Extraction de lexiques bilingues à partir de corpus comparables

Amir Hazem

UNIVERSITÉ DE NANTES
FACULTÉ DES SCIENCES ET DES TECHNIQUES

ÉCOLE DOCTORALE SCIENCES & TECHNOLOGIES
DE L'INFORMATION ET MATHÉMATIQUES – STIM

Année 2013

N° attribué par la bibliothèque

Extraction de lexiques bilingues à partir de corpus comparables

THÈSE DE DOCTORAT
Discipline : Informatique
Spécialité : Traitement Automatique du Langage Naturel

Présentée
et soutenue publiquement par

Amir HAZEM

Le 11 octobre 2013, devant le jury ci-dessous

Président	Éric GAUSSIER, Professeur des Universités, Université de Grenoble
Rapporteurs	Pierre ZWEIGENBAUM, Directeur de recherche, LIMSI-CNRS
	Philippe LANGLAIS, Professeur des Universités, Université de Montreal
Examinateurs	Béatrice DAILLE, Professeur des Universités, Université de Nantes
	Éric GAUSSIER, Professeur des Universités, Université de Grenoble
	Emmanuel MORIN, Professeur des Universités, Université de Nantes

Directeur de thèse : Emmanuel MORIN, Professeur des Universités, Université de Nantes

Remerciements

L'accomplissement de soi ne peut se faire qu'à travers les autres [1], c'est pourquoi j'aimerai à travers ces quelques lignes remercier tous ceux et toutes celles qui, de près ou de loin ont contribué à l'aboutissement de ce modeste travail. Long fut le dénouement ! Pour ceux qui me connaissent et pour les autres cette thèse est passée trop vite. En est la preuve mon envie insatiable de continuer, mais comme toute bonne chose a une fin, il est temps de mettre un terme à ces quatre années de recherches au combien enrichissantes.

Si cette période de doctorat fut pour moi une expérience riche en recherches scientifiques et en relations humaines, qui m'ont donné envie d'embrasser une carrière de chercheur mais aussi d'enseignant, c'est avant tout grâce à mon directeur de thèse et encadrant Emmanuel Morin, sans qui, rien n'aurait été possible. Emmanuel, je te suis infiniment reconnaissant pour m'avoir donné la possibilité de travailler dans les meilleures conditions qui soient. Merci pour ta disponibilité, ta gentillesse, ta patience, ta bonne humeur, tes remarques avisées ; je pourrais continuer comme cela pendant longtemps mais comme tu le dis souvent : *va à l'essentiel et sois concis*, alors je vais essayer d'appliquer ces conseils en commençant dès maintenant en te disant tout simplement un GRAND MERCI.

Au risque de rentrer dans une interminable anaphore où le syntagme "Je remercie" inonderait ces pages, car oui il y en a des personnes à remercier, je vais essayer d'alterner, sans grande innovation je suppose, des figures de style qui rendront je l'espère ces remerciements moins répétitifs.

Je tiens donc à remercier [2] les membres de mon jury de thèse pour leurs remarques pertinentes et constructives et tout particulièrement les relecteurs Pierre Zweigenbaum et Philippe Langlais pour leurs conseils et corrections au combien utile.

La recherche scientifique c'est aussi faire partie d'une équipe, c'est pourquoi je suis particulièrement honoré d'avoir côtoyé les membres de l'équipe TALN dans diverses réunions, conférences et enseignements. Je remercie bien sûr Béatrice Daille que j'espère enfin pouvoir tutoyer après le dépôt de ce manuscrit. Sans citer tout le monde, car oui il y en a du monde dans l'équipe, je remercie mes collègues : Nicolas, Chantal, Denis, notre regretté Sasha parti trop tôt, Solen, Emmanuel Planas, Christine, Colin, Laura et Florian... et mes camarades : Ophélie, Liza, Rima, James, Momo, Firas, Adrien, Prajol et Estelle ainsi que les anciens : Ramadan, Fabien, Matthieu et Sebastián.

Une pensée particulière pour mes collègues enseignants de l'IUT et de l'UFR des Sciences et Techniques de Nantes auprès de qui j'ai beaucoup appris. Sans oublier le personnel administratif et l'équipe technique du LINA pour leur sympathie et leur disponibilité.

1. Une pensée philosophique subjective qui n'engage que moi, d'autant plus que cet accomplissement n'est que partiel, puisqu'il ne concerne que le parcours professionnel néanmoins indispensable à tout un chacun.
2. Oui je me répète déjà...

Difficile est la rédaction d'un ouvrage, et encore plus quand on en a pas l'habitude, c'est pourquoi je suis profondément reconnaissant aux personnes qui ont participé à la relecture et à la correction de ce manuscrit. Je pense à Gabrielle, à Olivier, à Adrien, à Emmanuel, à ma mère et à mes deux relecteurs Pierre et Philippe et aux membres de mon comité de suivi de thèse : Éric Gaussier et Pierre-françois Marteau.

Durant ces 4 années j'ai eu souvent l'occasion de changer de bureau (3 en tout), j'aimerai donc rendre hommage à mes camarades de bureau pour leur compagnie des plus agréables. Merci à Olivier, Wence, Hafed, Luis, Momo, Firas, Rima, Khaled et Ahmed. Je n'oublierai pas les différentes discussions qu'on a pu avoir.

Un petit clin d'œil à Selma, Jalila, Guillaume et Bruno ainsi qu'à l'association des jeunes chercheurs : LOGIN pour les différentes activités organisées. Merci aussi à tous les doctorants du LINA, je ne vous oublierai pas.

Et pour finir, je ne remercierai jamais assez, et tous les mots du monde ne suffiront pas pour cela, celle qui a toujours été là, derrière moi, qui m'a toujours poussé et encouragé, soutenu dans les moments difficiles, qui a cru en moi, je pense bien sûr à ma mère, cette thèse est avant tout pour toi et à toi, alors merci du fond du cœur.

Liste de publications

Conférences d'Audience Internationale

[1] Amir Hazem, Emmanuel Morin : Word Co-occurrence Counts Prediction for Bilingual Terminology Extraction from Comparable Corpora. *In proceedings of the 6th International Joint Conference on Natural Language Processing.* ***IJCNLP 2013****.*

[2] Amir Hazem, Emmanuel Morin : Adaptive Dictionary for Bilingual Lexicon Extraction from Comparable Corpora. *In proceedings of the 12th International Conference on Language Resources and Evaluation.* ***LREC 2012*** *(Short paper).*

[3] Amir Hazem, Emmanuel Morin : QAlign: A New Method for Bilingual Lexicon Extraction from Comparable Corpora. *In proceedings of the 13th Conference on Intelligent Text Processing and Computational Linguistics.* ***CICLing 2012****.*

Conférences d'Audience Nationale

[4] A. Hazem, E. Morin : Extraction de lexiques bilingues à partir de corpus comparables par combinaison de représentations contextuelles. *In actes de la 20e Conférence sur le Traitement Automatique des Langues.* ***TALN 2013****.*

[5] A. Hazem, E. Morin and S. Peña Saldarriaga : Métarecherche pour l'extraction lexicale bilingue à partir de corpus comparables. *In actes de la 18e Conférence sur le Traitement Automatique des Langues.* ***TALN 2011****.*

[6] Bo Li, Éric Gaussier, Emmanuel Morin and Amir Hazem : Degré de comparabilité, extraction lexicale bilingue et recherche d'information interlingue. *In actes de la 18e Conférence sur le Traitement Automatique des Langues.* ***TALN 2011****.*

Ateliers (Workshop)

[7] Amir Hazem, Emmanuel Morin : A Comparison of Smoothing Techniques for Bilingual Lexicon Extraction from Comparable Corpora. *In proceedings of the 6th Workshop on Building and Using Comparable Corpora.* ***BUCC 2013****.*

[8] Amir Hazem, Emmanuel Morin : ICA for Bilingual Lexicon Extraction from Comparable Corpora. *In proceedings of the 5th Workshop on Building and Using Comparable Corpora.* ***BUCC 2012****.*

[9] Florian Boudin, Amir Hazem, Nicolas Hernandez and Prajol Shrestha. *Défi Fouille de Textes.* ***DEFT 2012.***

[10] A. Hazem, E. Morin and S. Peña Saldarriaga : Bilingual Lexicon Extraction from Comparable Corpora as Metasearch. *In proceedings of the 4th Workshop on Building and Using Comparable Corpora.* ***BUCC 2011.***

Table des matières

Liste des tableaux

Table des figures

Introduction

Les lexiques bilingues constituent une ressource fondamentale qui trouve son utilité dans plusieurs applications multilingues du traitement automatique des langues telles que la traduction automatique ou la recherche d'information interlingue. Il existe depuis plusieurs années maintenant un engouement des scientifiques et des chercheurs envers l'extraction terminologique et la création automatique de lexiques à partir de corpus multilingues. Cet intérêt particulier est motivé entre autres par la globalisation des échanges commerciaux des entreprises, et particulièrement dans leur domaine d'expertise où la maîtrise de la terminologie devient cruciale. La nécessité de disposer et de fournir des contenus textuels cohérents et la perpétuelle évolution de la terminologie rendent fastidieuse la création manuelle de telles ressources. Il est très rare de trouver des ressources linguistiques adaptées au domaine technique de l'entreprise. En plus de l'enjeu scientifique, la construction de lexiques terminologiques multilingues est devenue un enjeu économique.

