *rappel*. Lorsqu'il s'agit d'un système basé sur le modèle booléen qui donne des réponses de forme binaire (Oui/Non) (un document est pertinent ou pas pour une requête donnée) et que l'ensemble des documents pertinents est connu dans la base, il est possible de mesurer la performance du système à partir des paramètres définis par le tableau 6.1. La précision est définie par la pro-

TAB. 6.1 – *Table des mesures de performance d'un système*

	Pertinents	Non pertinents
Trouvés	a	b
Non trouvés	c	d

portion de documents pertinents retrouvés par rapport au nombre total de documents retrouvés par le système et le rappel par le rapport du nombre de documents pertinents retrouvés par rapport au nombre total de documents pertinents de la base :

$$Précision = \frac{a}{a+b} \quad Rappel = \frac{a}{a+c}$$

D'un autre point de vue, nous pouvons aussi définir les notions de 'bruit' et 'silence' qui sont respectivement complémentaires de la précision et du rappel :

$$Bruit = 1 - Précision \quad Silence = 1 - Rappel$$

L'idéal serait qu'un système donne un taux de rappel et de précision de 100%. Cela signifie que tous les documents de la base retrouvés sont pertinents pour chaque requête. En pratique, ces deux mesures ne sont pas indépendantes, elles varient en proportion inverse. Lorsque l'une augmente, l'autre diminue. Les valeurs exactes de la précision et du rappel ne sont pas accessibles pour de multiples raisons liées aux difficultés dans le jugement de la pertinence, que nous avons évoquées dans la section précédente : ces valeurs peuvent varier en fonction des personnes qui examinent les documents. De plus le nombre de documents retrouvés n'est pas fixe pour toutes les requêtes, il peut varier en fonction du type de système utilisé. Par exemple, le modèle probabiliste propose souvent une liste de réponses longue : en général, tous les documents de la base sont ordonnés. Une longue liste correspond en général à un ratio de rappel élevé mais un taux de précision bas.

Courbe précision-rappel

Pour toutes les raisons décrites ci-dessus, les valeurs exactes de précision et de rappel ne sont pas accessibles. Ainsi il est préférable de calculer des valeurs relatives afin de pouvoir comparer différents systèmes ou différents paramètres au sein d'un même système. Ces mesures sont en général calculées à plusieurs niveaux, *i.e.*, la précision aux N premiers documents retrouvés ('cut-off level N') ou la précision pour des valeurs prédéfinies du rappel (de 0% à 100% par intervalle de 10%). La performance d'un système peut être représentée ainsi sous la forme d'une courbe précision/rappel.

Il est fréquent d'appliquer l'interpolation sur cette courbe car les valeurs exactes du rappel peuvent ne pas être atteintes. L'interpolation consiste à lisser la courbe initiale pour qu'elle soit décroissante : la valeur interpolée de la précision pour un point de rappel i est la précision maximale obtenue pour un point supérieur ou égal à i. L'avantage de cette technique est de permettre de définir la précision pour des valeurs standardisées.

Lorsque l'on veut comparer deux systèmes ou deux méthodes de recherche d'information, il est difficile dans la pratique d'utiliser les courbes précision-rappel décrites précédemment comme seule base de comparaison. Il arrive parfois qu'un système présente une meilleure performance sur certaines requêtes par rapport à un autre système et que ce soit le cas contraire pour d'autres requêtes. Dans ce cas-là, l'évaluation de la performance peut s'effectuer sur la base d'une seule valeur. Nous présentons ici quelques mesures souvent utilisées : précision moyenne non interpolée et interpolée, R-précision et F-mesure.

Précision moyenne non interpolée

Lorsque les documents retrouvés sont classés par ordre décroissant en fonction de leur probabilité de pertinence dans l'ensemble des documents (ce qui est le cas des systèmes basés sur le modèles vectoriel et probabiliste), la précision moyenne non interpolée 'Uninterpolated Average Precision', mesure utilisée dans TREC (Hull, 1997; Hull & Grefenstette, 1996), permet d'évaluer la performance du classement. Cette mesure favorise les systèmes de recherche qui trouvent plus de documents pertinents parmi les premiers documents retournés.

L'idée est de calculer, pour chaque requête, les valeurs de précision obtenues sur les document pertinents en tenant compte de leur position dans la liste des documents retrouvés. La précision moyenne est obtenue en divisant la somme de ces différentes valeurs de précision par le nombre total de documents pertinents dans la base. Considérons une requête donnée Q pour laquelle les 3 documents pertinents sont trouvés aux rangs 1, 5 et 8. Les précisions obtenues pour chaque document pertinent sont respectivement 1, 0,4 et 0,375. La précision moyenne non interpolée est de 0,59 résultat de $(1 + 0,4 + 0,375)/3$.

Précision moyenne interpolée

La précision moyenne interpolée consiste à calculer tout simplement la moyenne de toutes les précisions aux différents seuils de rappel pour l'ensemble des requêtes. La précision moyenne sur 11 points (à partir des précisions obtenues aux seuils de rappel 0.0, 0.1... 1.0) est une mesure souvent utilisée dans les évaluations des différents systèmes de recherche d'information. La précision moyenne sur 10 points (0.1, 0.2... 1.0) est également utilisée.

R-précision

La R-précision mesure la précision obtenue pour un nombre de documents retournés. Ce nombre est fixé pour chaque requête en fonction du nombre de documents pertinents présents dans la base. Les rangs des documents pertinents retournés sont donc ignorés. La R-précision peut être intéressante dans la mesure où la base contient un nombre important de documents pertinents ce qui est le cas pour les collections TREC.

La R-précision est calculée pour chaque requête, la moyenne des R-précisions est obtenue en additionnant ces différentes valeurs et en divisant la somme par le nombre total des requêtes. Par exemple, soit une expérience réalisée avec trois requêtes dont la première a 30 documents pertinents dans la base de recherche, la deuxième 50 et la troisième 15. Pour ces requêtes le système trouve respectivement 10 documents parmi les 30 premiers documents retournés, 30 parmi les 50 premiers et 10 parmi les 15 premiers. La moyenne des R-précision de cette expérience est 0.533 ($\frac{\frac{10}{30}+\frac{30}{50}+\frac{10}{15}}{3}$).

F-mesure

La F-mesure proposée dans la thèse de Van Rijsbergen (van Rijsbergen, 1979) mesure l'efficacité globale d'un système de recherche d'information. Elle combine la précision (P) et le rappel (R) en une seule mesure. En général, on donne la même importance à ces deux paramètres :

$$F = \frac{2 \times P \times R}{P + R}$$

6.3 Collection OHSUMED

6.3.1 Base de documents

La collection des documents *OHSUMED* (ftp://medir.ohsu.edu/pub/ohsumed) est un ensemble des 348 566 références extraites de MEDLINE[6]. La base MEDLINE couvre des références et des résumés d'articles de plus de 4 600 revues internationales,

6. MEDLINE est la base de données bibliographique de la National Library of Medicine.

principalement en anglais dans le domaine bio-médical. Elle contient plus de 12 millions de références de 1966 à nos jours. Chaque référence est indexée manuellement par des termes *MeSH*[7] (Hersh *et al.*, 1994). Les références *OHSUMED* sont composées du titre et/ou du résumé des articles parus dans 270 revues médicales entre 1987 et 1991. La collection est présentée en fonction des types d'information contenus dans les documents. Le tableau 6.2 montre un document extrait de la collection.

.I (ordre d'apparition)	8
.U (identifiant MEDLINE)	87049096
.T (titre)	Intraosseous infusion of phenytoin
.W (résumé)	In the critically ill child, administration of fluids and medications via the intraosseous route often proves lifesaving. The authors describe the case of a child with status epilepticus in whom phenytoin was administered via the intraosseous route, and seizure resolution and therapeutic serum levels were achieved. Intraosseous drug administration should be reserved for the rare critically ill child in whom vascular access proves impossible.
.M (termes *MeSH*)	Case Report; Child, Preschool; Drug Administration Routes; Emergencies; Human; Male; Phenytoin/*AD/BL; Status Epilepticus/BL/*DT; Tibia
.A (auteur)	Walsh-Kelly CM; Berens RJ; Glaeser PW; Losek JD
.S (source)	Am J Emerg Med 8703; 4(6):523-4
.P (type de publication)	JOURNAL ARTICLE

TAB. 6.2 – *Exemple de document de la collection OHSUMED.*

6.3.2 Base de requêtes et jugement de pertinence

Dans le cadre de *TREC9*, 3 listes de requêtes ont été développées pour OHSUMED afin de permettre des expériences d'évaluation de systèmes de recherche d'information :

query.ohsu.* : un sous-ensemble de 63 requêtes extraites de la liste de requêtes originales dans la collection *OHSUMED*.

query.mesh.* : un ensemble de 4 904 termes correspondant aux concepts *MeSH* qui indexent les articles d'OHSUMED et à leur définition.

query.mesh-sample.* : 500 termes *MeSH* sélectionnés aléatoirement

Chaque document dans la collection est pondéré par sa pertinence par rapport à l'ensemble des requêtes citées ci-dessus. Pour la liste des requêtes *OHSUMED*, la pertinence des documents est validée par des experts[8]. Un document est considéré comme

7. Le nombre de concepts *MeSH* à l'époque était de 17 000 ; htpp://www.nlm.nih.gov/mesh/meshhome.html

8. http://trec.nist.gov/data/t9_filtering/README

pertinent pour la liste des requêtes *MeSH* s'il est indexé par le terme qui constitue la requête.

6.4 Expériences de recherche d'information translangue

6.4.1 Jeux de requêtes

L'objectif de ces expériences est d'évaluer dans quelle mesure notre modèle peut être utilisé pour la recherche translangue et plus précisément pour la reformulation des requêtes par la traduction de celles-ci. L'évaluation a porté dans une direction de langue, du français vers l'anglais. Il faut donc disposer d'un jeu de requêtes en français.

Les requêtes en anglais initiales utilisées sont les requêtes *MeSH* initiales de la collection *OHSUMED* (query.mesh*). Ces requêtes sont composées d'un terme *MeSH* et de sa définition. Par ailleurs, une partie (2 345 concepts et leur définition - MeSH scope note -) est traduite et disponible sur le site CISMeF[9]. Cette liste a été utilisée pour préparer le jeu de requêtes de ces expériences de la façon suivante. Les 4 904 requêtes *MeSH* de la collection *OHSUMED* ont été alignées sur cette liste de définitions traduites. Les requêtes pour lesquelles la traduction de la définition n'est pas disponible ont été écartées. Finalement, la liste résultante est une liste de requêtes en français qui sont la traduction des requêtes *OHSUMED*. Le nombre de requêtes obtenues est de 489.

Enfin nous nous situons dans le cas où le lexique partiel utilisé (section 4.2.2) ne peut pas aider l'utilisateur à traduire sa requête, c'est-à-dire dans le cas où au moins un mot de la requête est inconnu du lexique en question. Finalement 263 requêtes ont été retenues parmi les 489 précitées. Parmi ces 263 requêtes, 56 sont composées de mots inconnus et 207 contiennent au moins un mot inconnu du lexique partiel. La liste des 263 requêtes se trouve dans l'annexe B.1. L'idée initiale était d'effectuer les expériences de recherche en utilisant une liste de requêtes courtes composée des termes *MeSH* et une liste de requêtes longues composée de leur définition.

Dans quelques expériences initiales que nous avons effectuées en utilisant les définitions des concepts traduites par notre modèle, les résultats de recherche obtenus n'étaient pas satisfaisants. Une explication possible est que les concepts *MeSH* (scope note), en général composés de plusieurs phrases dont la plupart des composants sont des mots généraux[10]. Or notre modèle propose plutôt la traduction des mots spécialisés à partir des corpus spécialisés. L'utilisation de notre modèle pour traduire des mots généraux introduit plus du bruit que de bonnes traductions et n'est alors pas favorable. Nous avons finalement utilisé la liste des concepts *MeSH* pour les expériences de recherche d'information translangue. Il s'agit d'une liste de requêtes courtes cor-

9. Nous remercions SJ. Darmoni de nous les avoir envoyées.

10. Un exemple de la définition du concept *anthropométrie* (MSH259) : « Technique qui concerne la mesure de la taille, du poids, et des proportions de l'Homme ou de tout autre corps de primate. ».

respondant aux concepts *MeSH* exprimés par des termes complexes, *e.g.*, *anticorps antivirus hépatite, acides gras indispensables, vaccin antipoliomyélitique*, etc. Nous estimons que les requêtes courtes (quelques termes) correspondent plus à la situation réelle de recherche d'information.

6.4.2 Base des documents de recherche

Deux séries d'expériences différentes ont été réalisées. L'une a consisté à évaluer notre modèle en traduction de requêtes et l'autre en extension de requêtes. Pour toutes ces expériences, un outil de recherche d'information basé sur le modèle vectoriel (Sta, 1993; Sta, 1997) a été mis en œuvre. Il est basé sur le cosinus pondéré par *tf.idf*, calculé entre les requêtes et les documents. Ainsi chaque requête et document doivent préalablement être transformé en vecteur pondéré.

Puisque le jeu de requêtes est composé uniquement de concepts *MeSH* (section 6.4.1), dans les champs titre et résumé de chaque document, tous les termes ont été retenus en dehors des mots grammaticaux. Le champ correspondant aux termes *MeSH* désignés manuellement comme index du document par l'expert a été exclu afin de ne pas biaiser la performance de la recherche (Petras *et al.*, 2003).

Les termes simples sont les mots qui constituent les éléments des vecteurs de documents et de requêtes. Rappelons que l'objectif de ces expériences est de valider notre modèle de traduction dans le cadre de la recherche d'information translangue. L'intention n'est donc pas de construire un modèle de recherche mais de proposer un outil d'aide à la construction de requête qui peut être intégré dans un système de recherche d'information. Le fait que le terme simple soit central dans notre modèle de traduction fait que d'une part la traduction et l'extension de requêtes et d'autre part l'appariement avec les documents se font sur chacun des mots constituant les requêtes ou les documents, *i.e.*, mot par mot dans le cas des termes complexes.

Dans la suite, chaque série d'expériences est présentée avec les traitements effectués pour reformuler les requêtes ainsi que l'évaluation des résultats des recherches. Les résultats sont obtenus en utilisant la base de documents *OHSUMED*87 qui contient 54 709 documents dont 13 677 sont pertinents pour les 263 requêtes retenues. Cela correspond à une moyenne de 52 documents pertinents par requête.

6.4.3 Expériences sur la traduction de requêtes

Reformulation de requête par la traduction

Les expériences sur la traduction des requêtes consistent à évaluer trois méthodes différentes de traduction :

– Les mots de la requête sont traduits par le lexique bilingue partiel. Les mots inconnus (absents du lexique) sont ignorés. Cette méthode est nommée *DICO*.

– Les mots de la requête sont traduits d'abord par le lexique bilingue partiel *DICO*. La traduction des mots inconnus est ensuite ajouté automatiquement (non supervisé) en sélectionnant le premier candidat proposé par notre modèle. Cette méthode est nommée *DICO+Auto*.

– Les mots de la requête sont traduits par la méthode *DICO*. Les mots inconnus sont ensuite traduits par notre modèle avec une sélection semi-automatique de la bonne traduction parmi les 30 premiers candidats proposés (mode supervisé). Lorsque aucune traduction d'un mot inconnu n'est jugée pertinente, alors ce mot est ignoré. Cette méthode est nommée *DICO+Auto+UMLS*.

Le mode supervisé se déroule en deux phases. La première étape automatisée est guidée par le métathésaurus UMLS et la seconde manuelle par le dictionnaire Masson de médecine et de biologie (Manuila *et al.*, 1970).

L'algorithme de la première étape vise à mettre en valeur un couple de mots (un mot inconnu et un candidat à sa traduction) qui sont souvent attachés aux mêmes concepts dans l'UMLS. Nous nommerons cet algorithme 'coefficient d'association conceptuelle'. Pour chaque mot de requête inconnu du lexique initial, l'algorithme consiste à :

1. vérifier sa présence dans les termes français de l'UMLS et récupérer l'identifiant du ou des concepts (CUI) auxquelles ces termes sont associés ;

2. récupérer tous les termes anglais appartenant à ces concepts ;

3. calculer la fréquence d'apparition des premiers candidats à la traduction proposés par notre modèle dans les termes anglais récupérés ;

4. reclasser les candidats en ordre décroissant de leur fréquence d'apparition.

Parmi les candidats qui présentent une forte association conceptuelle avec un mot inconnu, nous avons retenu manuellement celui qui présente une ressemblance graphique (section 3.2.5) avec le mot français initial. Cette étape a été effectuée à l'aide du dictionnaire médical Masson pour aider à trancher dans le cas d'un mot ambigu. Par exemple, pour le mot inconnu *affectif* de la requête MSH102 *symptome, affectif* appartenant au concept C0001726, la traduction attendue est *affective* qui n'est pas présente parmi les premiers candidats proposés. C'est *bipolar* qui est le candidat reclassé en premier rang selon le coefficient d'association conceptuelle. En fait le couple *affectif/bipolar* dans ce cas a été mis en valeur par les synonymes (*affective psychosis*, *bipolar*, *bipolar affective disorder*, etc) appartenant à une autre classe C0005586 qui désigne *bipolar disorder/trouble bipolaire*.

Pour la requête MSH1952 *basedow, maladie*[11], l'indice d'association conceptuelle a permis de sélectionner la bonne traduction *graves* pour *basedow* qui est inconnu du lexique. De même, *RNA* a été reclassé en premier rang par rapport au rang

11. En principe, les termes français MeSH sont en majuscule et sans accent.

3 initialement proposé pour la requête MSH4048 *épissage ARN*. Dans certains cas le reclassement a fait ressortir des couples de mots qui constituent en effet un terme composé au lieu de bonnes traductions. C'est le cas de la requête MSH1013 *rhume banal* pour laquelle *banal* est absent du lexique initial et la bonne traduction *common* ne figure pas parmi les premières traductions proposées. Le candidat ayant le coefficient d'association conceptuelle le plus élevé est *cold*, qui est en effet un composant de la requête d'origine en anglais*commun cold*.

Pour les candidats d'un mot inconnu donné, qui ne figurent pas dans les termes anglais associés aux mêmes concepts que le mot inconnu, nous avons gardé leur rang initial calculé en fonction de la similarité croisée. De nouveau, nous avons eu recours au dictionnaire Masson pour vérifier les synonymes ou les différents sens des mots spécialisés.