La constitution des lexiques bilingues manuellement étant difficilement envisageable, les recherches se sont tournées vers l'exploitation de corpus pour construire ces lexiques. Le présent manuscrit aborde l'utilisation d'un type de corpus spécifique appelé corpus comparable pour la tâche d'extraction de lexiques bilingues de manière automatique. Un corpus comparable, par opposition à un corpus parallèle, peut être vu comme un ensemble de documents écrits en plusieurs langues traitant des mêmes sujets sans être en relation de traduction. Ce type de corpus constitue de par son abondance un réservoir d'information utile dans plusieurs domaines d'application tels que les mémoires de traduction ou la catégorisation multilingue. L'une des raisons principales de l'intérêt des chercheurs pour les corpus comparables par rapport à d'autres types de corpus est leur disponibilité grandement favorisée par l'essor du web, contrairement aux corpus parallèles par exemple qui restent plus difficiles à collecter et plus particulièrement en domaine de spécialité ou pour des couples de langues moins usités. Ces travaux se concentrent principalement sur l'extraction de termes spécialisés en exploitant des corpus comparables en langue de spécialité.

La plupart des travaux en acquisition de lexiques bilingues à partir de corpus comparables reposent sur l'hypothèse distributionnelle qui a été étendue au cadre bilingue. Deux mots ont de fortes chances d'être en relation de traduction s'ils apparaissent dans les mêmes contextes lexicaux. Ce postulat suppose une définition claire et rigoureuse du contexte et une connaissance parfaite des indices contextuels. Or, la

complexité et les spécificités de chaque langue font qu'il n'est pas aisé d'énoncer une telle définition qui garantisse une extraction de couples de traductions efficace dans tous les cas de figures. Toute la difficulté réside donc dans la manière de définir, d'extraire et de comparer ces contextes dans le but de construire des lexiques bilingues fiables.

Nous nous efforcerons tout au long des différents chapitres de cette thèse d'essayer de mieux comprendre cette notion de contexte, pour ensuite l'étendre et l'adapter afin d'améliorer la qualité des lexiques bilingues extraits. Une première partie des contributions vise à améliorer l'*approche directe* qui fait office de référence dans la communauté. Nous proposons plusieurs manières d'aborder le contexte pour mieux caractériser les mots du vocabulaire. Partant du fait que l'*approche directe* repose sur l'observation des cooccurrences des mots, nous proposons différentes techniques de ré-estimation de ces cooccurrences et montrons que ceci a un apport significatif sur la qualité des lexiques extraits. Une deuxième proposition, inspirée du domaine de la recherche d'information, consiste à présenter une méthode de combinaison de contextes qui tire parti des deux principales représentations contextuelles, à savoir la représentation dite graphique, qui repose sur une simple collecte des mots entourant un mot à caractériser selon une taille de fenêtre donnée, et une deuxième représentation dite syntaxique, qui repose sur les relations de dépendance syntaxique entre les mots. Nous montrons qu'une utilisation conjointe de ces deux représentations s'avère pertinente et améliore grandement les performances. Dans la deuxième partie des contributions, nous commençons par présenter une méthode qui vise à améliorer l'*approche par similarité inter-langue* aussi considérée comme état de l'art, où nous définissons une autre manière d'exploiter les k plus proches voisins d'un mot à traduire. Ensuite, une méthode nommée *Q-Align*, directement inspirée des systèmes de question/réponse est présentée. Celle-ci à l'instar des autres méthodes, se base sur une représentation locale du contexte et apporte ainsi une contribution supplémentaire et notamment en associant à chaque traduction le passage illustrant au mieux son emploi. Enfin, nous abordons un autre point essentiel qui est la représentation des mots du corpus dans différents espaces vectoriels. Après l'introduction des principes de base, nous présentons plusieurs transformations mathématiques et donc plusieurs représentations vectorielles et montrons leurs performances ainsi que l'intérêt de les combiner, notamment avec l'*approche directe*. Les différents résultats de ces contributions nous amènes à les regrouper dans une seule et même architecture. Ainsi, nous présentons comme dernière contribution un système multi-sources qui exploite les avantages des différentes approches proposées.

Ce manuscrit est constitué de quatre parties contenant chacune un ou plusieurs chapitres organisés comme suit :

La première partie est constituée de deux chapitres qui abordent dans un premier temps, les notions de corpus parallèles et comparables pour ensuite présenter les principales méthodes d'extraction de lexiques bilingues à partir de corpus comparables. Une attention particulière est donnée aux méthodes ayant pour but d'aligner des termes simples. Toutefois, une introduction aux méthodes d'alignement de termes complexes est présentée.

La deuxième partie (chapitre 3) propose une vue d'ensemble sur la définition du contexte telle que vue par les chercheurs en alignement multilingue. Nous revenons sur les notions de mots et de termes qui représentent la plus petite unité du contexte.

Ensuite, nous présentons différentes définitions ainsi que la manière de construire et de comparer les contextes.

La troisième partie (chapitre 4) présente une étude approfondie de l'*approche directe* qui constitue l'état de l'art en extraction de lexiques bilingues à partir de corpus comparables. Dans le but de mieux comprendre les lacunes de cette approche, nous étudions chacun de ses paramètres. Ceci servira de base pour justifier nos propositions.

La quatrième partie concerne les différentes contributions de ces travaux de thèse. Celle-ci est divisée en six chapitres. Les cinquième et sixième chapitres sont consacrés aux améliorations apportées autour de l'*approche directe*. Nous présentons dans le chapitre 5 une première méthode qui combine deux principales représentations contextuelles. Ensuite, nous présentons dans le chapitre 6 des techniques de ré-estimation et de lissage appliquées aux cooccurrences des mots dans le but de rendre plus fiable ces valeurs de cooccurrence. Les chapitres 7, 8 et 9 concernent la proposition de nouvelles méthodes. Dans le chapitre 7, nous présentons une méthode inspirée de l'*approche par similarité inter-langue* nommée *Métarecherche*. Le chapitre 8 est consacré à la méthode *Q-Align* qui constitue une autre manière d'appréhender l'extraction lexicale bilingue en se focalisant sur une représentation locale du contexte. Le chapitre 9 aborde l'exploitation et la combinaison de différents espaces vectoriels. Enfin, le chapitre 10 est consacré au système multi-sources qui regroupe les différentes contributions.

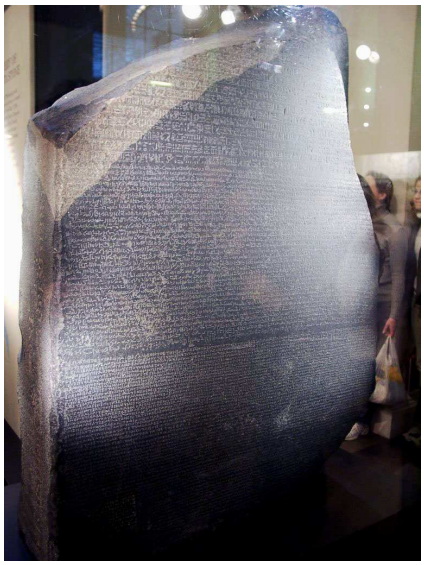
Nous concluons le présent manuscrit en dressant un bilan et en proposant plusieurs perspectives de recherche.

I

État de l'art

1

Corpus multilingue



Emprunté au latin *corpus iuris*, le terme *Corpus* (recueil) faisait référence à une collection de droit romain. Il est aussi défini comme un ensemble de textes, de documents fournis par une tradition ou rassemblés pour une étude linguistique.

Nous évoquons dans ce chapitre quelques unes des principales définitions du terme *Corpus* ainsi que son usage dans l'extraction de lexiques bilingues.

Introduction

L'extraction de lexiques bilingues à partir de corpus comparables introduit de manière intrinsèque l'utilisation de lexiques et de corpus. Nous aurons donc à traiter ces deux notions tout au long de notre étude. Parler du lexique d'une langue revient à faire référence à l'ensemble des mots de son vocabulaire d'une manière générale et à sa terminologie d'une manière plus spécifique. S'atteler à la tâche de l'extraction terminologique bilingue suppose donc l'utilisation de corpus de langue de spécialité. La terminologie est une partie intégrante de la science, d'ailleurs on ne pourrait parler de science sans parler de termes scientifiques ou techniques, et sans aller dans une pérégrination terminologique ou une étude scientifique approfondie, on associe souvent un terme au domaine auquel il fait référence ou auquel il appartient. Il existe depuis plusieurs années maintenant un engouement des scientifiques et des chercheurs envers l'extraction terminologique et la création automatique de lexiques à partir de corpus multilingues. [Catizone *et al.*, 1989] furent parmi les premiers à s'intéresser à l'extraction de lexiques multilingues à partir de corpus parallèles. Les lexiques extraits de corpus parallèles portaient généralement sur des termes simples, mais les recherches se sont dans un laps de temps relativement court, tournées vers des phénomènes plus complexes tels que l'extraction de collocations, d'expressions, de phrases, etc. [Daille *et al.*, 1994, Dagan et Church, 1994].

Au vu des différents problèmes dont souffrent les corpus parallèles, notamment du fait de leur rareté, et encore plus dans des domaines de spécialités et pour des couples de langues ne faisant pas apparaître l'anglais [Morin *et al.*, 2004], il devenait nécessaire d'y trouver une alternative. C'est ce qu'ont réussi à faire [Fung, 1995a] et [Rapp, 1995] puisqu'ils introduisirent en 1995 des méthodes permettant d'aligner des corpus non parallèles (corpus parallèles bruités puis corpus comparables). Il y eut par la suite un grand nombre de travaux portant d'abord sur les termes simples, puis sur les termes complexes, collocations, etc. [Rapp, 1999, Chiao, 2004, Déjean et Gaussier, 2002, Morin et Daille, 2004, Laroche et Langlais, 2010].

Les corpus multilingues trouvent aussi leur utilité dans les travaux de désambiguïsation du sens des mots [Brown *et al.*, 1991], en recherche d'information translingue CLIR (Cross Language Information Retrieval) [Oard et Diekema, 1998] et dans l'aide à la traduction. Si la plupart des ambiguïtés relèvent du niveau lexical, le recours à un corpus de même thématique dans une autre langue peut permettre de lever l'ambiguïté [Brown *et al.*, 1991]. Dans le domaine de la recherche d'information, il est souvent utile de rechercher des documents dans une langue à l'aide de requêtes dans une autre langue, autrement dit, d'effectuer des recherches multilingues sur le web [Véronis et Langlais, 2000]. Pour les traducteurs, les dictionnaires et lexiques bilingues sont parfois des ressources insuffisantes et s'appuyer sur des corpus multilingues en observant la langue dans son usage peut être un moyen très utile pour lever les ambiguïtés.

1.1 Corpus

Selon [Sinclair, 1996][1] *« Un corpus est une collection de données langagières qui sont sélectionnées selon des critères linguistiques et extra-linguistiques explicites pour servir d'échantillon d'emplois déterminés d'une langue »*. En somme, c'est un ensemble de documents traitant généralement d'un même thème, sujet ou domaine qui sera dédié à une utilisation spécifique. Les travaux en linguistique de corpus qui portaient majoritairement sur la langue anglaise, se sont ouverts depuis les années 1980 sur les langues européennes et asiatiques [McEnery et Xiao, 2007b]. De là, sont apparus les corpus multilingues qui représentent des ressources très utiles dans de nombreux domaines comme la traduction automatique, l'aide à la traduction, l'extraction d'informations multilingues, etc. [McEnery et Xiao, 2007b] listent trois types de corpus multilingues : ils appellent corpus de type A les corpus parallèles, corpus de type B les corpus comparables et corpus de type C la combinaison des types A et B. Il existe dans la littérature d'autres appellations pour les corpus multilingues. [Fung et Mckeown, 1997] parlent de corpus non-parallèles ou de corpus parallèles bruités et [Rapp, 1995] de corpus non-liés. Les corpus multilingues sont donc des corpus monolingues mis en relation.

1.1.1 Corpus parallèles

On ne peut parler de corpus parallèles sans citer l'exemple le plus emblématique et le plus ancien représentant des textes parallèles qui est la pierre de « Rosette » (figure 1.1). Découverte en 1799 par l'officier Napoléonien Pierre-François-Xavier Bouchard lors de travaux de terrassement dans une ancienne forteresse turque, la pierre de « Rosette » est un fragment de stèle d'origine égyptienne qui offre une représentation en deux langues (égyptien ancien et grec ancien) de trois systèmes d'écritures (hiéroglyphes, démotique et grec). Cette pierre qui relatait les honneurs rendus au roi Ptolémée V par les temples d'Égypte, permit à « Champollion » en 1822 de découvrir la clé du déchiffrement de l'écriture hiéroglyphique [Véronis et Langlais, 2000]. Si la pierre de « Rosette » constitue l'exemple le plus célèbre de corpus parallèles datant de l'Antiquité, il existe plusieurs autres inscriptions multilingues qui jalonnent à peu près toutes les périodes depuis l'apparition de l'écriture. Nous retrouvons ainsi des combinaisons de langues variées comme : sumérien/akkadien, babylonien/élamite, babylonien/hittite, vieux-perse/babylonien, etc.

Les corpus parallèles aussi appelés « Bi-textes », sont donc des paires de documents qui sont des traductions mutuelles. L'alignement ou le parallélisme peut être au niveau des textes, des phrases ou des mots. [Bowker et Pearson, 2002] définissent un corpus parallèle comme étant un ensemble de textes accompagnés de leurs traductions dans une ou plusieurs langues.

Bien qu'il y ait eu quelques tentatives de traduction automatique basées sur les corpus parallèles à la fin des années 1950 [Koutsoudas et Humecky, 1957], celles-ci se révélaient être sans réel succès, sans doute à cause des limitations de l'époque, à

1. A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language.

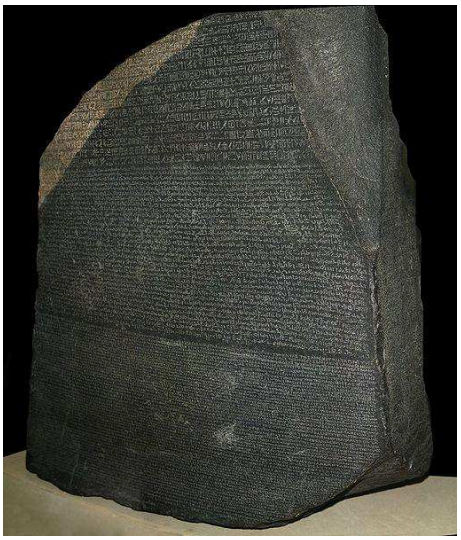


FIGURE 1.1 – Une reproduction de la pierre de Rosette au British Museum

savoir la capacité de stockage et de calcul des ordinateurs, ainsi que les difficultés de saisie de quantités importantes de textes. C'est seulement à partir des années 1980 que les textes parallèles ont commencé à être exploités de façon systématique dans le cadre du traitement automatique des langues. Parmi les corpus parallèles de référence nous pouvons citer :

- **Le corpus Hansard** : Créé dans les années 1980, ce corpus est composé de textes anglais et français extraits des transcriptions des débats du parlement canadien de 1970 à 1988. Ce corpus contient une dizaine de millions de mots [Véronis et Langlais, 2000].
- **Le corpus Europarl** : Ce corpus rassemble des textes du parlement européen dans 11 langues, avec plus de 20 millions de mots par langue [Koehn, 2004].
- **Le corpus de la Bible** : Ce corpus, construit en 1999, a été organisé par [Resnik *et al.*, 1999] en 13 langues.
- **Le corpus JRC-ACQUIS** : Ce corpus est constitué de textes des acquis communautaires de l'Union Européenne (EU) en majorité législatif dans 20 des langues officielles de l'UE, chaque langue contient une moyenne de 9 millions de mots [Patry et Langlais, 2010].

Voici quelques caractéristiques des corpus parallèles telles que définies par [Fung, 1998] :

- Les mots ont un seul sens par corpus ;
- Les mots ont une seule traduction possible ;
- Il n'existe pas de traduction manquante dans le corpus cible ;
- Les fréquences des mots dans le corpus bilingue sont comparables ;
- Les positions des mots dans le corpus bilingue sont comparables.