Parmi les 240 mots inconnus présents dans notre liste de requêtes, l'indice d'association conceptuelle issue de l'étape automatisée a permis de sélectionner la bonne traduction parmi les 30 premiers candidats pour 45 mots. Nous avons sélectionné manuellement la bonne traduction parmi les 30 premiers candidats pour 18 mots. Cette sélection semi-automatique nous a permis de rajouter la bonne traduction pour 63 mots des requêtes.

Cette supervision est certes loin d'être parfaite et ne résoud pas le problème de l'ambiguïté des requêtes mentionné dans la section 5.4. Dans ces expériences la désambiguïsation des requêtes aurait nécessité plus de compétence et notamment l'intervention de spécialistes du domaine.

Le tableau 6.3 donne quelques exemples de requêtes *OHSUMED* en anglais (MeSH EN), leur traduction dans INS2000 c'est-à-dire la version française du MeSH fournie par l'INSERM (Institut National de la Santé Et de la Recherche Médicale) (MeSH FR), les reformulations mot par mot par les trois méthodes de traduction : dictionnaire (DICO), dictionnaire en ajoutant la traduction des mots inconnus (DICO+Auto), dictionnaire en ajoutant la traduction des mots inconnus sélectionnés en fonction de leur fréquence d'apparition dans UMLS (DICO+Auto+UMLS).

Evaluation de la performance

L'évaluation des résultats obtenus sur les 263 requêtes est donnée dans le tableau 6.4. Trois mesures de pertinence sont présentés : la précision moyenne $\bar{P}$, la précision relative R-précision et les précisions $P(n)$ pour n documents retournés, calculées selon les critères présentés dans la section 6.2.2.

Sur l'ensemble des 263 requêtes, les meilleurs résultats sont obtenus par la méthode *DICO+Auto+UMLS*. Nous pouvons remarquer une amélioration nette par rapport aux deux autres méthodes de traduction : *DICO, DICO+Auto* quelle que soit la mesure d'évaluation utilisée. La précision moyenne et la R-précision augmentent d'environ 7%. La précision obtenue sur le premier document trouvé est améliorée de 14% en valeur absolue.

Requête 275	**antibodies, viral**
MeSH FR	anticorps antiviral
DICO	— —
DICO+Auto	antibody, alfacon
DICO+Auto+UMLS	antibody viral
Requête 3609	**poliovirus vaccine, oral**
MeSH FR	vaccin antipoliomyélitique sabin
DICO	vaccine
DICO+Auto	vaccine, mmr, live-attenuated
DICO+Auto+UMLS	vaccine poliovirus

TAB. 6.3 – *Exemples de requêtes : forme originale (MeSH EN) ; traduction humaine (MeSH FR) ; traduction par dictionnaire (DICO) ; traduction par dictionnaire + traduction automatique par notre modèle (DICO+Auto) et traduction par dictionnaire + traduction automatique + filtrage avec UMLS (DICO+Auto+UMLS).*

Méthodes	**DICO**	**DICO+Auto**	**DICO+Auto+UMLS**
Précision moy.	0.0793	0.0777	0.1506
R-précision	0.1012	0.0994	0.1698
$P(1)$	0.2069	0.2109	0.3394
$P(5)$	0.1435	0.1503	0.2555
$P(10)$	0.1238	0.1272	0.2198
$P(15)$	0.1119	0.1149	0.1970
$P(20)$	0.0988	0.1041	0.1759
$P(30)$	0.0917	0.0943	0.1557
$P(50)$	0.0699	0.0795	0.1230
$P(100)$	0.0495	0.0587	0.0902
$P(200)$	0.0384	0.0445	0.0654
$P(500)$	0.0188	0.0230	0.0331
$P(1000)$	0.0082	0.0126	0.0174

TAB. 6.4 – *Précisions pour les trois méthodes de traduction.*

La différence entre la méthode *DICO+Auto* et la traduction par dictionnaire (*DICO*) n'est pas significative bien que nous constations que l'ajout de la traduction proposée par notre modèle améliore légèrement les résultats en termes de précision pour les n premiers documents retournés. Les valeurs de la précision moyenne et de la R-précision diminuent légèrement lorsque l'on ajoute aux requêtes de façon automatique la traduction des mots inconnus.

L'explication vient du fait que sans validation manuelle, l'ajout automatique de la traduction d'un mot inconnu par la méthode *DICO+Auto* introduit du bruit. Par exemple, parmi les requêtes dans lesquelles un mot inconnu est mal traduit, on peut citer : MSH1464 *électrophorèse gel polyacrylamide* où *polyacrylamide* est traduit par *agarose* ; MSH1137 où *cristallographie* est traduit par *virology* ; MSH346 où *rendez-vous et programmes* est traduit par *substitute, care*.

Lorsqu'on examine la liste des candidats à la traduction proposée par notre modèle, on s'aperçoit que la bonne traduction d'un mot inconnu de la requête ne se trouve pas toujours en première position. Cela rend nécessaire un filtrage des candidats. La méthode *DICO+Auto+UMLS* dite 'supervisée' permet de sélectionner la bonne traduction parmi les candidats dans la liste proposée et améliore ainsi la performance des résultats.

Les courbes précision-rappel interpolées pour les trois méthodes de traduction sont présentées dans la figure 6.1. Ces courbes sont obtenues sur l'ensemble des requêtes en calculant, pour un point de rappel précis, la moyenne des précisions interpolées. Les différents seuils de rappel sont 0.0[12], 0.1... 1.0. Cette mesure permet de mieux illustrer les différences entre les méthodes. L'évolution des courbes précision/rappel confirme l'apport de la méthode *DICO+Auto+UMLS* sur les résultats de la recherche. Cette méthode permet en effet d'améliorer en même temps la précision et le rappel. Les traductions erronées sont filtrées manuellement à l'aide du métathésaurus UMLS et les traductions correctes sont retenues. Les performances des deux autres méthodes *DICO, DICO+Auto* sont très proches et toutes deux donnent des performances nettement inférieures à celles de la méthode *DICO+Auto+UMLS*. Cette constatation milite donc en faveur de l'utilisation des traductions proposées par notre modèle dans un cadre supervisé.

6.4.4 Expériences sur l'extension de requêtes

Reformulation des requêtes par extension

Les expériences réalisées sur l'extension de requêtes dans le cadre translangue consistent à reformuler une requête en y ajoutant toutes les traductions proposées pour les mots la composant. Les mots de la requête sont traduits d'abord par le lexique bi-

12. En réalité, il est impossible de calculer la précision pour le rappel 0. La valeur donnée ici correspond à la précision moyenne interpolée obtenue sur le premier document trouvé pour l'ensemble des requêtes utilisées.

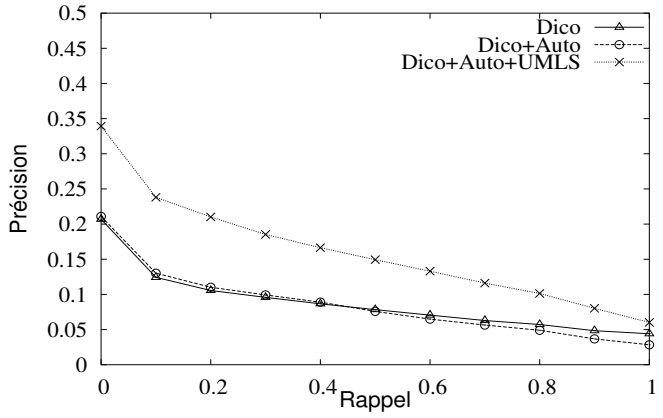


FIG. 6.1 – *Courbes Rappel-Précision interpolée des trois méthodes de traduction pour l'ensemble des 263 requêtes.*

lingue comme décrit dans la section précédente 6.4.3. A partir de ces requêtes traduites, deux méthodes d'extension ont été appliquées :

- Les traductions d'un mot inconnu de la requête sont ajoutées automatiquement (non supervisé) en sélectionnant les 5 premiers candidats proposés par notre modèle. Cette expérience est nommée *EXT-Auto*.

- Les traductions d'un mot inconnu de la requête sont ajoutées par notre modèle en sélectionnant parmi les 30 premiers candidats proposés ceux qui lui sont jugés sémantiquement proches. Lorsque aucune traduction d'un mot inconnu n'est jugée pertinente, la requête n'est pas étendue. Cette méthode est nommée *EXT-Auto-UMLS*.

Dans la méthode supervisée *EXT-Auto-UMLS*, pour filtrer les candidats proposés, nous avons utilisé de nouveau l'indice de coefficient d'association conceptuelle exploité pour les expériences sur la traduction de requête (section 6.4.3). Les candidats à la traduction d'un mot inconnu ayant une forte association conceptuelle avec ce dernier ont été sélectionnés manuellement pour étendre la requête. A ce stade, pour l'ensemble des 260 mots inconnus, 253 candidats ont été retenus pour étendre les requêtes.

Le tableau 6.5 illustre l'exemple de deux requêtes étendues par les méthodes décrites ci-dessus.

Evaluation de la performance

Les résultats obtenus par les deux méthodes d'extension sur l'ensemble des requêtes sont présentés dans le tableau 6.6. Afin d'évaluer l'effet de l'extension, nous utilisons comme mesure de la pertinence la précision moyenne interpolée, la R-précision

MeSH EN	antibodies, fungal
MeSH FR	anticorps, antifongique
DICO	—, antifungal
EXT-Auto	*antibody, antigen, result, high, positive*, antifungal
EXT-Auto-UMLS	*antibody, antigen, immunodeficiency*, antifungal
MeSH EN	abnormalities, drug-induced
MeSH FR	malformation origine chimique
DICO	malformation, origin, —
EXT-Auto	malformation, origin, *food, effect, service, protein, feedback*
EXT-Auto-UMLS	malformation, origin, *chemical, drug, abuse*

TAB. 6.5 – *Exemples d'extension pour les requêtes MSH273, MSH6 : forme originale (MeSH EN) ; traduction humaine (MeSH FR) ; traduction par dictionnaire (DICO) ; extension par traduction automatique (EXT-Auto) et extension par traduction automatique + filtrage avec UMLS (EXT-Auto-UMLS).*

et les précisions sur les n premiers documents trouvés, obtenues par la méthode *DICO* comme base de comparaison. Nous pouvons constater que l'application de l'extension

Méthodes	**DICO**	**EXT-Auto**	**EXT-Auto-UMLS**
Précision moy.	0.0793	0.0603	0.1488
R-précision	0.1012	0.0775	0.1674
$P(1)$	0.2069	0.1488	0.3293
$P(5)$	0.1435	0.1107	0.2529
$P(10)$	0.1238	0.0952	0.2177
$P(15)$	0.1119	0.0887	0.1950
$P(20)$	0.0988	0.0780	0.1739
$P(30)$	0.0917	0.0712	0.1544
$P(50)$	0.0699	0.0606	0.1247
$P(100)$	0.0495	0.0440	0.0902
$P(200)$	0.0384	0.0351	0.0660
$P(500)$	0.0188	0.0190	0.0335
$P(1000)$	0.0082	0.0110	0.0171

TAB. 6.6 – *Précisions pour les deux méthodes d'extension comparées à la traduction par dictionnaire.*

supervisée *EXT-Auto-UMLS* permet d'améliorer globalement et significativement les résultats quelles que soient les mesures d'évaluation utilisées. La précision moyenne interpolée atteint 14,88% et augmente de 7 points par rapport à la valeur de base obtenue par la méthode *DICO*. La R-précision augmente de 6 points pour atteindre 16,74%. La moyenne des précisions pour le premier document trouvé, obtenue sur l'ensemble des 263 requêtes, atteint 32,93% et augmente de 12 points.

La méthode d'extension automatique *EXT-Auto* ne permet pas d'améliorer les résultats. La perte en performance peut être due au bruit introduit en ajoutant les 5 premières traductions de façon non contrôlée. Cependant, la différence avec la méthode *DICO* sans extension est significative mais faible. La figure 6.2 illustre les courbes rappel-précision interpolée pour l'ensemble des 263 requêtes.

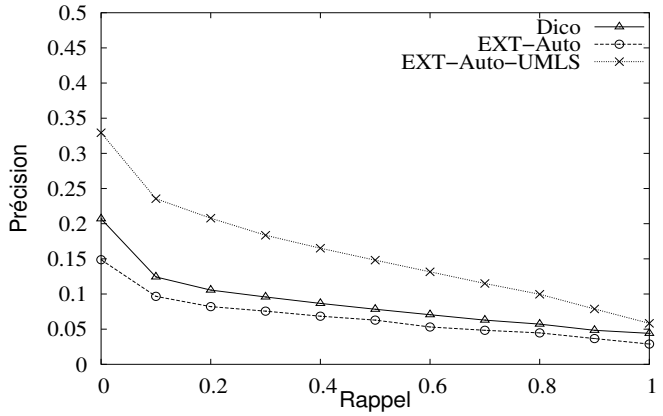


FIG. 6.2 – *Courbes Rappel-Précision interpolée des méthodes avec extension et de celle sans extension pour l'ensemble des 263 requêtes.*

Les résultats montrent que d'un côté, la méthode *EXT-Auto-UMLS* a permis d'améliorer de façon globale et significative la précision et le rappel sur l'ensemble des requêtes. Cette amélioration se traduit par le fait que la courbe précision/rappel sur l'ensemble de requêtes est systématiquement au-dessus de celles des deux autres méthodes. Par ailleurs l'apport de l'extension automatique d'un mot inconnu sans supervision est plutôt négatif.

6.5 Conclusion

Nous avons mise en œuvre et expérimenté dans ce chapitre notre modèle d'acquisition lexicale bilingue pour la recherche d'information translangue. Les expériences sur la traduction et l'extension de requête ont montré l'apport positif de notre modèle mais uniquement dans un cadre supervisé. Il s'agit de traduire ou d'étendre les mots d'une requête qui sont inconnus dans les ressources bilingues disponibles. Les bonnes performances obtenues reposent ainsi principalement sur le choix de la bonne traduction ou des bonnes extensions parmi une liste réduite. Cette remarque s'inscrit dans l'optique mentionnée initialement annonçant que cette approche est destinée à fournir une assistance pour des opérations de construction de lexique bilingue ou de recherche d'information translangue dans le domaine médical. Ce modèle s'adresse

ainsi à l'utilisateur en lui permettant de sélectionner la traduction ou les mots proches qui correspondent le plus à sa requête initiale. L'intérêt d'exploiter des corpus en recherche d'information réside entre autres dans le fait que le corpus a de plus grandes chances de contenir les traductions des néologismes que les lexiques existants.

Le fait que la collection OHSUMED soit destinée à la recherche d'information monolingue dans TREC fait que les expériences réalisées dans le cadre translangue sur cette collection sont rares. Citons néanmoins des expériences rapportées dans (Eichmann & Ruiz, 1998). Elles ont montré des résultats comparables à ceux de nos expériences. Avec une liste de 106 requêtes OHSUMED (6.3.2), traduites manuellement en français et en espagnol, l'objectif était de trouver les documents pertinents en anglais dans la base. La traduction des requêtes a été effectuée à l'aide du métathésaurus UMLS, en combinant plusieurs stratégies de sélection de la bonne traduction. La meilleure performance obtenue en terme de précision moyenne interpolée était de 0,1493 (atteignant 61% de la performance en recherche monolingue) et la précision pour les 5 et 10 premiers documents trouvés était de 0,1717 (56%) et 0,1358 (51%) respectivement.

Les performances médiocres de l'utilisation automatiques de notre modèle ont plusieurs explications.

La baisse de performance avec l'extension automatique des requêtes peut être le fait que dans nos expériences, les mots utilisés pour l'extension de requête sont extraits à partir de documents différents de ceux de la collection de OHSUMED sur lesquels est effectuée la recherche. A l'inverse, la pratique courante en extension de requête (section 5.5.2) est d'ajouter des mots extraits à partir des documents de la collection de recherche (par exemple la technique *relevance feedback* décrite dans les sections 5.5.2 et 5.5.3). Beaucoup de mots considérés sémantiquement proches et ajoutés à la requête initiale par notre modèle de traduction, sont alors absents dans les documents jugés pertinents et par conséquent inopérants. Les recherches sur l'application à la traduction et l'extension de requêtes doivent être poursuivies dans la direction d'une expérimentation de techniques de reformulation de la requête plus sophistiquées comme *relevance feedback*, utilisation d'un thésaurus bilingue ou d'un réseau sémantique (section 5.5).

Il faut noter de plus que certains mots inconnus et leur traduction sont tous deux absents des corpus sur lequel a été appliqué le modèle. Il en va ainsi par exemple pour le mot *dinitrochlorobenzene* de la requête MSH1303. Ce mot est alors ignoré par notre modèle.

Par ailleurs, étant donné que le lexique partiel est composé de mots spécialisés en majorité, beaucoup de mots généraux dans les requêtes sont ainsi considérés comme inconnus. Voici quelques requêtes dans lesquelles le mot inconnu est un mot général ou un nom propre : *rhume* $\underline{banal}$, *symptome* $\underline{affectif}$, $\underline{suivi}$ $\underline{soins}$ *patient*, *chromosome* $\underline{philadelphie}$ etc. Or nous avons montré au chapitre 4 la difficulté de la traduction des mots généraux par l'analyse des cooccurrences. La bonne traduction de ces mots ne se trouve donc pas toujours en première position sur la liste proposée. Cela explique que la méthode de traduction automatique n'améliore pas sensiblement les performances.

Enfin, pour certaines requêtes composées de termes complexes du type nom propre, (ce qui est fréquent dans le domaine médical (Bodenreider & Zweigenbaum, 2000) pour désigner les maladies, les virus ou syndromes), la méthode de traduction mot à mot pose des problèmes notamment lorsque ces termes et leurs traductions peuvent être représentés sous différentes formes dans le corpus. Par exemple, on trouve différentes formes de *basedow* dans la requête MSH1952 : *basedow, maladie*, *graves-basedow/grave-basedow* dans le corpus français et *graves/grave/basedow* dans le corpus anglais. Ainsi pour la requête MSH441 *ehlers danlos, syndrome*, *ehlers danlos* peuvent s'écrire sous la forme *ehlers-danlos* ou *meekeren-ehlers-danlos*, ce qui rend leur appariement plus difficile.

Une piste qui reste à explorer concerne l'application de notre modèle à l'indexation automatique des documents de langues différentes. L'idée consiste à partir d'un terme *MeSH*, à rechercher les documents pertinents selon le processus de recherche précité : traduction et extension contrôlées. Dans le cas où un document est déclaré pertinent, il pourra être indexé par le terme *MeSH* en question (et par le concept associé), c'est-à-dire par la requête qui a permis de le trouver. Nous avons montré au chapitre 4, que parmi les premiers candidats proposés par notre modèle, beaucoup d'entre eux sont sémantiquement proches ou sont des dérivations. Ces relations entre mots peuvent être exploitées ici pour résoudre la difficulté de recherche due au fait que les mots employés pour rechercher un document ne sont pas toujours présents dans celui-ci. Dans ce cas-là, le système de recherche utilise les candidats proches d'un terme d'indexation comme variantes pour retrouver les documents pertinents et par conséquent les indexer par ce terme. L'approche similaire a été expérimentée et évaluée sur l'indexation des documents monolingue dans (Pouliquen, 2002).

Chapitre 7

Discussion et conclusion

7.1 Introduction

Dans cette thèse, nous avons mis en œuvre et évalué un modèle d'acquisition lexicale bilingue à partir de corpus comparables dans le domaine médical. Ce modèle repose sur l'hypothèse de la symétrie distributionnelle, notion que nous avons avancée au chapitre 1. Elle postule que si deux mots de langues différentes ont des distributions contextuelles proches dans les deux directions de langues alors ces mots sont traductions mutuelles.

Nous avons présenté au chapitre 2 un état de l'art sur l'acquisition de lexique en y insistant sur le rôle central qu'y joue le corpus. Nous avons distingué à cette occasion l'acquisition lexicale monolingue de l'acquisition lexicale multilingue. La problématique de l'acquisition monolingue est plutôt liée à celle de l'extraction terminologique et de la structuration des termes. Cette problématique n'a pas été traitée car l'objectif est ici de mettre en correspondance traductionnelle deux mots issus de corpus et non de déterminer si ces mots sont des termes. La problématique en effet est bien celle de l'acquisition multilingue dont la difficulté est d'apparier des mots de langues différentes.

Nous avons ensuite montré l'intérêt qu'apporte ce modèle en exploitant diverses ressources bilingues du Web (publications en ligne, dictionnaires spécialisés, etc.), en le validant sur deux principales applications : la construction ou l'actualisation d'un lexique bilingue médical (chapitre 4) et la recherche d'information translangue (chapitre 6).

7.2 Synthèse

7.2.1 Acquisition lexicale bilingue

Il faut mentionner avant tout le niveau de difficulté que présente la recherche de la traduction d'un mot en corpus comparables. La traduction d'un mot dans un corpus comparable ne peut pas être repérée par sa position textuelle comme dans un corpus parallèle. La difficulté est essentiellement due à l'espace de recherche important qui ne peut pas être réduit comme dans le cas des corpus parallèles alignés au niveau des phrases. Dans ce dernier cas, l'espace de recherche de la bonne traduction est limité aux phrases traduites. Dans le cas de corpus comparables, l'espace de recherche est l'ensemble des mots du corpus en langue cible (section 2.4.1). Ainsi dans une des expériences menées au chapitre 4, l'objectif est pour un mot donné de trouver sa bonne traduction parmi plus de 70 000 candidats.

Plusieurs expériences ont été menées faisant varier différents paramètres et évaluant leur impact sur les performances. Nous passons en revue ici les principaux résultats.

Similarité croisée

Jusqu'à présent, tous les modèles développés dans le cadre de l'extraction de lexique à partir de corpus comparables exploitent la proximité distributionnelle dans une seule direction de langues (langue A $\rightarrow$ langue B) à l'aide d'une mesure de similarité que nous qualifions de classique. Nous avons proposé dans ce travail une nouvelle approche qui exploite les proximités distributionnelles des deux directions de langues (langue A $\leftrightarrow$ langue B) par l'intermédiaire d'une similarité que nous avons qualifiée de croisée.

Les expériences menées sur deux corpus du français vers l'anglais ont montré que l'application du calcul de la similarité croisée améliore les performances par rapport à celles du calcul de la similarité classique. Cette amélioration est encore plus importante lorsque la taille de corpus est plus importante. L'application de la similarité croisée permet d'extraire de façon significative plus de traductions correctes proposées au rang 1, par rapport à la similarité classique.

Le fait de prendre en compte les similarités calculées dans les deux directions de langues accorde la même importance aux distributions des mots dans les deux parties monolingues des corpus. Elle favorise ainsi les couples de mots ayant une similarité forte dans chaque direction (langue A $\leftrightarrow$ langue B) plutôt que les mots n'ayant une similarité forte que dans une seule direction.

Fréquence des mots

Les modèles développés à partir de l'analyse distributionnelle nécessitent un nombre minimum de cooccurrences afin de permettre le calcul statistique. En géné-

ral, l'évaluation de ces modèles est limitée aux mots les plus fréquents (Rapp, 1999; Déjean & Gaussier, 2002; Chiao & Zweigenbaum, 2002a). Or les mots inconnus des lexiques préexistants comme les néologismes, sont souvent peu fréquents dans les corpus et donc peu aptes à être traités par ces modèles.

Nous avons ainsi proposé, à côté de la présentation classique des résultats par rangs des candidats proposés, une présentation supplémentaire en fonction des tranches de fréquences homogènes des mots à traduire (section 4.3.3). L'intérêt d'une telle présentation est de pouvoir mesurer les performances d'une méthode sur les différentes tranches de fréquence. Les meilleures performances sont constatées pour la traduction des mots les plus fréquents. Néanmoins, il est intéressant de noter que notre modèle affiche une performance uniforme sur les mots appartenant aux tranches de fréquences moyennes. Les très basses fréquences restent peu performantes. La portée de notre modèle ne concerne donc pas uniquement les mots très fréquents mais également les mots de moyenne ou faible fréquence.

Taille des corpus

Dans les approches fondées sur l'analyse statistique de corpus, la taille des corpus est un paramètre décisif. L'idée sous-jacente est en effet que plus un corpus est de taille importante, mieux il peut faire ressortir les régularités entre un mot et sa traduction.

Les résultats ont confirmé l'apport positif de la taille des corpus. La performance de notre modèle sur les 30 premiers candidats pour l'actualisation de lexique médical [1] atteint 40% lorsque l'on utilise un corpus de 29 536 583 mots par rapport à 28% avec un corpus de 8 244 043 mots.

Autres paramètres évalués

Dans plusieurs expériences initiales, nous avons aussi montré l'effet de certains paramètres sur les performances de notre modèle. Nous avons ainsi constaté que l'ajout des mots généraux dans les vecteurs de contexte permet de mieux représenter le mot à traduire ainsi que sa traduction. Cela favorise alors l'appariement du vecteur d'un mot avec celui de sa traduction. Ainsi la performance sur les 20 premiers candidats atteint 100% lorsque l'on inclut les mots généraux dans le vecteur, tandis que 60% des mots ont été correctement traduits si l'on n'utilise que les mots spécialisés.

Deux tailles de fenêtre de contexte ont aussi été examinées. Les résultats obtenus pour l'extraction d'un lexique des mots fréquents ont montré l'intérêt d'utiliser une fenêtre réduite composée de 5 mots. 55% des mots ont été bien traduits parmi les 30 premiers candidats proposés contre 48% en utilisant une fenêtre de 7 mots. Il

1. Les résultats sont obtenus en appliquant le calcul de la similarité croisée avec la liste M (mots médicaux) comme lexique de référence et la liste U (mots hors lexique) comme candidats à la traduction (chapitre 4).

semble que la fenêtre de 5 mots favorise les relations syntagmatiques qui forment souvent des collocations. Pour le mot *sommeil*, nous trouvons par exemple *trouble, apnée, syndrome* comme trois premiers mots cooccurrants avec une fenêtre de 5 mots. Tandis qu'avec la fenêtre de 7 mots, nous trouvons *nuit, stress, activité*, comme cooccurrents significatifs.

Parmi les différentes combinaisons de mesures de similarité et de pondérations examinées, les résultats obtenus par *Jaccard* en combinaison avec *tf.idf* s'avèrent meilleurs que pour les autres combinaisons.

7.2.2 Reformulation de requête pour la recherche d'information translangue

Les expériences effectuées sur la collection OHSUMED ont pour but d'évaluer dans quelle mesure l'ensemble des traductions proposées par notre modèle peut être utilisé pour la recherche d'information translangue et quel effet il peut avoir sur les performances de la recherche. Deux séries d'expériences, dont l'une concerne la traduction de requêtes, et l'autre l'extension de requêtes, ont été réalisées.

Traduction de requête

L'ajout supervisé de la traduction d'un mot inconnu de la requête, proposée par notre modèle, a permis d'améliorer globalement les performances de la recherche. La supervision consiste à choisir la traduction parmi les candidats proposés. Les différentes mesures de la précision sont toutes améliorées : précision moyenne non interpolée, R-précision et précisions pour n documents trouvés. L'ajout automatique de la traduction d'un mot inconnu (première proposition de notre modèle) n'améliore pas la précision de manière significative. Cela semble introduire autant ou plus de bruit que de bonnes traductions.

Extension de requête

Les expériences sur l'extraction de lexique ont montré que les candidats à la traduction proposés par notre modèle sont souvent liés sémantiquement. Cela nous a amenée à appliquer notre modèle pour étendre les requêtes en ajoutant les premiers candidats proposés comme extension des mots inconnus. L'ajout de plusieurs candidats proches d'un mot inconnu de la requête dans un cadre supervisé améliore significativement les performances quelle que soit la mesure de précision utilisée. Ici aussi la supervision manuelle est nécessaire. Lorsqu'on ajoute automatiquement les candidats proposés sans filtrage manuel, les performances se dégradent.

Ces résultats s'inscrivent dans l'esprit dans lequel est développé notre modèle. Il s'agit de proposer un outil d'aide à la construction de lexique bilingue spécialisé ou à la recherche d'information translangue. Il propose des traductions candidates ou des

mots proches dans une autre langue. Ces propositions sont soumises au jugement de l'utilisateur. Ainsi de tels outils sont destinés à l'utilisateur possédant un minimum de connaissances du domaine et de compétences linguistiques sur la langue cible.

7.3 Discussion et perspectives

Plusieurs choix théoriques et pratiques effectués lors de ce travail ont fait l'objet d'explications et de justifications notamment au chapitre 1. Cette section propose une discussion autour de ces différents aspects et évoque aussi différentes pistes d'amélioration.

7.3.1 Comparabilité des corpus

La notion de comparabilité des corpus est une notion assez vague et peut prendre de multiples formes : comparabilité des vocabulaires, des genres, etc. A notre sens, la comparabilité des corpus bilingues doit être examinée sous l'angle de l'objectif visé et de la méthode qui les exploite et est liée à la notion de représentativité. Si on adopte le point de vue statistique, le corpus de domaine peut être considéré comme un échantillon de l'ensemble du discours du domaine en question. La représentativité d'un corpus est alors définie sur l'usage des mots de ce domaine. Mais cet ensemble est un ensemble idéal, inaccessible et trop peu contraignant. La constitution d'un corpus consiste en effet à choisir des critères d'appartenance à ce corpus. Un corpus est donc choisi le plus représentatif possible de l'usage des mots selon des critères définis. Il faut donc s'assurer que ces critères soient les mêmes pour les deux corpus comparables. Si ce n'est pas le cas, le décalage introduit va fausser la comparaison des mots. Choisir des critères, c'est introduire des contraintes de sélection qui à notre sens participent à la comparabilité des corpus.

La notion de comparabilité doit aussi être liée à l'objectif visé et à la méthode employée pour y parvenir. La question est de savoir ici quels sont les critères qui permettent de dire que deux corpus sont de bonnes ressources pour l'extraction de traductions. Cette remarque appliquée aux modèles basés sur l'analyse distributionnelle nous amène à proposer quatre critères de comparabilité.

Comparabilité de la couverture lexicale

Des corpus comparables doivent avoir des couvertures lexicales proches. En d'autres termes, leurs vocabulaires doivent être similaires (à la traduction près). Sachant que la comparaison des mots des deux corpus s'effectue sur des contextes communs contenus dans l'intersection des vocabulaires, un différentiel de couverture lexicale trop important limite ces contextes communs et ignore les contextes différents.

Comparabilité des fréquences relatives des mots

Des distributions de fréquence du vocabulaire très différentes dans des corpus de langues différentes peuvent influencer fortement les résultats.

Par exemple, dans notre modèle la similarité entre deux mots est plus importante si un mot de contexte commun est rare relativement à un corpus. Le système de pondération utilise en effet idf qui attribue plus d'importance aux mots rares qu'aux mots fréquents.

La notion de rareté est dépendante du corpus traité. Il se peut qu'un mot soit fréquent relativement à un corpus d'une langue et rare relativement à un autre corpus de l'autre langue, *e.g.* le cas du mot *plasmatique* qui est fréquent dans le corpus français utilisé dans nos expériences mais dont la traduction *plasmatic* est absente du corpus anglais. L'explication d'un tel phénomène est multiple. Il peut s'agir de réelles différences d'usage dans le domaine considéré. Il peut par contre s'agir d'un biais introduit dans un corpus. Un mot dans une langue peut être sous-représenté ou sur-représenté dans le corpus par rapport à son usage réel en considérant que l'usage réel est relatif aux contraintes de sélection choisies.

La comparabilité doit donc porter sur les fréquences des mots. Il s'agit ici de fréquences relatives (nombre d'occurrences d'un mot sur la taille du corpus) car il n'est pas nécessaire que les tailles des corpus soient comparables.

Comparabilité des cooccurrences relatives

Même si les vocabulaires de deux corpus sont proches et ont des fréquences relatives proches rien n'est dit sur leurs répartitions dans les corpus. Or l'approche générale se fonde sur la comparaison des distributions contextuelles qui s'appuie elle-même sur la comparaison des cooccurrences. Dans le cas idéal, si deux mots sont souvent (respectivement rarement) présents conjointement (au sens de la définition du contexte) dans un corpus, on s'attend à ce qu'ils soient souvent (respectivement rarement) présents conjointement dans un corpus comparable.

Ici aussi il s'agit de cooccurrences relatives. Il faut normaliser les cooccurrences pour qu'elles restent comparables d'un corpus à l'autre (Baayen, 2001).

Comparabilité des similarités distributionnelles entre mots

L'hypothèse de base sur laquelle est fondée l'approche générale est la proximité des distributions entre les langues. Cela peut aussi être un critère de comparabilité. En effet, le fait que certains mots soient proches de leur traduction au sens de la proximité de leurs distributions, laisse présager le même comportement sur les mots inconnus. A notre avis, la comparabilité doit être mesurée dans les deux directions de langues. Une mesure de comparabilité pourrait être d'ailleurs directement déduite de la similarité croisée.

Un critère serait par exemple que pour un ensemble de mots plus la similarité croisée met en évidence de bonnes traductions, plus les corpus sont comparables, sachant que dans l'idéal la similarité croisée trouverait les bonnes traductions pour l'ensemble des mots concernés. La vérification de ce critère est équivalente à la vérification de bonne traduction qui est justement la tâche pour laquelle on cherche un 'bon' corpus comparable. Cette comparabilité mesure le fait que les distributions contextuelles des mots sont positionnées de façon similaire les unes par rapport aux autres dans les deux corpus.

Vers des tests de comparabilité de corpus

Nous pensons que ces différents critères pourraient servir à effectuer un ensemble de tests sur un couple de corpus, qui permettraient d'évaluer leur aptitude à constituer une bonne ressource pour l'extraction de traductions. L'idée de ce type de test serait de mesurer l'écart de comparabilité entre deux corpus. La comparabilité maximale serait définie et atteinte par le parallélisme du corpus : le corpus parallèle étant le corpus comparable idéal. Dans ce cas idéal, tous les critères de comparabilité atteignent leur maximum : la couverture lexicale, les fréquences relatives, les cooccurrences relatives ou les distributions contextuelles.

Néanmoins, la comparabilité définie ainsi fait intervenir une ressource lexicale bilingue partielle qui y joue un rôle central. Une des questions délicates à traiter est de déterminer sur quels mots de cette ressource les critères doivent être étudiés : les mots généraux, spécialisés, non ambigus, etc. On pourrait ainsi tester la comparabilité sur des mots du lexique en espérant qu'elle se propage aux mots inconnus.

7.3.2 Contexte de cooccurrence

L'ajout de mots généraux dans le vecteur de contexte a également un impact sur les performances. Il améliore l'appariement des vecteurs. En fait, une explication est que l'utilisation d'un vecteur de taille plus grande (par ajout des mots généraux) permet de mieux représenter les mots et leurs traductions, ce qui favorise leur appariement. D'autres expériences sont envisageables en n'utilisant que les mots généraux dans la construction de vecteur de contexte.

Une fenêtre de n mots est souvent utilisée pour définir le contexte. En effet, différentes tailles de fenêtres permettent de mettre en valeur différents types de relations (sections 2.3.1, 3.2). Les deux tailles de fenêtres de contexte examinées ici ont donné des résultats similaires sauf pour l'extraction du lexique des mots fréquent dans laquelle la fenêtre plus petite de 5 mots améliore les performances par rapport à la fenêtre de 7 mots. En examinant les vecteurs construits avec la fenêtre plus petite, on observe plus de relations syntagmatiques comme N+N, N+Adj. Tandis que dans les vecteurs construits avec la fenêtre de 7 mots, nous trouvons plus des relations comme N+V, V+N.

Il serait intéressant d'effectuer d'autres expériences avec une fenêtre plus petite de 3 mots ou avec une fenêtre plus grande, *e.g.*, phrase, paragraphe. D'autre part, les bons résultats obtenus sur des contextes de taille de fenêtre réduite laissent penser que l'application des restrictions syntaxiques sur les contextes aurait des effets positifs. Cela nécessiterait néanmoins des corpus étiquetés.

7.3.3 Ressources lexicales partielles

La couverture et la qualité des ressources lexicales utilisées pour la traduction de vecteurs de contexte sont un des paramètres décisifs de la qualité de notre modèle. En effet, l'utilisation d'un lexique bilingue partiel est nécessaire dans les approches fondées sur les corpus comparables. Il permet d'assurer le passage d'une langue à une autre. Plus le lexique est complet, mieux il permet de représenter les mots par leur contextes, ce qui conditionne leur appariement.