On retrouve les corpus parallèles dans plusieurs applications liées à la lexicographie et la terminologie, à la traduction ou encore à la recherche d'information, etc. Ainsi les corpus parallèles peuvent servir à :

- L'extraction de dictionnaires ;

- L'extraction de listes terminologiques bilingues ;
- La constitutions de mémoires de traduction ;
- La constitution de ressources pour des langues peu dotées ;
- L'apprentissage de modèles de langage pour la traduction automatique ou la recherche d'information translingue ;
- L'extraction de connaissances pour la recherche d'information multilingue ;
- L'aide à la désambiguïsation, etc.

L'extraction de lexiques bilingues à partir de corpus a initialement été entreprise en considérant des corpus parallèles. Ces textes restent néanmoins des ressources difficiles à obtenir [Fung, 1998, Morin, 2009], c'est pourquoi des chercheurs se sont tournés vers l'exploitation des corpus comparables, ressource que nous présentons dans la section suivante.

1.1.2 Corpus comparables

Face aux insuffisances des corpus parallèles, les recherches se sont d'abord tournées vers les corpus parallèles bruités [Fung, 1995b], c'est-à-dire des corpus composés de documents en relation de traduction mais ne respectant pas toutes les contraintes des corpus parallèles. Par la suite, les recherches se sont penchées sur des corpus non-parallèles [Fung, 1995a, Rapp, 1995] avant d'exploiter les corpus comparables comme ressources linguistiques pour l'extraction lexicale.

Selon [Teubert, 1996] : *« Les corpus comparables sont des corpus composés de deux ou de plusieurs langues ayant une composition ou une structure similaire (ou quasi-similaire) »*.

[Bowker et Pearson, 2002] présentent les corpus comparables comme étant des corpus composés de documents en plusieurs langues, qui ne sont pas des traductions, mais qui partagent certaines caractéristiques. Ces caractéristiques peuvent être qualitatives : extra-linguistiques (auteur, période, thème) et/ou catégories pré-établies (genre, type de discours...), mais aussi quantitatives, et donc basées sur les mesures de fréquence de certains traits linguistiques.

Selon [Déjean et Gaussier, 2002] : *« Deux corpus de deux langues l1 et l2 sont dits comparables s'il existe une sous-partie non négligeable du vocabulaire du corpus de langue l1, respectivement l2, dont la traduction se trouve dans le corpus de la langue l2, respectivement l1 »*.

[Goeuriot, 2009] précise qu'il existe deux principales catégories de corpus comparables. Elle parle de corpus généralistes composés généralement d'articles de journaux, et de corpus comparables spécialisés qui comme leur nom l'indique, sont des corpus composés de documents spécialisés dans un domaine spécifique, souvent scientifique.

Voici quelques caractéristiques des corpus comparables telles que définies par [Fung, 1998] :

- Les mots ont plusieurs sens par corpus ;
- Les mots ont plusieurs traductions possibles ;
- Il peut y avoir des traductions manquantes dans le corpus cible ;

- Les fréquences des mots dans le corpus bilingue ne sont pas comparables ;
- Les positions des mots dans le corpus bilingue ne sont pas comparables.

1.2 Corpus parallèles versus corpus comparables

Les corpus parallèles sont une ressource fondamentale utilisée dans la plupart des systèmes de traduction automatique. Leur structure généralement basée sur un alignement au niveau de la phrase permet de construire des modèles probabilistes robustes. Cela dit, la difficulté de disposer de cette ressource surtout dans les domaines spécialisés, a fait émerger les corpus comparables qui sont devenus une alternative aux corpus parallèles. Si les corpus comparables sont encore loin d'égaler les performances des corpus parallèles, leur capacité à capturer une information plus proche du cadre réel de traduction pouvant éviter les erreurs de traduction littérales, a poussé certains chercheurs à exploiter conjointement les corpus parallèles et comparables pour tirer avantage de ces deux ressources. Ainsi, un troisième axe de recherche fait apparaître plusieurs travaux qui visent à améliorer les systèmes de traduction automatique et l'extraction terminologique bilingue [Fung et Cheung, 2004, Munteanu et Marcu, 2005, Munteanu et Marcu, 2006, Abdul-Rauf et Schwenk, 2009, Smith *et al.*, 2010, Morin et Prochasson, 2011, Hunsicker *et al.*, 2012]. [Munteanu et Marcu, 2005, Munteanu et Marcu, 2006] par exemple, présentent des méthodes d'extraction de phrases ou de segments parallèles à partir de corpus comparables. [Babych *et al.*, 2008] utilisent des stratégies de détection de traductions indirectes. Ces stratégies sont apprises à partir de corpus parallèles et généralisées en utilisant des corpus comparables. L'intégration de cette méthodologie dans l'architecture d'un système de traduction automatique a montré des résultats prometteurs. [Afli *et al.*, 2012] quant à eux, extraient des données parallèles à partir de corpus comparables multimodaux. Récemment, [Aker *et al.*, 2013] ont présenté une méthode d'extraction terminologique bilingue basée sur un classifieur SVM entraîné sur un corpus parallèle, avec des résultats proches de 100% de précision sur une vingtaine de langues européennes. Ces quelques exemples constituent un petit échantillon des travaux qui visent à enrichir les méthodes d'alignement existantes par les deux types de corpus présentés. S'affranchir des corpus parallèles ou comparables semble être une idée désuète dans le cas où les deux ressources sont disponibles.

2

Méthodes d'alignement à partir de corpus comparables

Introduction

Les lexiques bilingues sont une ressource importante pour différentes applications relevant du traitement automatique des langues comme en traduction assistée par ordinateur ou en recherche d'information interlangue. Bien que les travaux s'appuyant sur des corpus parallèles aient montré de très bons résultats, ce type de corpus reste difficile à collecter [Fung et Yee, 1998] et plus particulièrement quand il s'agit de traiter des corpus spécialisés ou des couples de langues rares ou moins usitées [Morin *et al.*, 2004]. L'exploitation des corpus comparables a marqué un tournant dans la tâche d'extraction de lexiques bilingues, et suscite un intérêt constant depuis le milieu des années 1990 grâce à l'abondance et la disponibilité de tels corpus [Rapp, 1995, Fung, 1995a, Rapp, 1999, Déjean *et al.*, 2002, Gaussier *et al.*, 2004, Morin *et al.*, 2004, Laroche et Langlais, 2010]. L'essor du Web ayant sensiblement facilité la collecte de grandes quantités de données multilingues, les corpus comparables se sont naturellement imposés comme une alternative aux corpus parallèles. Ils ont donné lieu à plusieurs travaux dont le dénominateur commun est l'hypothèse selon laquelle les mots qui sont en correspondance de traduction ont de grandes chances d'apparaître dans les mêmes contextes [Rapp, 1999]. Cette hypothèse découle directement de la proposition souvent citée de [Firth, 1957] : *« On reconnaît un mot à ses fréquentations »* [1]. [Rapp, 1995] et [Fung, 1995a] ont été les premiers à introduire les corpus comparables. Ils se sont appuyés sur l'idée de caractérisation du contexte des mots, contrairement aux travaux s'appuyant sur les corpus parallèles, qui eux se basaient sur des informations positionnelles. En 1998, [Fung, 1998] a introduit l'*approche directe*, reprise dans de nombreux travaux, notamment ceux de [Rapp, 1999]. Cette approche se base sur la comparaison des contextes des termes afin d'extraire des couples en relation de traduction. Par la suite, une partie des travaux a porté sur l'adaptation et l'amélioration de cette méthode à différents types de corpus (corpus de langue générale ou de spécialité) et à différentes langues et différents

1. *« You shall know a word by the company it keeps »*

types de termes (termes simples, termes complexes, collocations, etc.) [Déjean et Gaussier, 2002], [Morin et Daille, 2004]. De nouvelles méthodes ont également été proposées telles que l'approche par similarité interlangue [Déjean et Gaussier, 2002], l'utilisation de l'Analyse en Composantes Canoniques (CCA) [Haghighi *et al.*, 2008]. Récemment, [Li et Gaussier, 2010] et [Li *et al.*, 2011] se sont intéressés à l'aspect inverse qui consiste à améliorer la comparabilité des corpus comparables afin d'augmenter l'efficacité des méthodes d'extraction de lexiques bilingues. Il sera présenté dans ce qui suit, l'état de l'art des principales méthodes appliquées à l'extraction de lexiques bilingues à partir de corpus comparables.