Notons aussi un problème qui a perturbé l'évaluation des expériences d'acquisition lexicale. Certaines traductions erronées dans le lexique bilingue partiel ont invalidé des propositions pourtant correctes de notre modèle. Prenons l'exemple du mot français *trouble*, absent du lexique médical et dont la traduction proposée par le lexique général est *indistinct*. Les premiers candidats à la traduction proposés par notre modèle (*disorder, disease, syndrome*), contiennent la bonne traduction ou des synonymes en médecine, mais sont cependant considérés comme inconnus lors de l'évaluation.

7.3.4 Traitements linguistiques

Nous avons pris le parti de ne pas procéder à une analyse linguistique des corpus. Les seules analyses concernent la segmentation, l'éviction des mots vides ou grammaticaux et la mise au singulier. Or il est acquis que la richesse des informations linguistiques provenant des textes et la diversité des relations que les termes entretiennent entre eux en corpus peuvent servir à l'acquisition lexicale et à la recherche d'information. L'analyse linguistique de corpus permet l'exploitation de ces informations. Le rôle des techniques de traitement automatique des langues (aux différents niveaux : morphologique, syntaxique, sémantique et pragmatique) dans l'accès aux documents ont été étudiés notamment dans (Gaussier *et al.*, 2003).

La possibilité de travailler sur des corpus analysés et en particulier étiquetés est une voie d'amélioration de notre modèle. Un corpus étiqueté syntaxiquement permet en effet de faciliter le repérage de termes candidats. La mise en correspondance avec des mots d'une autre langue permet alors d'aligner un mot et sa traduction même si l'un est composé et l'autre simple.

L'étiquetage sémantique est une solution au problème de la polysémie qui n'est pas traité dans notre modèle. Or, la polysémie est à notre sens l'un des problèmes majeur de l'approche distributionnelle. Elle intervient à deux niveaux. Quand un mot de

contexte est polysémique, il intervient comme mot de contexte quelque soit sa signification. Ainsi deux mots sont tout de même rapprochés s'ils partagent la même graphie comme contexte mais avec des significations différentes. D'autre part, le vecteur de contexte d'un mot polysémique mélange les contextes de toutes ses significations. Le problème dans un cadre multilingue est que la polysémie dans une langue n'est en général pas vraie dans une autre langue. Les vecteurs de contexte de ses multiples traductions seront difficilement rapprochés de son vecteur de contexte.

La racinisation est un traitement linguistique qui peut également améliorer les résultats de l'approche distributionnelle. Elle permet de regrouper les mots d'une même famille flexionnelle et dérivationnelle. Dans nos expériences sur l'acquisition lexicale bilingue, beaucoup des candidats proposés pour un mot donné sont des variantes flexionnelles ou dérivationnelles de celui-ci (section 4.4). En recherche d'information, la racinisation permet d'augmenter les chances d'appariement entre une requête et un document car elle réduit le nombre des variantes sous lesquelles un mot peut apparaître dans un document. Les résultats des expériences sur la recherche d'information monolingue (français) et bilingue (anglais-français) dans (Gaussier *et al.*, 2000) ont montré l'amélioration des performances en utilisant les analyses morphologiques pour l'indexation des requêtes et des documents. La reconnaissance des entités nommées est un autre voix d'amélioration particulièrement importante dans le domaine médical. En effet, en médecine l'usage des entités nommées est fréquent. Elles servent à désigner des maladies, des médicaments, etc.

7.3.5 Retour sur la notion de la symétrie distributionnelle

L'exploitation de la symétrie distributionnelle pour l'acquisition lexicale bilingue est nous semble-t-il l'apport essentiel de ce travail. Revenons sur cette notion introduite au chapitre 1 (1.5.3). Dans un cadre bilingue, la symétrie distributionnelle postule que (pour certains mots) si un mot et sa traduction ont des distributions proches dans une direction de traduction alors ils ont également des distributions proches dans l'autre direction de traduction. La proximité des distributions est associée à une mesure de similarité. Pour chaque mot du premier corpus, cette mesure définit un ordre sur les mots du deuxième corpus et vice-versa.

Nous avons exploité la symétrie distributionnelle sur une partie de ces ordres, énonçant que si un mot est dans les n premiers rangs d'un ordre issu d'un autre mot et vice et versa (n = 30), les mots en question sont traductions l'un de l'autre. Une autre version de la symétrie distributionnelle serait d'examiner les ordres dans leur totalité en examinant non seulement les rangs de deux mots dans les ordres définis par une mesure de similarité mais aussi les rangs des autres mots (au lexique de transfert près). Par exemple, notre modèle examine le rang de 'liver' dans l'ordre des mots proches de 'foie' et vice-versa. On l'a montré, les rangs des deux mots peuvent être exploités pour mettre en évidence la relation de traduction qui les lie. Mais les autres mots proches ne sont pas exploités et la question qui se pose est de savoir si leur prise en compte est

intéressante ou non.

Une autre façon d'envisager la symétrie distributionnelle est de considérer les ordres établis par une similarité dans chacun des corpus pris séparément puis de considérer qu'ils sont là aussi identiques (Déjean & Gaussier, 2002). A ce stade nous ne pouvons affirmer si ces deux façons de procéder sont équivalentes, si l'une est plus puissante que l'autre ou si elles sont complémentaires.

Portée de la symétrie distributionnelle

La portée ou le champ de validité de la symétrie distributionnelle est mal définie à ce jour. Est-elle valide pour tous les mots ou partiellement ? Si sa validité est partielle, quels sont les critères qui définissent son champ de validité ? On a vu que la symétrie distributionnelle n'est plus vraie pour les mots polysémiques mais nous ne savons pas aujourd'hui si ce sont les seuls cas. D'autre part, il semble a priori qu'elle puisse être rétablie pour les mots polysémiques sur un corpus désambiguïsé.

Généralisation de la notion de symétrie distributionnelle

La symétrie distributionnelle peut être généralisée selon plusieurs axes. Elle peut être généralisée à plusieurs corpus comparables qu'ils soient dans deux langues différentes, dans la même langue ou dans plusieurs langues.

Dans le cadre monolingue, cette généralisation s'applique à des relations sémantiques symétriques comme la synonymie, la collocation ou la "traduction" des mots entre par exemple des niveaux de langue différents (langue technique vs langue grand public, langue familière vs langue soutenue). Notons que les critères de comparabilité définis précédemment ne restent pas nécessairement valides dans un cadre monolingue.

Dans le cadre multilingue, on peut imaginer traiter plus de deux langues. Il est probable en effet que l'hypothèse de la symétrie distributionnelle reste valide sur plus de deux langues. Le problème de l'acquisition lexicale en plus de deux langues est identique à celui de l'acquisition bilingue mais plus complexe à traiter. En dehors des difficultés linguistiques propres à chaque langue, la prise en compte de la symétrie distributionnelle entre n langues nécessite le calcul sur $2\frac{n!}{(n-2)!2!}$ directions de langues (nombre d'arrangements de 2 parmi n langues). Cela revient à calculer autant de similarités pour ensuite appliquer la similarité croisée sur les rangs résultants. Prenons le cas de trois langues, par exemple : le français, l'anglais et l'allemand. Il faut dans ce cas calculer 6 similarités classiques dans chaque direction de langues. Une autre option est de considérer la symétrie sur les langues deux à deux, donc de traiter isolément les couples de langues. L'intérêt de prendre en compte l'ensemble des langues simultanément est que si la symétrie distributionnelle interlangue s'avère une hypothèse valide, cela peut mettre en évidence directement des triplets valides. On peut espérer par exemple aligner simultanément les mots $foie$, $liver$, $leber$, à partir de trois corpus comparables.

C’est sur cette perspective que nous concluons. On le voit, la notion de symétrie distributionnelle ouvre des perspectives intéressantes notamment dans le cadre multilingue mais également monolingue. Cette notion demande encore à être précisée tant sur le plan théorique que sur le plan expérimental.

Bibliographie

ABNEY S. (1991). Parsing by Chunks. In S. A. ROBERT BERWICK & C. TENNY, Eds., *Principle-Based Parsing: Computation and Psycholinguistics*, p. 257–278, Boston: Kluwer Academic Publishers.

ARAÚJO M., NAVARRO G. & ZIVIANI N. (1997). Large Text Seraching allowing Errors. In *Proceedings of WSP'97*, p. 2–20, Valparaíso Chile: Carleton University Press.

ARNOLD D., BALKAN L., MEIJER S., HUMPHREYS R. L. & SADLER L. (1994). *Representation and Processing*, In *Machine Translation: an Introductory Guide*, chapter 3, p. 37–62. NCC Blackwell Ltd.

ASSADI H. (1997). Une méthode et des outils pour la construction d'une ontologie du domaine à partir de textes. Application à la consultation d'une documentation technique. In (IC, 1997), p. 363–374.

ATKINS S. (1990). Corpus Lexicography: the Bilingual Dimension. *Linguistica Computazionale*, **6**, 43–64. Special issue dedicated to Bernard Quemada.

BAAYEN R. H. (2001). *Word frequency distribution*. Number 18 in Text, Speech and Language Technology. Dordrecht: Kluwer Academic Publishers.

BAEZA-YATES R. & RIBEIRO-NETO B. (1999). *Modern Information Retrieval*. ACM Press.

BALLESTEROS L. & CROFT W. (1997). Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval. In *Proceedings of the 20th ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 84–91, Philadelphia, PA.

BALLESTEROS L. & CROFT W. (1998). Resolving Ambiguity for Cross-Language Retrieval. In *Proceedings of the 21th International ACM-SIGIR Conference Research and Development in Information Retrieval*, p. 64–71.

BALLESTEROS L. & CROFT W. B. (1996). Dictionary Methods for Cross-Lingual Information Retrieval. In *Proceedings of the 7th International DEXA Conference on Database and Expert Systems*, p. 791–801.

BENSON M. (1989). The structure of the collocational dictionary. *International Journal of Lexicography*, **2**, 1–14.

BODENREIDER O. & ZWEIGENBAUM P. (2000). Identifying proper names in parallel medical terminologies. In A. HASMAN, B. BLOBEL, J. DUDECK, R. ENGELBRECHT, G. GELL & H.-U. PROKOSH, Eds., *Medical Infobahn for Europe — Proceedings of MIE2000 and GMDS2000*, p. 443–447, Amsterdam: IOS Press.

BOITET C. (2001). Méthodes d'acquisition lexicale en TAO : des dictionnaires spécialisés propriétaires aux bases lexicales généralistes et ouvertes. In D. MAUREL, Ed., *Actes de TALN 2001 (Traitement automatique des langues naturelles)*, Tours: ATALA Université de Tours.

BOURIGAULT D. (1994). Extraction et structuration automatiques de terminologie pour l'aide à l'acquisition de connaissances à partir de textes. In *Actes du 9^e congrès Reconnaissance des Formes et Intelligence Artificielle - AFCET*, p. 1123–1132, Paris: AFCET.

BOURIGAULT D., GONZALEZ-MULLIER I. & GROS C. (1996). Lexter, a natural language processing tool for terminology extraction. In *Proceedings of the 7th EURALEX Internaional Congress on Lexicography*, p. 771–779, Goteborg.

BOURIGAULT D. & JACQUEMIN C. (1999). Term Extraction + Term Clustering: An Integrated Platform for Computer-Aided Terminology. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL'99)*, p. 15–22, Bergen.

BOURIGAULT D. & JACQUEMIN C. (2000). Term extraction and automatic indexing. In R. MITKOV, Ed., *Handbook of Computational Linguistics*. Amsterdam: John Benjamins.

BRASCHLER M., KRAUSE J., PETERS C. & SCHÄUBLE P. (1999). Cross-Language Information Retrieval (CLIR) Track Overview. In *Proceedings of the 7th Text REtrieval Conference*, p. 25–32, Washington DC.

BRAY T. (1996). Measuring the web. In *Proceedings of the 5th International World Wild Web Conference*, p. 993–1005, Paris.

BROWN P. F., COCKE J., PIETRA S. D., PIETRA V. J. D., JELINEK F., LAFFERTY J. D., MERCER R. L. & ROOSSIN P. S. (1990). A Statistical Approach to Machine Translation. *Computational Linguistics*, **16**(2), 79–85.

BROWN R. D. (1998). Automatically-Extracted Thesauri for Cross-Language IR: When Better is Worse. In *Proceedings of the First Workshop on Computational Terminology (COMPUTERM'98)*, Montreal, Canada.

CANCEDDA N., DÉJEAN H., ÉRIC GAUSSIER, RENDERS J.-M. & VINOKOUROV A. (2003). Report on CLEF-2003 experiments: two ways of extracting multilingual resources from corpora. In C. PETERS, Ed., *Proceedings of Cross Language Evaluation Forum (CLEF2003)*, Trondheim, Norway: Springer.

CARBONELL J., YANG Y., FREDERKING R., BROWN R. D., GENG Y. & LEE D. (1997). Translingual Information Retrieval: a Comparative Evaluation. In *Proceedings of Fifteenth International Joint Conference on Artificial Intelligence IJCAI-97*, volume I, p. 708–715, Nagoya, Japan.

CATACH N. (1963). Un point d'histoire de la langue : la bataille de l'orthographe aux alentours de 1900. *Le Français moderne*, **31**(2), 116.

CATIZONE R., RUSSELL G. & WARWICK S. (1989). Deriving Translation Data from Bilingual Texts. In *Proceedings of the First International Lexical Acquisition Workshop*, Detroit MI.

CAVNAR W. B. & TRENKLE J. M. (1994). N-gram based text categorization. In *Proceedings of the 3rd Annual Symposium on Document Analysis and Information retrieval*, p. 161–169.

CHEN H.-H. & BIAN G.-W. (1998). White page construction from web pages for finding people on the internet. *Computational Linguistics and Chinese Language Processing*, **3**(1), 77–100.

CHEN J. & NIE J.-Y. (2000). Parallel web text mining for cross-language ir. In *Proceedings of RIAO'2000 Content-Based Multimedia Information Access*, volume 1, p. 62–78, Paris.

CHIAO Y.-C. & STA J.-D. (2002). Recherche d'information multilingue et terminologie. In F. SEGOND, Ed., *Multilinguisme et traitement de l'information*, chapter 5. Hermès Science Publications.

CHIAO Y.-C., STA J.-D. & ZWEIGENBAUM P. (2004). A novel approach to improve word translations extraction from non-parallel, comparable corpora. In *Actes International Joint Conference on Natural Language Processing*, Hainan, China: AFNLP.

CHIAO Y.-C. & ZWEIGENBAUM P. (2002a). Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the American Medical Informatics Association 2002 Annual Symposium*, p. 150–154, San Antonio, Texas.

CHIAO Y.-C. & ZWEIGENBAUM P. (2002b). Looking for French-English translations in comparable medical corpora. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, volume 2, p. 1208–1212, Taipei, TAIWAN.

CHIAO Y.-C. & ZWEIGENBAUM P. (2003). The effect of a general lexicon in corpus-based identification of French-English medical word translations. In R. BAUD, M. FIESCHI, P. LE BEUX & P. RUCH, Eds., *The New Navigators: from Professionals to Patients,Actes Medical Informatics Europe*, volume 95 of *Studies in Health Technology and Informatics*, p. 397–402, Amsterdam: IOS Press.

CHIAO Y.-C. & ZWEIGENBAUM P. (2004). Aligning words in French-English non-parallel medical texts: Effect of term frequency distributions. In M. MUSEN, Ed., *Actes 10 th World Congress on Medical Informatics*, San Francisco, Ca. *À paraître*.

CHIEN L.-F. & PU H.-T. (1996). Important issues on chinese information retrieval. *Computational Linguistics and Chinese Language Processing*, **1**(1), 205–221.

CHOUEKA Y. (1988). Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In *Proceedings of the RIAO'88 Conference on User-Oriented Content-Based Text and Image Handling*, p. 609–623, Cambridge, MA.

CHOUEKA Y., KLEIN S. & NEUWITZ E. (1983). Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus. *Journal of the Association for Literary and Linguistic Computing*, **4**(1), 34–38.

CHURCH K. W. (1988). A stochastic parts program and noun phrase parser for unrestricted text. In *2nd Conference on Applied Natural Language Processing*, p. 136–143, Austin TE.

CHURCH K. W. (1993). Char_align: A program for aligning parallel texts at the character level. In *Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio.

CHURCH K. W., GALE W., HANKS P. & HINDLE D. (1991). Using statitics in lexical analysis. In U. ZERNIK, Ed., *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, p. 115–164. Hillsdale, NJ: Lawrence Erlbaum Ass.

CHURCH K. W. & HANKS P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, **16**(1), 22–29.

CLAVEAU V., SÉBILLOT P., BOUILLON P. & FABRE C. (2001). Acquérir des éléments du lexique génératif : quel résultats et à quel coûts? *Traitement Automatique des Langues*, **42**(3), 729–753.

CLEVERDON C., MILLS J. & KEEN M. (1996). *ASLIB CRANFIELD RESEARCH PROJECT: Design factors determining the performance of indexing systems*. Rapport interne, National Institute of Standards and Technology NIST.

CONDAMINES A. & REBEYROLLE J. (1997). Construction d'une BCT à partir de textes : expérimentation et définition d'une méthode. In (IC, 1997), p. 191–206.

CÔTÉ R. A. (1996). *Répertoire d'anatomopathologie de la SNOMED internationale, v3.4*. Université de Sherbrooke, Sherbrooke, Québec.

DAGAN I. & CHURCH K. (1997). Termight: Coordinating man and machine in bilingual terminology acquisition. *Machine Translation*, **12**(1-2), 89–107.

DAGAN I., CHURCH K. W. & GALE W. A. (1993). Robust bilingual word alignment for machine aided translation. In *Proceedings of the 1st Workshop on Very Large Corpora*, p. 1–8, Columbia.

DAILLE B. (1994). *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*. Rapport interne, Université de Paris 7. Thèse de Doctorat en Informatique Fondamentale.

DAILLE B. (1996). Study and implementation of combined techniques for automatic extraction of terminology. In J. KLAVANS & P. RESNICK, Eds., *The Balancing Act : Combining Symbolic and Statistical Approaches to Language*, p. 49–66. MIT Press.