2.1 Travaux fondateurs

2.1.1 Premiers travaux de Reinhard Rapp (1995)

[Rapp, 1995] fut l'un des premiers à proposer une méthode exploitant les corpus comparables pour l'extraction de lexiques bilingues. Cette méthode est basée sur l'hypothèse qu'il existe une corrélation entre les modèles de cooccurrences des mots qui sont des traductions les uns des autres, et cooccurrent dans des textes de langues différentes. [Rapp, 1995] utilise des matrices pour représenter les cooccurrences entre les mots. Les tables 2.1a et 2.1b illustrent la représentation d'un vocabulaire anglais de 6 mots et de leur traduction en allemand. Dans ces matrices, les entrées appartenant à ces paires de mots qui cooccurrent plus souvent que par chance sont marquées par une étoile. En général, l'ordre des mots des lignes et colonnes de la matrice de cooccurrences est indépendant, si maintenant l'ordre des mots de la matrice des mots anglais est permuté jusqu'à ce que le modèle résultant soit le plus similaire à celui de la matrice des mots allemand (voir table 2.1c), alors ceci augmente la vraisemblance que les mots anglais et allemands soient en correspondance de traduction. Le mot anglais n de la matrice 2.1a est donc la traduction du mot allemand n de la matrice 2.1b. Une simulation a été conduite dans le but de vérifier l'hypothèse concernant la similarité des modèles de cooccurrences. Dans cette expérience, pour un vocabulaire anglais-allemand équivalent, deux matrices de cooccurrence ont été construites et comparées. une liste de 100 mots a été utilisée pour le vocabulaire anglais et une traduction de cette liste a été construite pour le vocabulaire allemand. Le calcul des cooccurrences des mots anglais, respectivement allemands a été effectué sur des corpus de 33 et 46 millions de mots.

- Corpus anglais : The Brown corpus du journal de Wall Street, encyclopédie Grolier's Electronic et les résumés scientifiques dans différents domaines ;
- Corpus allemand : Journaux de Frankfurter Rundschau, Die Zeit et Mannheimer Morgen.

Le calcul de la fréquence des cooccurrences a été effectué pour chaque couple de mots dans le corpus anglais respectivement allemand. Les études concernant les cooccurrences notamment par [Wettler et Rapp, 1993] ont montré qu'il est préférable de réduire l'influence de la fréquence des cooccurrences des couples de mots pour la prédiction des associations de mots. Les meilleurs résultats ont été obtenus en utilisant la formule suivante :

(a)		1	2	3	4	5	6
blue	1		*			*	
green	2	*		*			
plant	3		*				
school	4						*
sky	5	*					
teacher	6				*		

(b)		1	2	3	4	5	6
blau	1		*	*			
grun	2	*				*	
Himmel	3	*					
Lehrer	4						*
Pflanze	5		*				
Schule	6				*		

(c)		1	2	5	6	3	4
blue	1		*	*			
green	2	*				*	
sky	3	*					
teacher	4						*
plant	5		*				
school	6				*		

TABLE 2.1 – Représentation des correspondances entre mots anglais et allemands

$$A_{i,j} = \frac{(f(i\&j))^2}{f(i)f(j)} \tag{2.1}$$

où $f(i\&j)$ est la fréquence de cooccurrence des mots i et j, $f(i)$ et $f(j)$ la fréquence du mot i respectivement j. Par souci de comparaison, la formule originale de cooccurrence (2.2), et une formule similaire à celle de l'information mutuelle (2.3) ont aussi été utilisées.

$$A_{i,j} = f(i\&j) \tag{2.2}$$

$$A_{i,j} = \frac{f(i\&j)}{f(i)f(j)} \tag{2.3}$$

Le but est de déterminer dans quelle mesure la similarité entre les deux matrices des corpus anglais et allemands dépend de l'ordre des mots. La mesure de similarité entre matrices qui a été utilisée est la somme en valeur absolue des différences des valeurs des matrices aux positions correspondantes.

$$s = \sum_{i=1}^{N} \sum_{j=1}^{N} |E_{i,j} - G_{i,j}| \tag{2.4}$$

avec $E_{i,j}$ qui correspond à une valeur à la i ème ligne et j ème colonne de la matrice de représentation des mots anglais E et $G_{i,j}$, qui correspond à une valeur à la i ème ligne et j ème colonne de la matrice de représentation des mots allemands G.

La simulation a été effectuée en permutant aléatoirement l'ordre des mots dans la matrice du vocabulaire allemand, puis en calculant la similarité s avec la matrice du vocabulaire anglais. Pour chaque permutation, l'auteur a déterminé combien de mots c ont été décalés à des positions différentes de celles de la matrice allemande originale. La simulation a été poursuivie jusqu'à ce que pour chaque valeur de c, un ensemble de 1000 valeurs de similarité ait été construit. Il a été constaté que même pour des corpus anglais/allemand non reliés entre eux, les modèles de cooccurrences des mots sont fortement corrélés. Une croissance monotone indique qu' il est possible en principe de trouver une correspondance entre mots de deux matrices de langues différentes, en permutant aléatoirement l'une des deux matrices jusqu'à ce que la fonction de similarité s atteigne un minimum et indique une similarité maximale, ceci dit, il y a un risque de trouver des minima locaux des fonctions de similarités.

Comme perspectives, il y a à résoudre :

- Le problème du traitement limité à un certain type de vocabulaire ;
- Le problème de l'ambiguïté de la traduction des mots.

2.1.2 Premiers travaux de Pascale Fung (1995)

[Fung, 1995a] rejoint [Rapp, 1995] dans l'idée d'exploiter les corpus comparables. Elle souligne que les algorithmes existant à cette époque se basaient sur des statistiques fréquentielles de cooccurrences, de longueur ou de position extraites à partir des corpus parallèles. À partir de ces observations, elle formule l'hypothèse qu'il existe une corrélation statistique entre un mot et sa traduction dans des corpus non-parallèles (comparables). [Fung, 1995a] utilise les notions de ***contextes productifs*** et de ***contextes rigides***. Elle suggère que si des mots sont représentés par des contextes productifs alors leurs traductions seront aussi représentées par des contextes productifs, et de la même manière, pour les mots de contextes rigides, leurs traductions auront aussi des contextes rigides. Pour mesurer le degré de productivité du contexte d'un mot dans un domaine donné, l'auteur propose une mesure appelée ***hétérogénéité du contexte***.

[Fung, 1995a] définit pour un mot w son vecteur d'hétérogénéité de contexte comme étant une paire ordonnée (x, y) telle que :

- a : nombre de mots différents précédant immédiatement w ;
- b : nombre de mots différents suivant immédiatement w ;
- c : nombre d'occurrences de w dans le corpus ;
- $x = \frac{a}{c}$, représente l'hétérogénéité à gauche du mot w ;
- $y = \frac{b}{c}$, représente l'hétérogénéité à droite du mot w.

Pour mesurer la distance entre deux vecteurs de contexte hétérogènes, [Fung, 1995a] a choisi la distance euclidienne représentée dans la formule ci dessous :

$$\varepsilon = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{2.5}$$

L'equation 2.5 suppose que l'ordre des bigrammes dans les deux langues soit similaire, or ceci n'est pas toujours le cas, car les noms peuvent apparaître avant

ou après un verbe comme sujet ou objet. Partant de cette observation, une autre mesure euclidienne prenant en compte cette caractéristique a été proposée :

$$\varepsilon = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (x_1 - y_2)^2 + (y_1 - x_2)^2} \tag{2.6}$$

Pour les expérimentations, l'auteur a utilisé un corpus anglais/chinois concernant les débats législatifs de l'assemblée de Hong Kong entre 1988 et 1992. En prenant comme exemple le mot anglais ***air*** appartenant à la partie anglaise du corpus, son hétérogénéité contextuelle est calculée comme suit :

- $x = \frac{119}{176}$, $y = \frac{47}{176}$, $(x, y) = (0.676, 0.267)$

L'hétérogénéité contextuelle de la traduction du mot ***air*** est donnée par :

- $x = \frac{29}{37}$, $y = \frac{17}{37}$, $(x, y) = (0.784, 0.459)$

Prenons maintenant le mot z qui correspond à la traduction chinoise du mot ***adjournment*** qui apparaît pratiquement le même nombre de fois que le mot ***air*** :

- $x = \frac{37}{175}$, $y = \frac{16}{175}$, $(x, y) = (0.211, 0.091)$

Nous pouvons constater que bien que ***air*** et z apparaissent respectivement 176 et 175 fois dans le corpus, leur hétérogénéité contextuelle est notablement différente, indiquant que ***air*** a beaucoup plus de contexte productif que z. Ceci se confirme avec la distance euclidienne qui est de 0.221 pour ***air*** et sa bonne traduction alors qu'elle est de 0.497 entre ***air*** et z.

[Fung, 1995a] a montré qu'il existe une corrélation statistique entre des couples de mots qui sont en relation de traduction grâce à la mesure de ***l'hétérogénéité du contexte***. Elle a aussi expliqué qu'il y avait différents moyens d'améliorer cette mesure en incluant des informations linguistiques concernant l'anglais et le chinois, telles que les correspondances de classe et d'ordre entre les mots, ainsi que l'utilisation de contextes plus larges.