DAILLE B. (1999). Identification des adjectifs relationnels en corpus. In P. AMSILI, Ed., *Actes de TALN 1999 (Traitement automatique des langues naturelles)*, p. 105–114, Cargèse: ATALA.

DAILLE B., GAUSSIER E. & LANGÉ J.-M. (1994). Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of the 15th International Conference on Computational Linguistics*, p. 515–521, Kyoto, Japan.

DAL G. & JACQUEMIN C. (1999). *FRANLEX. Base de données lexicales sur la morphologie dérivationnelle en français contemporain*. WWWpage http://m17.limsi.fr/Individu/jacquemi/FRANLEX/index.html, SILEX - LIMSI-CNRS. Visité le 17/08/99.

DALADIER A. (1990). Aspects constructifs des grammaires de Harris. *Langages*, (99), 57–84. A. Daladier (resp.).

DARMONI S. J., LEROY J.-P., THIRION B., BAUDIC F., DOUYERE M. & PIOT J. (2000). CISMeF: a structured health resource guide. *Methods of Information in Medicine*, **39**(1), 30–35.

DAVID S. & PLANTE P. (1990). *Termino Version 1.0*. Rapport de recherche, Centre d'Analyse de Textes par Ordinateur, Université du Québec, Montréal.

DAVIS M. W. (1998). On the Effective Use of Large Parallel Corpora in Cross-Language Information Retrieval. In (Grefenstette, 1998a), chapter 2, p. 11–23.

DAVIS M. W., DUNNING T. E. & OGDEN W. C. (1995). Text Alignment in the Real World: Improving Alignments of Noisy Translations Using Common Lexical Features, String Matching Strategies and N-Gram Comparisons. In *Proceedings of the Seventh Conference of the European Chapter of Association for Computational Linguistics*.

DAVIS M. W. & OGDEN W. (1997). QUILT: Implementing a Large-scale Cross-Language Text Retrieval System. In *Proceedings of the 20th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97)*, p. 92–98, Philadelphia, PA, USA: ACM.

DEERWESTER S., DUMAIS S. T., FURNAS G. W., LANDAUER T. K. & HARSHMAN R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, **41**(6), 391–407.

DEGOULET P. & FIESCHI M. (1991). *Traitement de l'information médicale. Méthodes et applications hospitalières*. Masson.

DRAPER S. (1998). Mizzaro's framework for relevance.

DUNNING T. E. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**(1), 61–74.

DÉJEAN H. & GAUSSIER E. (2002). Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica, numéro spécial sur Alignement lexical dans les corpus multilingues*, p. 1–22.

EICHMANN D. & RUIZ M. E. (1998). Cross-language information retrieval with the UMLS metathesaurus. In *Proceedings of the 21th ACM SIGIR*, p. 72–80.

ENGUEHARD C. (1994). Automatic natural acquisition of a terminology. In *Proceedings of the 2nd International Conference of Quantitative Linguistics (QUALICO'94)*, p. 83–88, Moscow.

ENGUEHARD C., MALVACHE P. & TRIGANO P. (1992). Indexation de textes : l'apprentissage automatique de concepts. In *Actes du XVème colloque international en linguistique informatique*, p. 1197–1202, Nantes.

FLUHR C. (2000). Indexation et recherche d'information textuelle. In J.-M. PIERREL, Ed., *Ingénierie des langues*, chapter 10.

FLUHR C., SCHMIT D., ORTET P., ELKATEB F., GURTNER K. & RADWAN K. (1998). Distributed crosslingual information retrieval. In (Grefenstette, 1998a), chapter 4, p. 42–50.

FOX E. A. (1980). Lexical relations: Enhancing effectiveness of information retrieval systems. *SIGIR Forum*, **15**(3), 5–36. DBLP, http://dblp.uni-trier.de.

FRANZ M., MCCARLEY J. S. & ROUKOS S. (1999). Ad hoc and Multilingual Information Retrieval at IBM. In *Proceedings of the 7th Text REtrieval Conference*, p. 104–115, Washington DC.

FUJII A. & ISHIKAWA T. (1999). Cross-Language Information Retrieval using Compound Word Translation. In *Proceedings of the 18th International Conference on Computer Processing of Oriental Languages (ICCPOL'99)*, p. 105–110.

FUNG P. (1998). A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-parallel Corpora. In L. G. DAVID FARWELL & E. H. HOVY, Eds., *Proceedings of the third conference of the Association for Machine Translation in the Americas (AMTA'98)*, volume 1529, p. 1–17, Langhorne, PA: Springer.

FUNG P. & MCKEOWN K. (1997). Finding Terminology Translations from Non-Paralle Corpora. In *Proceedings of the 5th Annual Workshop on Very Large Copora*, volume 1, p. 192–202, Hong Kong.

FUNG P. & YEE L. Y. (1998). An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics*, volume 1, p. 414–420, Montreal.

GARCIA D. (1998). *Analyse automatique des textes pour l'organisation causale des actions. Réalisation du système informatique COATIS*. Rapport interne, Université de Paris-Sorbonne 4. Thèse de Doctorat en Informatique.

GAUSSIER E. (1998). Flow network models for word alignment and terminology extraction from bilingual corpora. In C. BOITET & P. WHITELOCK, Eds., *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*, p. 444–450, San Francisco, CA: Morgan Kaufmann Publishers.

GAUSSIER E., GREFENSTETTE G., HULL D. & ROUX C. (2000). Recherche d'information en français et traitement automatique des langues. *Traitement automatique des langues*, **41**(2), 473–493.

GAUSSIER E., JACQUEMIN C. & ZWEIGENBAUM P. (2003). Traitement automatique des langues et recherche d'information. In ÉRIC GAUSSIER & M.-H. STEFANINI, Eds., *Assistance intelligente à la recherche d'informations*, chapter 2, p. 71–96. Paris: Hermès-Lavoisier.

GAUSSIER E. & LANGÉ J.-M. (1995). Modèles statistiques pour l'extraction de lexiques bilingue. *Traitement automatique des langues*, **36**(1-2), 133–156.

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

GREFENSTETTE G. (1994b). *Explorations in Automatique Thesaurus Discovery*. Boston/Dordrecht/London: Kluwer Academic Publishers.

GREFENSTETTE G. (1996). Evaluation techniques for automatic semantic extraction : comparing syntactic and window based approaches. In B. BOGURAEV & J. PUSTEJOVSKY, Eds., *Corpus Processing for lexical acquisition*, Language, Speech and Communication, chapter 11, p. 205–216. Cambridge, Massachusetts: MIT Press.

GREFENSTETTE G. (1998a). *Cross-Language Information Retrieval*. London: Kluwer Academic Publishers.

GREFENSTETTE G. (1998b). The problem of cross-language information retrieval. In (Grefenstette, 1998a), p. 1–9.

GREFENSTETTE G. & NIOCHE J. (2000). Estimation of English and non-English language use on the WWW. In *Proceedings of RIAO 2000: Content-Based Multimedia Information Access*, p. 237–246, Paris, France: C.I.D.

GRUNDY V. (1996). L'utilisation d'un corpus dans la rédaction du dictionnaire bilingue. In B. H. & T. P, Eds., *Les dictionnaires bilingues*, p. 127–149. Louvain-la-Neuve, Duculot.

HABERT B. (1998). *Des mots complexes possibles aux mots complexes existants : l'apport des corpus*. Rapport interne, Université Lille III - Charles de Gaulle. Habilitation à diriger des recherche en linguistique.

HABERT B., GRABAR N., JACQUEMART P. & ZWEIGENBAUM P. (2001). Building a text corpus for representing the variety of medical language. In *Corpus Linguistics 2001*, Lancaster.

HABERT B. & JACQUEMIN C. (1993). Noms composés, termes, dénominations complexes : problématiques linguistiques et traitement automatiques. *Traitement automatique des langues TAL*, **34**(2), 5–42.

HABERT B. & JARDINO M. (2003). Compte rendu de R. Harald Baayen, Word Frequency Distribution. *Traitement automatique des langues*, **43**(2), 209–211.

HABERT B., NAULLEAU E. & NAZARENKO A. (1996). Symbolic word clustering for medium-size corpora. In J.-I. TSUJII, Ed., *Proceedings of the 16 th COLING*, p. 490–495, Copenhagen, Denmark.

HABERT B., NAZARENKO A. & SALEM A. (1997). *Les linguistiques de corpus*. Paris: Armand Colin.

HALLIDAY M. & HASAN R. (1976). *Cohesion in English*. English Language Series. London: Longman.

HARMAN D. (1991). How effective is suffixing. *Journal of the American Society for Information Science*, **42**, 7–15.

HARPER D. J. & VAN RIJSBERGEN C. (1978). An evaluation of feedback in document retrieval using co-occurrence data. *Journal of Documentation*, **34**(3), 189–216.

HARRIS Z., GOTTFRIED M., RYCKMAN T., MATTICK JR P., DALADIER A., HARRIS T. & HARRIS S. (1989). *The Form of Information in Science, Analysis of Immunology Sublanguage*, volume 104 of *Boston Studies in the Philosophy of Science*. Dordrecht & Boston: Kluwer Academic Publishers.

HARRIS Z. S. (1988). *Language and information*. New York: Columbia University Press.

HARRIS Z. S. (1991). *A theory of language and information. A mathematical approach*. Oxford: Oxford University Press.

HEAPS H. S. (1978). Information Retrieval: Computational and Theoretical Aspects.

HEARST M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In A. ZAMPOLLI, Ed., *Proceedings of the 14 th COLING*, p. 539–545, Nantes, France.

HERSH W., BUCKLEY C., LEON T. & HICKAM D. (1994). OHSUMED: an interactive retrieval evaluation and new large test collection for research. In *Proceedings of the 17th Annual ACM SIGIR Conference*, p. 192–201.

HERSH W. R., BALL A., DAY B., MASTERSON M., ZHANG L. & SACHEREK L. (1999). Maintaining a catalog of manually-indexed, clinically-oriented World Wide Web content. *Journal of the American Medical Informatics Association*, **6**(suppl), 790–794.

HIEMSTRA D. (1998). Multilingual domain modeling in Twenty-One: automatic creation of a bi-directional translation lexicon from a parallel corpus. In H. V. H. PETER-ARNO COPPEN & L. TEUNISSEN, Eds., *Proceedings of the eightth CLIN meeting*, p. 41–58.

HIEMSTRA D. & DE JONG F. (1999). Disambiguation strategies for cross-language information retrieval. In *European Conference on Digital Libraries*, p. 274–293.

HIEMSTRA D., DE JONG F. & KRAAIJ W. (1997). A domain specific lexicon acquisition tool for cross-linguage information retrieval. In L. DEROYE & C. CHRISMENT, Eds., *Proceedings of RIAO97 Conference on Computer-Assisted Searching on the Internet*, p. 217–232, Montreal, Canada.

HINDLE D. (1989). Acquiring disambiguation rules from text. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*.

HULL D. & GREFENSTETTE G. (1996). Querying Across Languages: A Dictionary-Based Approach to Multilingual Information Retrieval. In H.-P. FREI, D. HARMAN, P. SCHÄBLE & R. WILKINSON, Eds., *Proceedings of the 19th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96)*, p. 49–57, Zurich, Switzerland: ACM. Special Issue of the SIGIR Forum.

HULL D. A. (1997). Using structured queries for disambiguation in cross-language information retrieval. In *AAAI Symposium on CrossLanguage Text and Speech Retrieval*: American Association for Artificial Intelligence.

IC (1997). *Actes des 1es Journées Ingénierie des Connaissances*, Roscoff, France.

IRAM S., OHTAKE K., MASUYAMA S. & YAMAMOTO K. (1999). Identifying Translations of Compound Nouns Using Non-aligned Corpora. In *Proceedings of the Workshop MAL'99*, p. 108–113.

JACQUEMIN C. (1997a). Guessing morphology from terms and corpora. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97)*, p. 156–167, Philadelphia, PA.

JACQUEMIN C. (1997b). *Variation terminologique : Reconnaissance et acquisition automatique de termes et de leurs variantes en corpus*. Mémoire d'habilitation à diriger des recherches, Université de Nantes.

JACQUEMIN C. & ZWEIGENBAUM P. (2000). Traitement automatique des langues pour l'accès au contenu des documents. In J. LE MAÎTRE, J. CHARLET & C. GARBAY, Eds., *Le document en sciences du traitement de l'information*, chapter 4, p. 71–109. Toulouse: Cepadues.

JANG M., MYAENG S. H. & PARK S. Y. (1999). Using mutual information to resolve query translation ambiguities and query term weighting. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, p. 223–229.

KAJI H., KIDA Y. & MORIMOTO Y. (1992). Learning translation templates from bilingual text. In *Proceedings of the 14th International Conference on Computational Linguistics*, p. 672–678, Nantes.

KAVANAGH J. (1995). The text analyzer: A tool for extracting knowledge from text. Master's thesis, University of Ottawa. Disponible à http://www.site.uottawa.ca/~kavanagh/Thesis/.

KITTREDGE R. & LEHRBERGER J. (1982). *Sublanguage : Studies of Language in Restricted Domains*. New York: Walter de Gruyter.

KLAVANS J. & TZOUKERMANN E. (1990). The BICORD System: Combining Lexical Information from Bilingual Corpora and Machine Readable Dictionaries. In H. KARLGREN, Ed., *Proceedings of the 13th International Conference on Computational Linguistics*, volume 3, p. 174–179, Helsinki, Finland.

KRAFT D. H. & BUELL D. A. (1983). Fuzzy sets and generalized boolean retrieval systems. *International Journal of Man-Machine Studies*, **19**(1), 45–56.

KRAIF O. (2001a). *Constitution et exploitation de bi-textes pour l'aide à la traduction*. Rapport interne, Université de Nice Sophia Antipolis. Thèse de Doctorat en Sciences du Langage.

KRAIF O. (2001b). Exploitation des cognats pour l'alignement : architecture et évaluation. *Traitement automatique des langues*, **42**(3).

KUPIEC J. (1993). An Algorithm for finding Noun Phrase Correspondences in Bilingual Corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, p. 17–22, Columbus, Ohio: Ohio State University.

LEBART L. & SALEM A. (1988). *Analyse statistique des données textuelles : questions ouvertes et lexicométrie*. Paris: Dunod.

LEBART L. & SALEM A. (1994). *Statistique textuelle*. Paris: Dunod.

LIN D. (1997). Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of ACLEACL-97, Meeting of the Association for Computational Linguistics*, p. 64–71, Madrid.

LIN D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL'98)*, p. 768–774, Montreal.

LITTMAN M. L., DUMAIS S. T. & LANDAUER T. K. (1998). Automatic Cross-Language Information Retrieval using Latent Semantic Indexing. In (Grefenstette, 1998a), chapter 5, p. 51–62.

LOSEE R. M. (1998). *Text Retrieval and Filtering: Analytic Models of Performance*, volume 3 of *Information Retrieval*. Dordrecht & Boston: Kluwer Academic Publishers.

MANN G. S. & YAROWSKY D. (2001). Multipath Translation Lexicon Induction via Bridge Language. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

MANUILA A., MANUILA L., NICOLE M. & LAMBERT H. (1970). *Dictionnaire Français de Médecine et de Biologie*, volume 1–4. Paris: Masson & Cie.

MARCUS M. P., SANTORINI B. & MARCINKIEWICZ M. A. (1994). Building a large annotated corpus of English: The Penn Treebank. In S. ARMSTRONG, Ed., *Using Large Corpora*, p. 273–290, Cambridge: MIT Press.

MATSUMOTO Y., ISHIMOTO H. & UTSURO T. (1993). Structural matching of parallel texts. In *Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics*, p. 23–30, Columbus.

MCKEOWN K. R. & RADEV D. R. (2000). *Collocations*, In *A Handbook of Natural Language Processing*, chapter 15. Marcel Dekker.

MELAMED I. D. (1995). Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons. In *Proceedings of the 3rd Workshop on Very Large Corpora*.

MELAMED I. D. (2000). Models of translational equivalence among words. *Computational Linguistics*, **26**(2), 221–249.

MILLER G., BECKWITH R., FELLBAUM C., GROSS D. & MILLER K. (1990). *Five papers on WordNet*. Rapport interne, Cognitive Science Laboratory, Princeton University.

MIZZARO S. (1997). Relevance: the whole history. *Journal of the American Society for Information Science*, **48**(9), 810–832.

MIZZARO S. (1998). How many relevances in information retrieval? *Interacting with Computers*, **10**(3), 305–322.

MORIN E. (1998). Prométhée : un outil d'aide à l'acquisition de relations sémantiques entre termes. In (Zweigenbaum, 1998), p. 172–181.

MULLER C. (1997). *Principes et méthodes de statistique lexicale*. Coll. Langue, Linguistique, Communication. Paris: Hachette Université.

NAZARENKO A. & HAMON T. (2002). Structuration de terminologie : quels outils pour quelles pratiques ? *Traitement automatique des langues*, **43**(1), 7–18.

NEWMAN P. (1987). Foreign language identification: First step in translation process. In *Proceedings of the 28th Annual Conference of the American Translators Association*, p. 509–516.

NIE J.-Y. (2001). A general logical approach to inferential information retrieval. *Encyclopedia of Computer Science and Technology*, **44**, 203–226.

NIE J.-Y., SIMARD M. & FOSTER G. (2000). Multilingual information retrieval based on parallel texts from the web, Cross-Language Information Retrieval and Evaluation. In C. PETERS, Ed., *Proceedings of Cross Language Evaluation Forum, CLEF2000*, p. 188–201, Lisbon: Springer.

NLM (2000). *UMLS Knowledge Sources Manual*. National Library of Medicine, Bethesda, Maryland, 11th edition. www.nlm.nih.gov/research/umls/.

OARD D. W. (1998). A comparative study of query and document translation for cross-language information retrival. In *Proceedings of 3rd Conference of the Association for Machine Translation in America*, p. 478–483.

OARD D. W. & DORR B. J. (1996). *A Survey of Multilingual Text Retrieval*. Technical Report UMIACS-TR-96-19, Institute for Advanced Computer Studies and Computer Science Department, University of Maryland, College Park, MD.

OGAWA Y., MORITA T. & KOBAYASHI K. (1991). A fuzzy document retrieval system using the keyword connection matrix and a learning method. *Fuzzy Sets and Systems*, **39**, 163–179.

OUESLATI R. (1999). *Aide à l'acquisition de connaissances à partir de corpus*. Rapport interne, Université Louis Pasteur Strasbourg. Thèse de Doctorat en Informatique.

OZDOWSKA S. (2004). Appariement bilingue de mots par propagation syntaxique à partir de corpus français/anglais alignés. In *Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL)*, Fèz, Maroc: Laboratoire Parole et Langage Aix-en-Provence, France.

PALMER D. & BURGER J. (1997). Chinese word segmentation and information retrieval. In *AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*. Electronic Working Notes.

PEARSON J. (1999). Comment accéder aux éléments définitoires dans les textes spécialisés. *Terminologie et intelligence articfielle, actes du colloque de Nantes; Rint Réseau international de néologie et de terminologie*, (19), 21–28.

PETER F. BROWN, STEPHEN DELLA PIETRA V. J. D. P. & MERCER R. L. (1994). The Mathematic of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, **19**(2), 263–311.

PETERS C. & SHERIDAN P. (2001). Accès multilingue aux système d'information. In *Workshop of the 67th IFLA Council and General Conference*, Boston.

PETRAS V., PERELMAN N. & GEY F. C. (2003). UC Berkeley at CLEF 2003 - Russian Language Experiments and Domain-Specific Cross-Language Retrieval. In C. PETERS, Ed., *Working notes for the Workshop of Cross-Language Evaluation Forum CLEF2003*, p. 116–128, Trondheim, Norway: Springer-Verlag Berlin Heidelberg.

PICCHI E. & PETERS C. (1998). Cross-language information retrieval: A system for comparable corpus querying. In (Grefenstette, 1998a), chapter 7, p. 81–90.

POULIQUEN B. (2002). *Indexation de textes médicaux par extraction de concepts, et son utilisation*. Rapport interne, Université de Rennes I. Thèse de Doctorat en Génie Biologie et Médicale.

QIU Y. & FREI H. (1993). Concept based query expansion. In *Proceedings of 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*.

RAJMAN M., BESANÇON R. & CHAPPELIER J.-C. (2000). Le modèle DSIR : une approche à base de sémantique distributionnelle pour la recherche documentaire. *Traitement Automatique des Langues*, **41**(2/2000).

RAJMAN M. & BONNET A. (1992). Corpora-base linguistics: new tools for natural language processing. In *Proceedings of the 1st Annual Conference of the Association for Global Strategic Information*, Bad Kreusnach, Germany.

RAPP R. (1995). Indentifying word translation in non-parallel texts. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, student session*, volume 1, p. 321–322, Boston, Mass.

RAPP R. (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, College Park, MD.

RESNIK P. (1993). *Selection and Information: A Class-based Approach to Lexical Relationships*. PhD thesis.

RESNIK P. (1995). Disambiguating noun grouping with respect to wordnet senses. In *Proceedings of the 3th Workshop on Very Large Corpora*, Cambridge, USA.

RESNIK P. (1999). Mining the web for bilingual text. In *Proceedings of the International Conference of the Association of Computational Linguistics*, Maryland.

RESNIK P. & MELAMED I. (1997). Semi-automatic acquisition of domain-specific translation lexicons. In *Proceedings of the 7th ACL Conference on Applied Natural Language Processing*, Washington, DC.

RILOFF E. (1993). Automatically contructing a dictionary for information extraction tasks. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, p. 811–816: AAAI Press/MIT Press.

ROBERTS R. P. & MONTGOMERY C. (1996). The use of corpora in bilingual lexicography. In *Proceedings of the Seventh EURALEX International Congress on Lexicography*, p. 457–464, Göteborg: Göteborg University.

ROBERTSON S., WALKER S., JONES S., HANCOCK-BEAULIEU M. M. & GATFORD M. (1994). Okapi at TREC3. In D. K. HARMAN, Ed., *NIST Special Publication 500-226: the Third Text REtrieval Conference TREC-3*, p. 109–126, Gaithersburg, Maryland: Department of Commerce, National Institute of Standards and Technology.

ROBERTSON S. E. & SPARCK JONES K. (1976). Relevance weighting of search terms. *Journal of American Society for Information Science*, p. 129–146.

ROCCHIO J. (1971). Relevance feedback in information retrieval. In G. SALTON, Ed., *SMART Retrieval System*, chapter Experiments in Automatic Document Processing, p. 313–323. Englewood Cliffs, New Jersey: Prentice Hall, Inc.

ROMESBURG H. C. (1990). *Cluster Analysis for Researchers*. Malabar, FL: Krieger.

SAGER N. (1987). Information formatting of medical literature. In N. SAGER, C. FRIEDMAN & M. S. LYMAN, Eds., *Medical Language Processing: Computer Management of Narrative Data*, chapter 10, p. 197–220. Reading, MA: Addison Wesley.

SALTON G. (1969). A new comparison between conventional indexing (MEDLARS) and automatic text processing (SMART). *Journal of Americain Society for Information Science*.

SALTON G. (1970). Automatic processing of foreign language documents. *Journal of the American Society for Information Science*, **21**(3), 187–194.

SALTON G. & BUCKLEY C. (1988). Term-weighting approaches in auomatic retrieval. *Information Processing and Management*, **5**(24), 513–528.

SALTON G. & BUCKLEY C. (1990). Improving retrieval performance by relevance feedback. *Journal of American Society for Information Science*, **4**(41), 288–297.

SALTON G., FOX E. A. & WU H. (1983). Extended boolean information retrieval. In *Communications of the ACM*, volume 26, p. 1022–1036.

SALTON G., WONG A. & YANG C. (1974). A vector space model for automatic indexing. *Cornell University Press*.

SAVOY J. (2002). *Morphologie et recherche d'information*. Cahier de recherche en informatique CR-I-2002-01, Faculté de Droit et des Science Économiques, Université de Neuchatel, Suisse.

SCHÜTZE H. & PEDERSEN J. O. (1997). A cooccurrence-based thesaurus and two applications to information retrieval. *Information Processing and Management*, **33**(3), 307–317.

SHANNON C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, **27**, 379–423.

SHERIDAN P., BALLERINI J. P. & SCHÄUBLE P. P. (1998). Building a large multilingual test collection from comparable news documents. In (Grefenstette, 1998a), chapter 11, p. 137–149.

SINCLAIR J. (1991). *Corpus, Concordance, Collocation*. Oxford University Press.

SINCLAIR J. (1996). *Preliminary recommandations on corpus typology*. Rapport interne, EAGLES (Expert Advisory Group on Language Engineering Standards, Pisa.

SINCLAIR J., HANKS P., FOX G., MOON R. & P.STOCK (1987). *Collins COBUILD English Language Dictionary*. Glasgow: Collins.

SMADJA F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics, Special Issue on Using Large Corpora*, **19**(1), 1–38.

SMADJA F., MCKEOWN K. R. & HATZIVASSILOGLOU V. (1996). Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, **22**(1), 1–38.

SPARCK JONES K. (1979). Experiments in relevance weighting of search terms. *Information Processing and Management*, **15**, 133–144.

SPARCK JONES K. & VAN RIJSBERGEN C. J. (1975). Report on the need for and provision of an "ideal" information retrieval test collection. In *British Library Research and Development Report*, number 5266.

STA J.-D. (1993). Information filtering: a tool for communication be tween researchers. In *Proceedings of INTERCHI'93*, Amsterdam.

STA J.-D. (1997). *Acquisition terminologique en corpus : aspects linguistiques et statistique*. Rapport interne, Université Paris 7. Thèse de Doctorat en Linguistique Informatique.

STA J.-D. & CHIAO Y.-C. (2001). Knowledge acquisition from a text by a linguistic and statistical method. In *Proceedings of the 7th International Workshop on Parsing Technologies*, Bejing, China.

SÉGUÉLA P. & AUSSENAC N. (1999). Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine. In R. TEULIER, Ed., *Actes de IC'99*.

TEUBERT W. (2001). Corpus Linguistics and Lexicography. *Text Corpora and Multilingual Lexicography: Special Issue of International Journal of Corpus Linguistics*, **6**(1), 125–154.

TOUSSAINT Y., NAMER F., DAILLE B., JACQUEMIN C., ROYAUTÉ J. & HATHOUT N. (1998). Une approche linguistique et statistique pour l'analyse de l'information en corpus. In (Zweigenbaum, 1998), p. 182–191.

VAN DER EIJK P. (1993). Automating the acquisition of bilingual terminology. In *Proceedings of the 6th Conference of the European Chapter of the ACL (EACL'93)*, p. 113–119, Utrecht, Netherland.

VAN RIJSBERGEN C. J. (1979). *Information retrieval*. London: Butterworths edition.

VÉRONIS J. (2000a). From Rosetta stone to the information society: A survey of parallel text processing. In (Véronis, 2000b), chapter 1, p. 1–24.

VÉRONIS J. (2000b). *Parallel Text Processing. Alignment and Use of Translation Corpora*, volume 13 of *Text, Speech and Language Technology serie*. Kluwer Academic Publishers, nancy ide and jean véronis edition.

VÉRONIS J. & IDE N. (1991). An assessment of semantic information automatically extracted from machine readable dictionaries. In *Proceedings of the 5 th EACL*, p. 227–232, Berlin, Germany.

VÉRONIS J. & LANGLAIS P. (2000). Evaluation of parallel text alignment systems: ARCADE. In (Véronis, 2000b), chapter 19.

WILLETT P. (1981). A fast procedure of the calculation of similarity coefficients in automatic classification. *Information Processing and Management*, **17**, 53–60. academic press.

WONG S., ZIARKO W. & WONG P. (1985). Generalized vector space model in information retrieval. In *Proceedings of the 8th ACM SIGI Conference on Research and Development in Information Retrieval*, p. 18–25, N.Y. USA.

WU D. (1997). Stochastic Inversion Transduction Grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, **23**(3), 377–404.

WU D. (2000). Bracketing and aligning words and constituents in parallel text using Stochastic Inversion Transduction Grammars. In (Véronis, 2000b), chapter 7.

WU D. & XIA X. (1994). Learning an English-Chinese Lexicon from a parallel corpora. In *Proceedings of the first conference of the Association for Machine Translation in the Americas (AMTA'94)*, p. 206–213, Columbia.

XU J. & CROFT W. B. (1996). Query Expansion Using Local and Global Document Analysis. In H.-P. FREI, D. HARMAN, P. SCHÄUBLE & R. WILKINSON, Eds., *Proceedings of the 19th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96)*, p. 4–11, Zurich, Switzerland: ACM. Special Issue of the SIGIR Forum.

YANG D.-H., GOMEZ P. C. & SON M. (2000). An algorithm for predicting the relationship between lemmas and corpus size. *Journal of Electronics and telecommunications research institute*, **22**(2), 20–31.

YANG Y., CARBONELL J. G., BROWN R. D. & FREDERKING R. E. (1998). Translingual information retrieval: Learning from bilingual corpora. *Artificial Intelligence Journal*, **103**(1-2), 323–345. Special Issue: Best of IJCAI-97.

ZIGLER D. (1991). *The Automatic Identification of Languages using Linguistic Recognition Signals*. PhD thesis, State University of New York, Buffalo.

ZIPF G. K. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley.

P. ZWEIGENBAUM, Ed. (1998). *Actes de TALN 1998 (Traitement automatique des langues naturelles)*, Paris. ATALA.

ZWEIGENBAUM P., BAUD R., BURGUN A., NAMER F., ÉRIC JARROUSSE, GRABAR N., RUCH P., DUFF F. L., THIRION B. & DARMONI S. (2003). UMLF : construction d'un lexique médical francophone unifié. In *Actes des 10 Journées Francophones d'Informatique Médicale*, Tunis.

ÉMILE BENVENISTE (1966). Formes nouvelles de la composition nominale. *Bulletin de la Société de Linguistique de Paris*, **LXI**(1), 82–95.

Annexe A

Annexe : Expériences d'extraction lexicale bilingue spécialisée

A.1 Texte extrait du corpus CISMeF

Régénération du foie. Le foie est un organe doué d'une fantastique capacité à régénérer. Ainsi, si on réalise une hépatectomie (ablation) de 70% de la masse hépatique chez le rat, il existe une récupération intégrale en 7 à 10 jours. Sur un foie au repos, on observe très peu de mitose spontanée dans les hépatocytes (environ 1 mitose pour 20 000 cellules). Après hépatectomie, chaque hépatocyte peut se diviser 1 à 2 fois permettant ainsi la récupération de la masse hépatique. On a pu réaliser jusqu'à 12 hépatectomies itératives et observer une régénération sans aucun problème. Un seul hépatocyte peut donc se diviser de façon itérative pour former 50 foies complets. Ces chiffres montrent la fantastique capacité de régénération du foie. La pratique de transplantation hépatique illustre parfaitement l'adaptation du foie à son environnement. Ainsi, le foie d'un gros chien transplanté chez un petit chien diminue de taille pour se conformer à son nouvel environnement. Il en est de même chez l'homme où des transplantations de foie de babouin se sont suivies d'une adaptation de la taille du viscère transplanté. Ceci montre l'existence de mécanismes de régulation de la croissance hépatique pour que la taille de ce viscère s'adapte parfaitement à son hôte . Les mécanismes de la régénération hépatique sur un foie sain Dans ce cas , la régénération se fait à partir des hépatocytes restant qui se divisent de façon itérative. Les hépatocytes quiescents sont en phase G0. Le TNF $\hat{I}\pm$ (Tumor Necrosis Factor alpha) qui est sécrété par les cellules de Küpffer permet de faire entrer l'hépatocyte dans le cycle cellulaire. Il existe au moins 70 gènes différents qui sont activés rapidement après l'hépatectomie. Ces gènes constituent les immediate early genes dont les protéines co rrespondantes sont très variées (facteurs de transcription, tyrosine phosp hatase, protéines sécrétées, protéines cellulaires.). Il est bien sûr hors de propos de présenter tous ces gènes. Nous parlerons de deux facteurs de transcription : NFϰB et STAT3. NFϰB (pour Nuclear Factor fo r the ϰ chain of B cells) est activé 30 minutes après l'hépatectomie et ceci pendant 4 à 5 heures. STAT3 est activé 1 à 2 heures après l'hépa tectomie et ceci pendant 4 à 6 heures.

FIG. A.1 – *Extrait du document indexé par le terme **foie** : www.chups.jussieu.fr/polys/histo/histoApp/POLY.Chp.1.3.html*

A.2 Texte extrait du corpus CLINIWEB

REGENERATION-ASSOCIATED SERPIN-1 The mammalian liver has an extraordinary capacity for regeneration. In the rat, the liver regenerates the most of its original mass within several days following hepatectomy; regeneration is virtually complete by 2 weeks after surgery. Studies of gene expression after hepatectomy have focused on events occurring the first few hours after surgery. For example, the augmenter of liver regeneration ALR (600924) was isolated in this way. New et al. (1996) isolated a gene encoding a plasma protein by constructing and screening a cDNA library with RNA isolated from liver at 48 hours after 70 to 90% hepatectomy. New et al. (1996) stated that the expression of acute phase inflammatory proteins should be substantially diminished, thereby reducing the 'background' and facilitating the identification of genes associated with regeneration. They identified several clones that were upregulated in the regenerating liver. They isolated 1 clone, termed 'regeneration-associated serpin-1' (RASP1), that was expressed in normal liver but was upregulated approximately 3- to 4-fold by 48 hours after hepatectomy. DNA sequence analysis showed that the RASP1 gene encodes a novel 436-amino acid secreted protein. Moderate homology was found with several members of the serpin family of serine-protease inhibitors. The 1.7-kb RASP1 mRNA was highly expressed in rat liver, but not in brain, heart, kidney, lung, testis, or spleen. It was found in normal and hepatectomy rat plasma. REFERENCES 1. New, L.; Liu, K.; Kamali, V.; Plowman, G.; Naughton, B. A.; Purchio, A. F.: cDNA cloning of rasp-1, a novel gene encoding a plasma protein associated with liver regeneration. Biochem. Biophys. Res. Commun. 223: 404-412, 1996. PubMed ID : 8670294.

FIG. A.2 – *Extrait d'un document indexé par le terme* ***liver*** *: http://www.ncbi.nlm.nih.gov:80/entrez/dispomim.cgi?id=602455*

A.3 Liste des mots vides français

Liste des mots vides français			
a	à	accidentellement	accompagnant
actuellement	affectant	afin	agissant
ai	ailleurs	ainsi	and
anormalement	any	apparaissant	approximativement
après	associant	assurer	atteignant
attendant	attendu	au	aujourd'hui
auquel	aussi	autre	autres
aux	auxquelles	auxquels	avait
avant	avec	avoir	ayant
bien	c	c'	ça
car	causant	ce	ceci
cela	celle	celles	celui
cependant	certain	certaine	certaines
certains	ces	cet	cette
ceux	chaque	chez	ci
cinq	cinquièmement	cliniquement	cliquez
combien	comme	comment	commettre
compliquant	comprenant	concernant	concordant
conduisant	consultant	contre	correctement
couvrant	d	d'	dans
de	debout	dedans	dehors
delà	déjà	dépassant	depuis
derrière	des	dès	désormais
desquelles	desquels	dessous	dessus
deux	deuxièmement	devant	devers
devra	directement	divers	diverse
diverses	dix	doit	donc
donne	donnez	dont	douze
du	duquel	durant	elle
elles	émotionnellement	en	encore
entraînant	entraîner	entre	envers
environ	ès	essentiellement	est
et	étant	etc	été
etre	être	êtes	eu
eux	éventuellement	évoquant	excepté
exceptionnellement	exclure	exigeant	extra
extrêmement	facilitant	faussement	finement
fortement	fortuitement	fréquemment	gênant
généralement	grossièrement	habituellement	hélas
hormis	hors	hui	huit
			../..