2.1.3 Travaux de Fung et McKeown (1997)

Dans la continuité des travaux de [Rapp, 1995] sur les corpus dits non parallèles (comparables) où ce dernier proposait une méthode se basant sur les cooccurrences entre mots pour construire un modèle interlingue, d'une manière un peu différente [Fung et Mckeown, 1997] présentent une méthode qui essaie de caractériser les corrélations entre les termes en utilisant une matrice de relations inter-mots, appelée aussi ***WoRM*** (Word Relation Matrix). Cette matrice est un modèle statistique construit à partir de corpus non parallèles pour des termes techniques.

L'idée d'utiliser cette méthode a d'abord été construite sur une observation des termes dans un contexte monolingue pour ensuite se voir étendre à un contexte bilingue. Les auteurs utilisent deux jeux de tests. Un premier jeu concernant un corpus monolingue anglais/anglais, et un second jeu sur un corpus bilingue anglais/japonais. La partie anglaise a été extraite du ***Wall Street Journal*** (WSJ)

entre 1993 et 1994, et la partie japonaise a été extraite à la même période du ***Nikkei Financial News***.

Le premier jeu de test est utilisé comme expérience de référence (*baseline*), le but étant de caractériser des termes dans les deux parties anglaises du corpus anglais/anglais (WSJ_1/WSJ_2), et d'essayer de retrouver pour chaque terme de la partie WSJ_1 son équivalent, c'est-à-dire lui même dans la partie WSJ_2. Comme exemple, les auteurs ont observé le terme ***debentures*** et ont illustré ses possibles corrélations avec tous les termes dans les deux parties du ***WSJ***. Elles ont listé les observations suivantes :

- *debentures* cooccurre plus souvent avec des termes comme : *million* et *due* dans les deux parties du corpus anglais ;
- *debentures* est moins lié à *engineering* car les deux termes n'apparaissent pas dans les mêmes contextes ;
- Étant donné tous les mots dans la partie une du corpus, *debentures* est fortement corrélé avec une sous-partie des termes du corpus (*million*, *due*, *convertible*, *subordinated*, etc.).

À partir des observations citées ci-dessus, [Fung et Mckeown, 1997] ont proposé un algorithme pour extraire des termes en relation de traduction à partir de corpus non-parallèles, ici le corpus anglais/japonais. L'algorithme se présente comme suit :

- Soit une liste de termes dont nous connaissons les traductions (*seed words*), ce sont les termes du dictionnaire bilingue ;
- Pour chaque terme e de la langue l_1, une corrélation est calculée avec chaque terme source du dictionnaire appartenant donc à la langue l_1, ceci donne un vecteur de relation appelé $WoRM_1$;
- De la même manière, pour chaque terme c de la langue l_2, est calculé un vecteur de relation $WoRM_2$ à l'aide des termes appartenant au dictionnaire ;
- Une corrélation est calculée entre les vecteurs $WoRM_1$ et $WoRM_2$. Si cette corrélation est élevée alors les termes e et c sont considérés comme une paire de traduction.

La corrélation inter-mots est une information statistique importante qui a été utilisée avec succès dans l'extraction lexicale à partir de corpus parallèles. Les auteurs calculent la corrélation $W(w_s, w_c)$ à partir d'un score de vraisemblance basé sur la cooccurrence des mots dans des segments. Ces segments peuvent être des phrases, des paragraphes ou des groupes de mots délimités par des points d'ancrage. Ci-dessous la manière de calculer la corrélation inter-mots :

- $Pr(w_s = 1) = \frac{a+b}{a+b+c+d}$;
- $Pr(w_c = 1) = \frac{a+c}{a+b+c+d}$;
- $Pr(w_s = 1, w_c = 1) = \frac{a}{a+b+c+d}$;
- a = le nombre de segments où w_s et w_c apparaissent ensemble ;
- b = le nombre de segments où seulement w_s apparaît ;
- c = le nombre de segments où seulement w_c apparaît ;
- d = le nombre de segments où aucun des deux mots w_s et w_c n'apparaît.

Toutes les mesures de corrélations utilisent les scores de vraisemblance cités ci-dessus. Les auteurs ont choisi l'information mutuelle comme mesure d'association considérant qu'elle était plus appropriée pour des termes techniques de moyenne fréquence. La mesure de l'information mutuelle se présente comme suit :

$$W(w_s, w_c) = Pr(w_s = 1, w_c = 1) \times \log_2(\frac{Pr(w_s = 1, w_c = 1)}{Pr(w_s = 1) \times Pr(w_c = 1)}) \quad (2.7)$$

Une fois que toutes les corrélations entre un terme source w_x et tous les termes sources du dictionnaire ont été calculées, la matrice de corrélation obtenue (WoRM) est représentée comme suit :

$$(W(w_x, w_{s1}), W(w_x, w_{s2}), ..., W(w_x, w_{sn}))$$

Plusieurs tailles de segments ont été testées, allant des différentes délimitations par des ponctuations, jusqu'aux phrases et aux paragraphes. Les auteurs ont conclu à partir des résultats de leurs expériences que la taille des segments la plus appropriée était calculée en fonction de la fréquence des mots du dictionnaire. La raison d'un tel choix est que si certains mots du dictionnaire sont très fréquents et que la taille des segments choisis est grande (taille d'un paragraphe), alors ces termes très fréquents vont apparaître dans presque chaque segment. Dans ce cas, les chances de cooccurrence des mots du dictionnaire sont biaisantes. Le choix donc des segments de petite taille parait plus logique. Inversement, pour des mots du dictionnaire moyennement voire peu fréquents, des segments de taille plus grande sont nécessaires. L'équation suivante montre le rapport entre la taille d'un segment et la fréquence d'un terme donné :

$$TailleSegment = \frac{1}{frequence(W_s)} \quad (2.8)$$

Une première expérience pilote a été menée sur le corpus anglais WSJ_1/WSJ_2 pour tester et montrer le pouvoir discriminant de la matrice de relations WoRM. L'évaluation a porté sur deux listes d'évaluation, une première liste (liste A) de 582 mots extraite de WSJ_1 et une seconde liste (liste B) de 687 mots extraite de WSJ_2. Le but étant de mesurer la similarité entre tous les mots de ces deux listes, et voir s'il était possible de retrouver les 582 mots de la liste A dans la liste B. Les auteurs ont aussi choisi 307 mots du dictionnaire de fréquences allant de (400-3900) [2]. Considérant que les mots du dictionnaire (*seed words*) ont une fréquence élevée, le choix s'est naturellement porté sur des segments de petite taille comme fenêtre contextuelle et de la distance euclidienne comme mesure de similarité.

Les auteurs obtiennent une précision de 21% au top1 et de 58% au top100 [3] . En éliminant des listes d'évaluation les mots polysémiques, ce qui revient à 445 mots sélectionnés manuellement à partir des 582 mots de départ de la liste A, les résultats s'améliorent avec 26% au top1 et 70% au top100.

Une seconde expérience a été menée sur le corpus bilingue anglais/japonais, Cette évaluation est présentée comme une tâche difficile [4]. 1416 entrées du dictionnaire en ligne EDICT avec des fréquences entre 100 et 1000 ont été choisies. Contrairement à

2. Un tel choix a été pris afin de minimiser le nombre de mots outils.

3. Une précision au *top n* veut dire que la bonne traduction se trouve dans la liste des n premiers candidats renvoyés par le système de traduction.

4. Les deux langues n'appartiennes pas à la même famille et les deux corpus ne traitent pas des mêmes thématiques.

la première expérience, les mots du dictionnaire ont une fréquence faible comparée à la taille du corpus [5], d'où le choix d'une grande taille de segments qui correspond à la taille d'un paragraphe, et du cosinus comme mesure de similarité. Les auteurs ont mené trois types d'évaluations. Elles ont choisi comme liste de départ 19 termes japonais et ont choisi de les comparer à trois listes de termes anglais. La première liste (liste A) contient les 19 termes anglais qui sont les bonnes traductions des termes japonais. La deuxième liste (liste B) contient 312 termes anglais (19 bonnes traductions + 293 autres termes). La troisième liste (liste C) contient 402 termes dont les 19 bonnes traductions.

Les résultats obtenus au top20 sont de 52% pour la liste A, 21,1% pour la liste B et 31,6% pour la liste C. La précision moyenne obtenue est d'environ 30% ce qui d'après les auteurs est loin d'être suffisant. Cependant, ces résultats sont les premiers concernant l'extraction terminologique bilingue à partir de corpus comparables. Ces expériences ont néanmoins montré que la plupart des bonnes traductions se trouve souvent dans les 20 premiers candidats (top20) et que cette méthode peut être un bon système d'aide à la traduction. Les auteurs ont conduit des tests dans ce sens en demandant à des traducteurs de s'aider des résultats obtenus par le système pour traduire les 19 termes. Ces expériences ont montré une amélioration de 50.9% en moyenne.