Liste des mots vides français (suite)			
il	ils	immédiatement	impliquant
in	incluant	initialement	irradiant
istère	j	j'	je
jouxtant	jusqu	jusque	l
l'	la	là	laquelle
le	lequel	les	lesquelles
lesquels	leur	leurs	liant
lors	lorsque	lui	ma
mais	mal	malgré	manquant
masquant	me	même	mêmes
merci	mes	mien	mienne
miennes	miens	mieux	min
moi	moins	mon	moyennant
moyennement	n	n'	ne
néanmoins	nécessitant	neuf	ni
no	non	normalement	nos
notre	nôtre	nôtres	nous
ô	of	on	ont
onze	orientat	ou	où
outre	par	parmi	partant
partiellement	pas	passé	pendant
per	périodiquement	permet	permettant
permettre	peu	peut	plaignant
plein	plus	plusieurs	pour
pourquoi	préalablement	précédant	premièrement
près	présentant	presque	principalement
probablement	proche	progressant	proprement
provenant	provient	provoquant	puisque
qu	qu'	quand	quant
quatre	quatrièmement	que	quel
quelle	quelles	quelque	quelques
quels	québ	qui	quoi
quoique	rarement	récemment	relevant
répéter	résultant	revoici	revoilà
rien	s	s'	sa
sans	sauf	se	selon
sept	seront	ses	seulement
sévèrement	si	sien	sienne
siennes	siens	sinon	situe
six	soi	soit	son
sont	sous	spécifiquement	spontanément
suis	suivant	sur	survenant
ta	te	tes	tien
			../..

Liste des mots vides français (suite)			
tienne	tiennes	tiens	toi
ton	tous	tout	toute
toutes	transversalement	très	trois
troisièmement	trop	tu	un
une	va	vers	voici
voilà	voir	vos	votre
vôtre	vôtres	vouloir	vous
vu	y		

A.4 Liste des mots vides anglais

Liste des mots vides anglais			
a	about	above	accordingly
across	after	afterwards	again
against	all	allows	almost
alone	along	already	also
although	always	am	among
amongst	amoungst	amount	an
and	another	any	anybody
anyhow	anyone	anything	anyway
anywhere	apart	appear	appropriate
are	around	as	aside
associated	at	available	away
awfully	b	back	be
became	because	become	becomes
becoming	been	before	beforehand
behind	being	below	beside
besides	best	better	between
beyond	bill	both	bottom
brief	but	by	c
call	came	can	cannot
cant	cause	causes	certain
changes	co	come	computer
con	consequently	contain	containing
contains	corresponding	could	couldnt
cry	currently	d	day
de	describe	described	detail
did	different	do	does
doing	done	down	downwards
due	during	e	each
eg	eight	either	eleven
else	elsewhere	empty	enough
et	etc	even	ever
every	everybody	everyone	everything
everywhere	ex	example	except
f	far	few	fifteen
fifth	fify	fill	find
fire	first	five	followed
following	for	former	formerly
forth	forty	found	four
from	front	full	further
furthermore	g	get	gets
			../..

Liste des mots vides anglais (suite)			
give	given	gives	go
gone	good	got	great
h	had	hardly	has
hasnt	have	having	he
hence	her	here	hereafter
hereby	herein	hereupon	hers
herself	him	himself	his
hither	how	howbeit	however
hundred	i	ie	if
ignored	immediate	in	inasmuch
inc	indeed	indicate	indicated
indicates	inner	insofar	instead
interest	into	inward	is
it	its	itself	j
just	k	keep	kept
know	l	last	latter
latterly	least	less	lest
life	like	little	long
ltd	m	made	make
man	many	may	me
meanwhile	men	might	mill
mine	more	moreover	most
mostly	move	mr	much
must	my	myself	n
name	namely	near	necessary
neither	never	nevertheless	new
next	nine	no	nobody
none	noone	nor	normally
not	nothing	novel	now
nowhere	o	of	off
often	oh	old	on
once	one	ones	only
onto	or	other	others
otherwise	ought	our	ours
ourselves	out	outside	over
overall	own	p	part
particular	particularly	people	per
perhaps	placed	please	plus
possible	probably	provides	put
q	que	quite	r
rather	re	really	relatively
respectively	right	s	said
same	second	secondly	see
			../..

Liste des mots vides anglais (suite)			
seem	seemed	seeming	seems
self	selves	sensible	sent
serious	seven	several	shall
she	should	show	side
since	sincere	six	sixty
so	some	somebody	somehow
someone	something	sometime	sometimes
somewhat	somewhere	specified	specify
specifying	state	still	sub
such	sup	system	t
take	taken	ten	than
that	the	their	theirs
them	themselves	then	thence
there	thereafter	thereby	therefore
therein	thereupon	these	they
thick	thin	third	this
thorough	thoroughly	those	though
three	through	throughout	thru
thus	time	to	together
too	top	toward	towards
twelve	twenty	twice	two
u	un	under	unless
until	unto	up	upon
us	use	used	useful
uses	using	usually	v
value	various	very	via
viz	vs	w	was
way	we	well	went
were	what	whatever	when
whence	whenever	where	whereafter
whereas	whereby	wherein	whereupon
wherever	whether	which	while
whither	who	whoever	whole
whom	whose	why	will
with	within	without	work
world	would	x	y
year	years	yet	you
your	yours	yourself	yourselves
z	zero		

A.5 Exemples de contextes

Phrases extraites des corpus comparables, en français et en anglais, contenant le mot **foie** et la traduction **liver**.

Extrait des contextes des occurrences de **foie** dans le corpus CISMeF		
. . . transformation nodulaire du	foie	sans cirrhose. Il s'agit . . .
. . . cancers primitifs du	foie	concerne les hépatocarcinomes et les cancers développés sur cirrhose . . .
. . . maladies inflammatoires chroniques du	foie	en dehors de la pathologie . . .
. . . vaisseaux portes et métastase dans le	foie	par l'intermédiaires des branches portales . . .
. . . régénération hépatique sur un	foie	sain se fait à partir des hépatocytes . . .
. . . lésions trouvées dans le	foie	dans l'hépatite alcoolique . . .
Extrait des contextes des occurrences de **liver** dans le corpus CLINIWEB		
. . . matched normal	liver	tissue, hepatocellular carcinoma showed extinction. . .
. . . bone marrow metastasis and in multiple	liver	metastases, but not in normal DNA. . .
. . . nodular cirrhosis of the	liver	and impaired tyrosine metabolism. . .
. . . primary cancer of the	liver	as complication of giant cell hepatitis . . .
. . . capacity for regeneration of the rat, the	liver	regenerates the most of its original mass . . .
. . . main cause of chronic	liver	after renal transplantation, hepatitis C virus infection . . .

A.6 Extrait des résultats numériques

vessie		⟷	**bladder**
bladder	.329899	uretère	.392350
ovarian	.263955	**vessie**	.356089
prostate	.263202	urètre	.325898
small	.172262	vésicale	.277441
screening	.168382	rectum	.254249
tip	.163207	rétention	.228185
metastatic	.163189	urèthre	.223598
colon	.162360	miction	.206950
tract	.148228	prostatique	.190150
primary	.141628	bassinet	.188052
centre	.139984	fuite	.180336
polycystic	.137587	vésical	.177973
urologic	.137163	vagin	.174719
breast	.134460	broncho-pulmonaire	.172011
incidence	.133789	cystite	.171868
multidisciplinary	.123596	testicule	.171031
occur	.121761	advanced	.170695
benign	.119633	colo-rectal	.169643
seen	.116697	vide	.169448
present	.115970	bandelette	.167586
large	.114844	localized	.165912
increased	.114734	res	.159725
death	.114686	excrétion	.155412
australian	.113161	métastatique	.155400
including	.112744	pancréas	.155187
cervical	.110990	pleine	.152671
wilm	.110088	primitif	.150244
advanced	.107717	metastatic	.148698
early	.107178	adénocarcinome	.148486
australia	.106176	localisé	.147633

névrome		⟷		**neuroma**
neurofibroma	.226270		neurofibrome	.387371
neuroma	.209898		schwannome	.314184
opl	.176061		méningiome	.253363
freckling	.141581		glosso-pharyngien	.231074
onl	.138826		**névrome**	.230712
phrenic	.135247		pathétique	.229973
sural	.120105		hypoglosse	.228722
sciatic	.115814		phrénique	.210559
hypoglossal	.102198		gliome	.207736
pi	.100021		cranien	.207172
entrapment	.095337		glossopharyngien	.198252
plexiform	.092165		neurinome	.195977
eighth	.088877		ulnaire	.190250
circumscribed	.086652		honteux	.189157
regenerative	.083715		plexiforme	.184791
lisch	.083256		oculo-moteur	.183106
decompression	.083062		crural	.182153
glucosamine	.083053		pneumogastrique	.180905
abducen	.082820		oculomoteur	.161644
trochlear	.082407		innervé	.159113
interpositional	.082286		sortent	.157353
long-standing	.081687		rétinoblastome	.151316
vagus	.081391		épendymome	.149753
fascicle	.081096		trochléaire	.147298
transection	.080328		wrisberg	.144231
sublamina	.080226		hering	.140899
schwannoma	.079802		olfactif	.134727
rootlet	.078090		desmoïde	.132218
peroneal	.077483		pudendal	.131629
iac	.077047		récurrent	.131546

abcès		⟷	**abscess**
abscess	.240050	**abcès**	.281893
inflammation	.179661	hémisphère	.195616
complication	.179097	fistule	.195462
infarction	.167155	cervelet	.190174
hemorrhage	.162057	dème	.164984
edema	.161793	dispensée	.159802
recurrent	.160324	péritonite	.156658
lead	.159208	palsy	.154213
gross	.152554	formateur	.151913
seen	.148098	viscère	.147448
rare	.146320	diplômante	.146485
vascular	.142042	cellulite	.143297
occur	.134969	septicémie	.139749
develop	.130308	cicatrice	.138162
embolism	.128819	aboutissant	.137360
formation	.128422	ects	.136644
pneumonia	.127645	perfectionnement	.135724
dr	.124507	flow	.135712
hypertension	.124274	cur	.133411
include	.123144	hématome	.133051
including	.122350	compliquer	.131534
icon	.120319	caillot	.130887
prevent	.120286	aboutit	.130760
caused	.120005	appendicite	.130206
characterized	.119439	plèvre	.128045
lesion	.118515	emblée	.127614
likely	.117921	évoluer	.125597
involvement	.116867	localisée	.125067
microscopic	.116619	empyème	.123080
presence	.116541	survenir	.121666

cervelet	⟷		**cerebellum**
hemisphere	.289538	**cervelet**	.249003
cortex	.221325	vermis	.212582
palsy	.204290	hémisphère	.207562
cerebellum	.176821	tente	.199394
abscess	.144341	pédoncule	.199019
real	.138921	bulbe	.187403
mri	.129040	neuronale	.184259
dissection	.119520	noyal	.172438
damage	.116906	mésencéphale	.158593
cerebral	.115225	gliale	.152060
maldevelopment	.112125	immature	.144901
story	.108911	encéphale	.141575
stem	.107364	hippocampe	.138580
inferior	.105255	ciliée	.138147
brainstem	.104685	protubérance	.138018
dura	.103258	fusiforme	.134330
abscesse	.102001	cérébelleux	.133083
stroke	.098240	tubercule	.129922
cn	.097403	villositaire	.126515
posterior	.096744	schwann	.126370
anterior	.096108	fibroblaste	.124988
superior	.095682	mature	.123596
metastase	.095396	glioblastome	.119209
hemorrhage	.095283	thalamus	.117537
angiopathy	.094267	méninge	.116702
meninge	.092644	présentatrice	.116153
regional	.090672	neurale	.115888
transplant	.088956	indifférenciée	.115728
brad	.087085	totipotente	.115314
involvement	.086942	astrocyte	.114949

chéloïde	⟷		**keloid**
keloid	.302487	indélébile	.387364
hyperintense	.215956	**chéloïde**	.269496
thick-walled	.159403	hypertrophique	.268516
camouflage	.144143	choriorétinienne	.242821
subpleural	.138756	inesthétique	.213728
diverge	.135869	désunion	.203808
hypointense	.134665	déprimée	.180780
stellate	.123315	atrophique	.171704
inspect	.113704	séquellaire	.158621
orphology	.112343	lifting	.151045
gumma	.110245	microfibrille	.150180
subperiosteal	.104328	radiaire	.136413
pathologically	.104018	dépigmentée	.134027
new-bone	.093504	laissera	.133590
infarcted	.091998	disgracieuse	.131416
armpit	.085997	élastine	.131386
fade	.085418	pigmentation	.130107
fibrous-obliterative	.084683	pigmentée	.127904
recurred	.084146	fibreuse	.125889
hypertrophic	.084066	radiodermite	.118481
unilocular	.080426	fibronectine	.114844
tubo-ovarian	.078802	cicatrice	.114421
brodie	.077718	trophicité	.107098
morpheaform	.076123	laminine	.103857
walled	.075946	cicatriciel	.098861
iagnosis	.075921	mutilante	.097070
rac-induced	.075423	stellaire	.095708
disciform	.074372	ride	.094979
choledochal	.072454	laspect	.094546
tumor-like	.070300	achromique	.094127

verrue	⟷		**wart**
verruca	.373566	condylome	.239264
hpv	.289615	**verrue**	.226858
wart	.227892	vulgaire	.131566
papilloma	.162763	hpv	.117459
genital	.153550	papillomavirus	.116495
ebola	.131130	lappareil	.116055
simplex	.128735	herp	.084655
inverted	.126571	papillome	.083037
syncytial	.126569	mycoplasme	.067747
lymphadenopathy-associated	.126172	plane	.067392
t-lymphotropic	.120505	tractus	.065471
nile	.119144	ventouse	.064565
norwalk	.118477	herpès	.062414
epstein-barr	.114636	verruca	.061414
visible	.112386	bourrelet	.061273
perianal	.105561	récurrence	.057031
condylomata	.103513	lichen	.053404
wort	.102380	hsv	.052781
papillomavirus	.101116	tubercule	.051972
ebv	.100435	prolapsus	.051941
hsv	.100018	vph	.051253
contagiosum	.099921	replis	.051127
acuminata	.099854	mac	.050748
simian	.099237	plantaire	.049746
rsv	.096171	périné	.046628
plantar	.086700	uro	.046144
condyloma	.083891	pore	.045456
infect	.083597	vul	.045104
lassa	.083189	molluscum	.044773
hcv-related	.083083	ambiguïté	.044186

sphingomyéline		⟷	**sphingomyelin**
sphingomyelin	.195748	tranforment	.392256
clinical-genetic	.190962	trihexosyl	.332927
inverse	.167927	gbo	.316916
coefficient	.156589	**sphingomyéline**	.240887
radiologic-pathologic	.154901	dacyl-coa	.210103
mucopolysaccharide	.151554	lactosyl	.181245
arachidonic	.138724	galactosyl	.178725
genetic-clinical	.134011	oligosaccharidique	.177321
hyaluronic	.132402	trihexosidase	.170780
trihexoside	.126179	glycolipide	.170383
globoside	.124094	cytosidose	.169551
pfr	.122749	phosphoalcool	.167258
trihexosyl	.121315	céramide	.166319
phytanic	.120784	sialyloligosaccharide	.165285
sphingosine	.118855	galnac	.159985
pipecolic	.115166	ß-galactosidase	.159923
hydroxylated	.107263	phosphorylcholine	.156695
ibotenic	.106068	ganglioside	.144969
galnacî	.103353	cdp-choline	.136480
dibasic	.102000	acyl-transférase	.134045
ascorbic	.100925	lafora	.132384
hydrochloric	.098902	sitostérol	.131799
acetic	.098897	valeurs-seuil	.131036
lactosyl	.098194	dihydrosphingosine	.129326
pharmacologically	.097687	trigly	.129149
galactose	.097332	cholesté	.129046
keto	.094197	amidification	.125628
niemann-pick	.093770	lcat	.125478
genotype-phenotype	.092727	sphingomyélinase	.122965
ribonucleic	.092050	phytosterol	.121697

apnée	⟷		**apnea**
apnea	.301230	paradoxal	.500003
sleep	.224927	dette	.444816
alport	.138080	éveil	.364057
marfan	.118274	privation	.338019
fragile	.116220	somnolence	.297823
rett	.104729	endormissement	.293834
digeorge	.104360	**apnée**	.282998
alagille	.093327	rêve	.281216
acquired	.090755	dõautogestion	.274354
lesch-nyhan	.084807	probléme	.272514
kallmann	.078541	sieste	.250273
opitz	.077291	ventilatoire	.244290
problem	.077025	lent	.229585
infant	.070793	ronflement	.208526
hprt	.067715	svs	.206899
paraneoplastic	.066673	balantidiose	.201436
cushing	.065362	trichomonose	.194869
congenital	.063886	réparateur	.194230
latency	.063514	réveil	.192424
wiskott-aldrich	.063482	perturbé	.185672
coffin-lowry	.061550	xlien	.179744
hygiene	.059515	comas	.172798
disorder	.059277	posté	.169410
molecule	.058837	leishmaniose	.169090
nephrotic	.058832	cdrom	.165420
multiple	.058379	sauveur	.164060
child	.055726	insomnie	.158440
including	.055472	irritabilité	.158160
anomaly	.054757	tarn	.157708
familial	.054563	chagas	.148899

mélanome		⟷	**melanoma**
melanoma	.355220	**mélanome**	.397835
stage	.285958	lymphome	.249872
metastatic	.274561	malin	.204179
ovarian	.245024	bénin	.185654
lymphoma	.230850	hodgkinien	.129887
advanced	.214352	primitif	.125964
early	.200569	mésothéliome	.120978
neoplasm	.189854	sarcome	.113238
lesion	.188275	carcinome	.105416
tip	.185482	peau	.099933
diagnosed	.177446	épithéliale	.094767
prostate	.176968	épaisseur	.088413
australian	.176029	tumorale	.087928
survival	.171879	opéré	.085623
australia	.170165	cancéreuse	.085479
screening	.169996	nodule	.082926
malignant	.169104	endothéliale	.081880
hyperthermia	.168980	cutané	.076637
cervical	.166909	prolifération	.076317
staging	.164857	embryonnaire	.076100
incidence	.162588	adénocarcinome	.071229
metastase	.160631	métastatique	.070737
benign	.159911	métastase	.070276
colorectal	.157015	dendritique	.070030
cutaneous	.154354	différenciation	.069335
incorporating	.151149	localisation	.068339
colon	.144415	histologique	.068064
primary	.144295	es	.066699
development	.141502	hématopoïétique	.066379
centre	.140314	nerveuse	.066342

ulcère		⟷	**ulcer**
ulcer	.275763	**ulcère**	.304020
gastric	.138056	duodénal	.246240
peptic	.123084	ulcéreuse	.233656
pylori	.108808	gastrique	.226854
duodenal	.107689	helicobacter	.211305
atresia	.070178	pylori	.211186
diagnosis	.067576	poussée	.189676
occur	.064931	gastrite	.187081
acute	.064455	contagieuse	.171984
leg	.063609	lassurance	.171280
incidence	.062918	lyme	.168316
severe	.059834	cicatrisation	.165449
obstruction	.057318	gastro-duodénal	.164627
seen	.055603	progression	.163995
reported	.055086	auto-immune	.163983
present	.054838	dassurance	.163506
treatment	.053552	crohn	.163499
similar	.052220	variante	.162039
early	.051649	parkinson	.160535
increased	.049761	opéré	.157957
pain	.049672	asthmatique	.157299
report	.049117	provoquer	.156733
finding	.048955	coronarienne	.155505
study	.048876	diagnostiquer	.153574
result	.048283	guérison	.153424
week	.047875	horton	.152570
include	.047690	traiter	.150786
stomach	.046770	mortelle	.149095
small	.045700	évolutive	.148561
lesion	.045699	suivis	.147693

sein	⟷		**breast**
center	.359986	prostate	.839669
breast	.357395	poumon	.721924
cancer	.257484	colorectal	.580114
lung	.232067	ligue	.434209
diagnostic	.185384	registre	.416689
prostate	.156130	**sein**	.400638
internal	.147301	thyroïde	.356426
test	.130226	col	.343880
laboratory	.126120	ovaire	.267845
national	.115082	lutte	.266968
pathology	.107917	côlon	.259189
topic	.098756	incidence	.231111
women	.090790	décès	.230034
pain	.090272	dépistage	.229809
breathing	.083377	thyroïdien	.224795
information	.083370	utérus	.205765
ovarian	.079307	utérin	.204189
risk	.078434	institute	.198675
medicine	.078290	colon	.192649
related	.077596	fréquent	.189726
imaging	.072632	mortalité	.187195
prevention	.070973	rectum	.185822
treatment	.069650	research	.174759
centre	.064468	foie	.174144
disease	.063200	lung	.173355
cell	.061494	cause	.172442
care	.056699	bronchique	.161577
service	.052652	augmentation	.158753
case	.052326	screening	.158261
screening	.051957	pancréas	.153452

crâne		⟷	**skull**
skull	.295435	purique	.327362
fracture	.174535	**crâne**	.307340
temporal	.160646	pyrimidique	.289764
cranial	.130708	urofrance	.264932
marrow	.125043	azotée	.264874
lesion	.122419	volontariat	.258215
single	.119185	servent	.237183
scan	.117678	informatisée	.230278
spine	.103832	assistance-emploi	.226290
posterior	.096984	los	.221474
pair	.084358	voûte	.210574
anterior	.083394	hyoïde	.208805
formation	.082989	servi	.201708
tongue	.081974	database	.187619
metastase	.080063	quarré	.187613
seen	.079594	plat	.187597
base	.076891	radius	.178573
mass	.073982	donnee	.177213
cyst	.070060	servir	.170904
sinus	.069079	humérus	.170838
transplantation	.066134	articule	.165213
facial	.065769	servant	.165057
frontal	.065767	alumnas	.163668
involvement	.065439	carpe	.163087
occur	.065099	rocher	.162674
hip	.064335	fémur	.162667
including	.062662	sacrum	.161477
hand	.061492	ostéoporotique	.161273
present	.060743	las	.158835
middle	.060154	fragilité	.157362

Annexe B

Annexe : Expériences de recherche d'information translangue

B.1 Liste des requêtes MeSH

Dans le thésaurus MeSH, les termes employés pour désigner les concepts sont présentés en majuscules et sans accent. Ce qui rend leur utilisation difficile dans les traitements automatiques. Nous avons décidé de les présenter en minuscules et avec accents pour une meilleure illustration.

Liste des requêtes MeSH		
MSH3688	preventive medicine	medecine préventive
MSH3296	otolaryngology	otorhinolaryngologie
MSH3971	research support	aide recherche
MSH850	child development	développement enfant
MSH1441	ehlers-danlos syndrome	ehler danlos syndrome
MSH3860	radiographic image enhancement	amélioration image radiographique
MSH280	antibody-producing cell	cellule productrice anticorps
MSH3609	poliovirus vaccine oral	vaccin antipoliomyélitique sabin
MSH912	chromatography ion exchange	chromatographie échange ion
MSH2954	molecular biology	biologie moléculaire
MSH2108	hepatitis b antigens	antigène hbv
MSH2598	legislation drug	législation produit chimique ou pharmaceutique
MSH2649	linoleic acid	acide linoléique
MSH4553	tissue preservation	conservation tissu
MSH2177	hospital record	archive hôpital
MSH4775	viral protein	proteine virale
MSH4224	social control formal	contrôle social formel
MSH346	appointment and schedule	rendez-vous et programme
MSH1041	concanavalin a	concanavaline a
MSH2798	medical oncology	oncologie médicale
MSH1994	health facility	équipement santé
MSH48	actuarial analysis	analyse actuarielle
MSH4463	temporal arteritis	artérite temporale
MSH2107	hepatitis b antibody	anticorps anti-hbv
MSH425	autoantigens	auto-antigène
MSH4048	rna splicing	épissage arn
MSH2547	kupffer cell	cellule kupffer
MSH2780	measle	rougéole
MSH3982	respiratory hypersensitivity	allergie respiratoire
MSH2124	herpes genitalis	herpès génital
MSH290	antigen-presenting cells	cellule présentant antigène
MSH2378	infusion parenteral	perfusion parentérale
MSH3227	oncogene proteins viral	protéines oncogène virale
MSH1044	condylomata acuminata	condylome acuminé
		../..

Liste des requêtes MeSH (suite)		
MSH1659	fatty acid	acide gras
MSH4482	tetradecanoylphorbol acetate	acétate tetradecanoylphorbol
MSH93	adrenoleukodystrophy	adrenoleucodystrophie
MSH1532	epidermolysis bullosa	épidermolyse bulleuse
MSH36	acid phosphatase	acid phosphatase
MSH1187	day care	soins jour
MSH1588	ethics professional	ethique professionnelle
MSH2340	immunosuppression	immunodepression
MSH383	aspartic acid	acide aspartique
MSH1013	common cold	rhume banal
MSH1351	dna mitochondrial	adn mitochondrial
MSH4340	street drug	produit illicite
MSH1874	genetic medical	génétique medicale
MSH2614	lethal dose 50	dose letale 50
MSH1606	exercise therapy	traitement par effort
MSH2111	hepatitis b surface antigens	antigène hbs
MSH250	anorexia nervosa	anorexie mentale
MSH4144	self-help group	groupe soutien social
MSH272	antibody bacterial	anticorps antibacterien
MSH274	antibody neoplasm	anticorps antitumoral
MSH1234	depressive disorder	trouble dépressif
MSH4273	spectrum analysis mass	analyse spectrale masse
MSH908	chromatography affinity	chromatographie affinité
MSH2115	hepatitis toxic	hépatite toxique
MSH907	chromatography	chromatographie
MSH1765	food microbiology	microbiologie alimentaire
MSH2320	immunoglobulin allotype	allotype immunoglobuline
MSH3705	professional practice	pratique professionnelle
MSH2689	longitudinal study	étude longitudinale
MSH1738	flow cytometry	cytométrie flux
MSH1137	crystallography	cristallographie
MSH3979	respiratory distress syndrome	détresse respiratoire syndrome
MSH911	chromatography high pressure liquid	chromatographie liquide haute pression
MSH4698	ursodeoxycholic acid	acide ursodésoxycholique
MSH277	antibody diversity	diversité anticorps
MSH2208	hydrochloric acid	acide chlorhydrique
MSH3300	outcome and process assessment health care	évaluation résultat et méthode soins
MSH2654	lipid bilayer	double couche lipidique
MSH3722	propionic acid	acide propionique
		../..

Liste des requêtes MeSH (suite)		
MSH1843	gastroenterology	gastroentérologie
MSH4774	viral hepatitis vaccine	vaccin antihépatite virale
MSH2159	home nursing	soins a domicile
MSH1354	dna ribosomal	adn ribosomique
MSH3575	plasminogen activator	activateur plasminogène
MSH726	cardiovascular agent	agent cardiovasculaire
MSH4142	self medication	auto-médication
MSH2781	measle vaccine	vaccin antimorbilleux
MSH1476	emergency	urgence
MSH2551	labor complication	accouchement compliqué
MSH3683	preoperative care	soins préopératoire
MSH1970	h-2 antigens	antigène h2
MSH2125	herpes simplex	herpès
MSH2955	molecular conformation	configuration moléculaire
MSH3633	population surveillance	surveillance population
MSH2286	iatrogenic disease	affection iatrogénique
MSH778	cell communication	communication cellulaire
MSH3643	postoperative care	soins postopératoire
MSH1244	desensitization immunologic	désensibilisation immunologique
MSH135	algorithm	algorithme
MSH446	bacterial adhesion	adhérence bactérienne
MSH4272	spectrum analysis	analyse spectrale
MSH3255	organ procurement	recueil organe transplantation
MSH271	antibody	anticorps
MSH2158	home care service	service soins domicile
MSH3188	nursing home	maison repos
MSH2109	hepatitis b core antigens	antigène hbc
MSH4626	truth disclosure	divulgation verité
MSH61	adenosine deaminase	adénosine deaminase
MSH4442	taurocholic acid	acide taurocholique
MSH392	asthma exercise-induced	asthme effort
MSH4395	surgery oral	chirurgie stomatologique
MSH142	allied health personnel	personnel santé auxiliaire
MSH3496	phosphatidic acid	acide phosphatidique
MSH1208	delivery of health care	délivrance soins
MSH2371	influenza vaccine	vaccin antigrippe
MSH2834	meningitis viral	meningite virale
MSH4040	right to die	droit à la mort
MSH1423	education	enseignement et éducation
MSH3694	private practice	pratique professionnelle privée
MSH1279	diet survey	enquête régime alimentaire
		../..

Liste des requêtes MeSH (suite)		
MSH1190	death certificate	certificat décès
MSH3674	pregnancy prolonged	grossesse prolongée
MSH2110	hepatitis b e antigens	antigène hbe
MSH3507	phosphonoacetic acid	acide phosphonoacétique
MSH2016	health survey	enquête santé
MSH2118	hepatolenticular degeneration	dégénérescence hépatolenticulaire
MSH4883	5 8 11 14 17-eicosapentaenoic acid	acide eicosapentaenoique-5 8 11 14 17
MSH2116	hepatitis viral human	hépatite virale humaine
MSH54	addison disease	addison maladie
MSH3680	prenatal care	soins prenatal
MSH176	aminocaproic acid	acide aminocaproique
MSH1464	electrophoresis polyacrylamide gel	électrophorèse gel polyacrylamide
MSH2013	health service research	recherche en santé publique
MSH492	bibliography	bibliographie
MSH2464	iodohippuric acid	acide iodohippurique
MSH74	administration intranasal	voie intranasale
MSH3336	palmitic acid	acide palmitique
MSH4835	wheat germ agglutinin	agglutinine germe blé
MSH4473	terminal care	soins aux mourant
MSH4882	3 4-dihydroxyphenylacetic acid	acide 3 4-dihydroxyphenylacétique
MSH3207	occupational medicine	médecine travail
MSH71	adjuvant immunologic	adjuvant immunologique
MSH1783	forensic medicine	médecine légale
MSH886	cholic acid	acide cholique
MSH2315	immunoassay	dosage immunologique
MSH4141	self disclosure	ouverture personnelle
MSH1598	euthanasia passive	euthanasie passive
MSH4072	s-adenosylmethionine	adénosylmethionine
MSH3865	radioimmunoassay	dosage radioimmunologique
MSH2009	health service accessibility	accessibilité service santé
MSH3816	purchasing hospital	achat de l'hopital
MSH521	biotechnology	biotechnologie
MSH3274	osmolar concentration	concentration osmolaire
MSH1066	continuity of patient care	suivi soins patient
MSH12	abscess	abcès
MSH1767	food service hospital	service nutrition hôpital
MSH66	adenovirus infection human	adénovirus infection humaine
MSH497	bile acid and salt	acide et sels biliaire
MSH276	antibody affinity	affinité anticorps
MSH2786	meclofenamic acid	acide méclofénamique
		../..

Liste des requêtes MeSH (suite)		
MSH1867	genetic counseling	conseil génétique
MSH989	coliphage	coliphage
MSH582	bone development	croissance osseuse
MSH2365	infectious mononucleosis	mononucléose infectieuse
MSH1455	electromagnetic fields	champ électromagnétique
MSH4773	viral envelope proteins	protéines virale enveloppe
MSH2370	influenza	grippe
MSH1257	diabete mellitus insulin-dependent	diabète insulinodépendant
MSH102	affective symptom	affectif symptome
MSH2223	hydroxyeicosatetraenoic acid	acide hydroxyéicosatétraenoique
MSH1463	electrophoresis agar gel	électrophorèse gel agar
MSH1768	food service	réstauration
MSH2437	interview	entretien
MSH4172	sexually transmitted disease	maladie sexuellement transmissible
MSH4315	stainless steel	acier inoxydable
MSH173	amino acid metabolism inborn error	aminoacidopathie héréditaire
MSH3488	philadelphia chromosome	chromosome philadelphie
MSH1586	ethics medical	éthique médicale
MSH2898	microbial sensitivity test	test sensibilité microbienne
MSH910	chromatography gel	chromatographie gel
MSH3327	p-aminohippuric acid	acide para-aminohippurique
MSH2799	medical record	dossier médical
MSH2201	hyaluronic acid	acide hyaluronique
MSH4064	rubella vaccine	vaccin antirubéoleux
MSH4080	salicylic acid	acide salicylique
MSH2625	leukoencephalopathy progressive multifocal	leucoencéphalopathie multifocale progressive
MSH913	chromatography thin layer	chromatographie couche mince
MSH3601	pneumonia viral	pneumopathie virale
MSH3945	regional medical program	programme medical régional
MSH3163	nordihydroguaiaretic acid	acide nordihydroguaiaretique
MSH720	cardiolipin	cardiolipide
MSH2008	health service	service santé
MSH1443	electric conductivity	conductivité électrique
MSH1755	folic acid	acide folique
MSH1356	dna superhelical	adn superhélicoidal
MSH1658	fatty acid desaturase	acyl-coa désaturase
MSH213	anemia hemolytic autoimmune	anémie hémolytique autoimmune
MSH1660	fatty acid essential	acide gras indispensable
MSH2298	image enhancement	agrandissement image
MSH4782	virus replication	réplication virale
		../..

Liste des requêtes MeSH (suite)		
MSH670	calorimetry indirect	calorimétrie indirecte
MSH4183	sick role	personnage du malade
MSH175	amino acids branched-chain	acides aminés à chaine ramifiée
MSH1167	cytopathogenic effect viral	pouvoir cytopathogène virus
MSH1883	germ-free life	élevage axenique
MSH2997	mumps	oreillon
MSH1661	fatty acid nonesterified	acide gras libre
MSH1763	food handling	traitement aliment
MSH73	administration inhalation	administration respiratoire
MSH1377	drug combinations	association médicamenteuse
MSH19	accident occupational	accident travail
MSH307	antibody antinuclear	anticorps antinucléaire
MSH2774	maternal-fetal exchange	échange foetomaternel
MSH4231	social support	soutien social
MSH704	carcinoembryonic antigen	antigène carcino-embryonnaire
MSH2450	intraoperative care	soins peropératoire
MSH63	adenosine diphosphate ribose	adénosine diphosphate ribose
MSH1723	fibrous dysplasia of bone	dysplasie fibreuse des os
MSH15	abstracting and indexing	analyse et indexation
MSH260	antibody anti-idiotypic	anti-anticorps
MSH162	ambulatory care facility	service soins ambulatoire
MSH1456	electromagnetic	électromagnétique
MSH1390	drug utilization	utilisation médicament
MSH1727	financial management	gestion financière
MSH1238	dermatitis atopic	dermatite atopique
MSH273	antibody fungal	anticorps antifongique
MSH1352	dna neoplasm	adn tumoral
MSH4106	schizophrenic psychology	psychologie des schizophrènes
MSH4772	viral core proteins	protéines virale capside
MSH1316	disability evaluation	évaluation incapacité
MSH2162	home for the aged	maison médicalisée personne âgée
MSH4263	specialty board	conseil de spécialité
MSH1471	embolism air	embolie gazeuse
MSH1952	grave disease	basedow maladie
MSH180	aminosalicylic acid	acide aminosalicylique
MSH3204	occult blood	sang occulté
MSH4704	utilization review	bilan opérationnel
MSH3217	oleic acid	acide oléique
MSH1854	gene amplification	amplification génique
MSH424	autoantibody	auto-anticorps
MSH3065	national health program	programme national santé
		../..

Liste des requêtes MeSH (suite)		
MSH52	acyltransferase	acyltransférase
MSH379	ascorbic acid	acide ascorbique
MSH4860	x-ray diffraction	diffraction rx
MSH909	chromatography ga	chromatographie phase gazeuse
MSH4181	sialic acid	acide sialique
MSH3987	respiratory syncytial viruse	virus respiratoire syncytial
MSH4586	training support	aide enseignement
MSH47	activity of daily living	activité quotidienne
MSH4137	self administration	auto-administration
MSH3254	organ preservation	conservation organe
MSH6	abnormality drug-induced	malformation origine chimique
MSH3193	nutrition survey	enquête nutritionnelle
MSH275	antibody viral	anticorps antiviral
MSH390	aspirin	acide acétylsalicylique
MSH567	blood transfusion autologous	transfusion sanguine autologue
MSH2470	ioxaglic acid	acide ioxaglique
MSH4234	society medical	association médicale
MSH2509	kainic acid	acide kainique
MSH790	cell transformation viral	transformation cellulaire par virus
MSH2998	mumps vaccine	vaccin antiourlien
MSH2105	hepatitis antibody	anticorps antivirus hépatite
MSH33	acetylglucosaminidase	aéetylglucosaminidase
MSH1303	dinitrochlorobenzene	dinitrochlorobenzene
MSH1662	fatty acid unsaturated	acide gras insaturé
MSH1197	decision making computer-assisted	aide décision ordinateur
MSH1286	dietary service	service diététique
MSH4063	rubella	rubéole
MSH1884	gestational age	âge gestationnel