2.1.4 Approche directe

Introduite par [Fung, 1998] et reprise par [Rapp, 1999] et ensuite par beaucoup d'autres chercheurs, l'*approche directe* part du principe que pour traduire un terme d'une langue à une autre, il faut dans un premier temps caractériser le mot à traduire par un vecteur représentatif de son contexte. Dans un deuxième temps, traduire ce vecteur en langue cible à l'aide d'un dictionnaire que l'on appelle aussi lexique de transfert ou lexique pivot. Puis, comparer ce vecteur avec tous les vecteurs de contexte des mots de la langue cible, et en extraire les n plus proches comme traductions candidates. Dans le domaine de l'extraction lexicale, les chercheurs entendent par contexte d'un mot ses fréquentations, c'est-à-dire les mots avec lesquels il cooccurre. Le choix de ces mots n'est pas effectué d'une manière arbitraire mais à l'aide d'une fenêtre contextuelle qui délimite, ou borne, le mot assujetti à la construction de son contexte. La taille de cette fenêtre, dont le choix est souvent fait empiriquement, va constituer le nombre de mots autour du mot à contextualiser. Par exemple, si on fait le choix d'une taille de fenêtre égale à 7, cela voudra dire que le vecteur de contexte du mot à traduire va comporter les 3 mots se trouvant à sa gauche et à sa droite, et ce, pour une occurrence du mot à traduire, et donc d'une manière générale, on répétera cette opération pour couvrir toutes les occurrences du mot à contextualiser. Les mots constituant le vecteur de contexte, que l'on peut aussi appeler ses voisins ou ses fréquentations, sont caractérisés ou pondérés par une mesure d'association. Le choix de cette mesure n'est pas toujours facile à justifier, néanmoins, de par la littérature, celles qui reviennent le plus souvent sont le test du rapport de vraisemblance [Dunning, 1993], l'information mutuelle [Fano, 1961] et le odds-ratio [Laroche et Langlais, 2010]. Ces mesures tentent d'estimer le lien ou la force qu'il peut y avoir entre un couple de mots. La comparaison entre deux vecteurs

5. Taille du corpus anglais/japonais = 7 millions de mots.

de contexte se fait sur la base d'une mesure de similarité. Les plus utilisées sont la mesure du cosinus [Salton et Lesk, 1968] et l'indice de Jaccard [Grefenstette, 1994b].

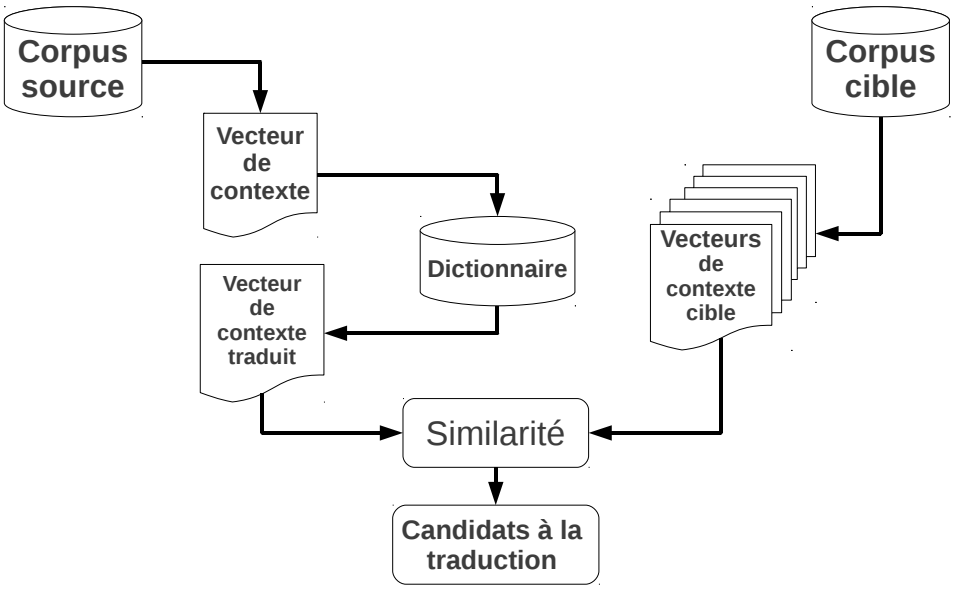


FIGURE 2.1 – Illustration de l'*approche directe*

Sur les corpus de langue générale [Fung, 1998] obtient une précision de 76% sur les 20 premiers candidats à la traduction sur des couples de mots de termes simples anglais/chinois, et ce, sur le corpus du « Wall Street Journal » et du « Nikkei Financial news ». [Rapp, 1999] quant à lui, obtient une précision de 89% sur les 10 premiers candidats sur des couples de termes simples anglais/allemand à partir d'un corpus journalistique de 85 millions de mots.

Sur les corpus de langue de spécialité, [Déjean et Gaussier, 2002], en utilisant un corpus médical de 8 millions de mots obtiennent une précision de 84% sur les 10 premiers candidats pour des couples anglais/allemand. [Chiao et Zweigenbaum, 2003] obtiennent 100% de précision pour les 25 premiers candidats sur un corpus médical pour des couples de mots français/anglais, et ce, en s'appuyant sur les mots de la langue générale pour la traduction des nouveaux mots dans un domaine de spécialité.

L'*approche directe* présente des lacunes concernant les mots de faible fréquence, c'est-à-dire les mots rares. [Pekar *et al.*, 2006] montrent que les mots de faible fréquence sont beaucoup moins bien traduits que les mots qui ont une fréquence élevée, et proposent une alternative à cela, en décrivant une extension de la mesure de similarité pour estimer la probabilité des cooccurrences des mots, avant de traduire ces vecteurs dans l'espace vectoriel des différentes langues.

Si les résultats obtenus jusqu'alors sont encourageants, l'*approche directe* présente deux inconvénients non négligeables. Le premier, est lié au lexique de transfert pour la traduction du contenu d'un vecteur de contexte. Le problème se pose si un ou plusieurs mots du vecteur de contexte ne sont pas présents dans le lexique pivot, ceci rend difficile le transfert en langue cible et conduit nécessairement à une perte

d'information voire, dans le cas le plus extrême, à une impossibilité de traduire l'ensemble du vecteur de contexte. Le deuxième manque concerne les ressources disponibles, car on constate une inadéquation des ressources existantes à un corpus donné.

2.2 Améliorations de l'approche directe

Après l'apparition de l'*approche directe* qui a marqué un tournant dans l'exploitation des corpus comparables, plusieurs propositions d'amélioration de cette méthode ont vu le jour, le principe étant de reprendre une ou plusieurs étapes de l'*approche directe* et d'essayer de les améliorer.

2.2.1 Ressources linguistiques

[Koehn et Knight, 2002] se sont intéressés à différents indices extraits des ressources linguistiques afin d'améliorer l'extraction terminologique bilingue en corpus comparables. Ils ont combiné plusieurs stratégies telles que la détection des cognats[6], les contextes similaires, la préservation des scores de similarité des mots et de leur fréquence. Chacune des stratégies est définie comme suit :

1. **Mots identiques** : Deux langues contiennent un certain nombre de mots identiques tels que *computer* ou *email* ;
2. **Écriture similaire** : Certains mots peuvent avoir une écriture similaire ou quasi-similaire étant donné les racines interlangues communes qui peuvent exister, par exemple *Freund* et *Friend* ou *Webseite* et *Website* ;
3. **Contextes similaires** : Il y a des chances pour que les traductions des mots qui occurrent dans certaines fenêtres contextuelles dans une langue source occurrent dans des contextes similaires dans une langue cible ;
4. **Mots similaires** : Les mots qui sont utilisés de la même manière dans une langue source (par exemple *wednesday* et *thursday*) devraient avoir des traductions qui sont utilisés de la même manière (par exemple *Mittwoch* et *Donnerstag*) ;
5. **Fréquence** : Dans des corpus comparables, des mots fréquents dans une langue source devraient avoir des traductions fréquentes dans une langue cible.

Les expériences ont montré que l'utilisation des cognats et des mots ayant des contextes similaires aidaient à améliorer significativement les résultats de l'alignement.

Outre les expériences de [Koehn et Knight, 2002], [Déjean et Gaussier, 2002] et [Chiao, 2004] ont utilisé respectivement le thésaurus multilingue MeSH et le métathesaurus de l'UMLS. Ils ont montré que s'appuyer sur de telles ressources complémentaires permettait d'améliorer les performances de l'extraction terminologique.

6. Cognats : mots ayant une graphie similaire ou quasi-similaire

2.2.2 Filtrage par contraintes syntaxiques et lexicales

[Sadat *et al.*, 2003] ont proposé d'améliorer l'*approche directe* en utilisant deux stratégies. La première consiste à appliquer l'*approche directe* dans les deux directions (Langue source <-> Langue cible) pour ensuite fusionner les scores. La deuxième consiste à filtrer les résultats renvoyés par l'*approche directe* en ne gardant que les traductions candidates qui respectent certaines règles préalablement définies. Ces contraintes sont de nature linguistique. Les auteurs ont expérimenté avec un corpus anglais/japonais et l'analyse morphologique des termes a conduit à extraire des règles de filtrage qui supposent qu'un nom anglais sera traduit par un nom japonais, un verbe anglais par un verbe japonais, etc. Dans ce cas, si le système renvoie des termes ne respectant pas ces règles, alors ces derniers ne seront pas considérés comme traductions candidates. Cette méthode a permis d'améliorer le gain en précision de 12% en recherche d'information interlingue.

2.2.3 Combinaison de contextes

[Shao et Ng, 2004] ont proposé une méthode qui combine l'information contextuelle à celle des translittérations, considérant qu'elles sont complémentaires. Partant du principe déjà énoncé par [Rapp, 1995] et [Rapp, 1999] qui consiste à dire que si le mot anglais e est une traduction du mot chinois c, alors leurs contextes sont similaires, les auteurs présentent ceci comme pouvant être un problème d'extraction de documents dans le domaine de la recherche d'information, où le contexte de e est considéré comme une requête $C(e)$ et le contexte de la traduction candidate est vu comme un document $C(c)$. Les auteurs s'appuient sur un modèle de langue dérivé de chaque document D puis estiment la probabilité $P(Q/D)$ qu'une requête Q soit générée par le document D. Dans le cas de la traduction, la probabilité estimée est $P(C(e)/C(c))$. De plus, les auteurs utilisent un modèle de translittération basé sur la prononciation. Chaque mot chinois est représenté par sa forme dite : "*Pinyin*", c'est-à-dire convertie dans sa forme romane. C'est une forme basée sur la phonétique. Ensuite, une estimation de $P(e/c) = P(e/pinyin)$ est calculée. La combinaison des deux modèles sur un corpus anglais/chinois a montré des résultats intéressants.

Récemment, [Andrade *et al.*, 2011] ont proposé une méthode qui combine quatre modèles statistiques (3 modèles à base de relations de dépendance syntaxique et un modèle à base de sac de mots.) et compare les dépendances lexicales pour identifier les traductions candidates. Cette approche a aussi fait état d'une amélioration significative des performances en comparaison avec l'*approche directe*.

2.2.4 Exploitation des termes du domaine

Partant de l'observation que l'*approche directe* souffrait de vecteurs de contexte bruités qui pouvaient affecter grandement les résultats, [Ismail et Manandhar, 2010] ont proposé une méthode qui permet de filtrer ce bruit et ainsi d'améliorer les

performances de l'*approche directe*. Leur technique est basée sur la notion de termes du domaine (*In-domain terms*), c'est-à-dire des termes pertinents et vus comme les plus importants du contexte d'un terme donné. Le choix de ces termes du domaine est effectué en utilisant une mesure d'association (le rapport de vraisemblance) avec un seuil t et en choisissant au maximum 100 termes. Le vecteur de contexte d'un terme à traduire W_S ne sera plus composé des termes apparaissant dans son contexte $CT(W_S)$, mais d'une sous-partie (*In-domain terms*) $IDT(W_S, W_R)$ avec W_R un terme fortement corrélé à W_S. Les termes du domaine de W_S sachant les termes de son contexte W_R sont donnés par :

$$ID(W_S, W_R) = CT(W_S) \cap CT(W_R) \tag{2.9}$$

Soit W_T la traduction potentielle de W_S, et soit $tr(W_R)$ la traduction équivalente de W_R. Le vecteur de termes du domaine en langue cible est représenté par $ID(W_T, tr(W_R))$. Les auteurs formulent l'hypothèse que la distribution des termes du domaine en langues source et cible est comparable. Pour mesurer la similarité entre les vecteurs de termes du domaine, ils utilisent une mesure de rang par histogramme (*Rank-binning*) [Wong, 2009].

Les expériences ont été menées sur un corpus anglais/espagnol de 500 000 phrases avec trois différentes listes d'évaluation. Les résultats obtenus sont meilleurs sur toutes les listes d'évaluations en utilisant l'approche proposée par les auteurs en comparaison avec l'approche directe. Une précision moyenne de 84,4% au top10 est obtenue alors que l'*approche directe* n'obtient que 59,6%.

2.2.5 Termes peu fréquents

Partant de l'observation que les mots de faible fréquence (ou mots dits rares) étaient plus difficile à traduire et afin de pallier cela, [Pekar *et al.*, 2006] ont proposé une extension du modèle de similarité pour estimer la probabilité de cooccurrence des mots [Dagan *et al.*, 1999]. Pour tenter de résoudre la faible densité de certains mots, les auteurs ont utilisé une distance moyenne (*distance-based averaging*) pour prédire les cooccurrences des mots qui n'apparaissent pas ensemble dans le corpus. Le principe est d'estimer la probabilité de cooccurrence de deux mots par analogie avec d'autres mots qui ont une distribution similaire (les k plus proches voisins) à ces derniers. Les auteurs proposent aussi de lisser les vecteurs de contexte des mots qui cooccurrent de manière faible. Cette méthode a montré une amélioration significative des résultats concernant les mots rares sur un corpus journalistique anglais/français/allemand/espagnol.

Récemment, [Prochasson et Fung, 2011] ont présenté une nouvelle méthode pour aligner les mots rares. Cette méthode est basée sur deux stratégies : la première consiste en l'utilisation de la similarité entre vecteurs de contexte des termes pour les aligner [Fung, 1998, Laroche et Langlais, 2010], et la deuxième suit l'hypothèse que des termes spécifiques et leur traduction devraient apparaître souvent dans les documents traitant des mêmes thématiques. Cette méthode est la première à obtenir des scores aussi importants entre 80 et 98% de F-mesure.

2.2.6 Symétrie distributionnelle

[Chiao, 2004] ont proposé un modèle d'alignement symétrique qui repose principalement sur l'analyse de cooccurrences. Le principe étant d'aligner les termes en utilisant l'*approche directe* dans les deux directions[7] (langue A <-> langue B) . Il s'agit de faire une correspondance croisée pour mettre en évidence les mots les plus proches dans les deux directions simultanément. La méthode se décompose en quatre étapes :

1. Prétraitement du corpus ;
2. Construction des vecteurs de contexte ;
3. Transfert des vecteurs de contexte ;
4. Calcul de la similarité croisée.

Les trois premières étapes sont similaires à l'*approche directe*, la principale contribution réside dans l'utilisation de la symétrie et l'introduction de la mesure de similarité croisée qui se formule comme suit :

$$MH(r_{sc}, r_{cs}) = \frac{1}{\frac{1}{2}(\frac{1}{r_{sc}} + \frac{1}{r_{cs}})} = \frac{2r_{sc}r_{cs}}{r_{sc} + r_{cs}} \tag{2.10}$$

où r_{sc} représente le rang calculé dans la direction d'une langue source vers une langue cible, et inversement r_{cs} représente le rang calculé de la langue cible vers la langue source. Les résultats des expériences menées par [Chiao, 2004] montrent une amélioration significative de la précision en obtenant un gain de 25% au top1 et de 20% au top30.

2.2.7 Relations de dépendance syntaxique

Dans un souci de mieux représenter le contexte d'un mot, plusieurs travaux se sont tournés vers les relations de dépendance syntaxique, notamment [Gamallo, 2008a,Garera *et al.*, 2009] où l'idée n'est plus seulement de représenter le contexte par les mots avoisinants, mais de rajouter une information supplémentaire qui spécifie le type de relation entre les mots.

[Gamallo, 2008a] ont proposé d'améliorer l'*approche directe* en utilisant une autre manière de considérer le contexte d'un mot à traduire. Le contexte n'est plus représenté par une fenêtre contextuelle (sac de mots) mais par des relations de dépendance syntaxique extraites à l'aide d'expressions régulières. Prenons la phrase suivante comme exemple :

a_D **man_N** with_P a_D **green_A jacket_N see_V** yesterday_R a_D **big_A dog_N**

7. Ceci rejoint l"approche proposée dans [Sadat *et al.*, 2003].

Le but est d'extraire des relations entre les lemmes en utilisant les catégories grammaticales. Les relations sont notées comme des triplets (head,rel,dep) avec :

- Head : représente la tête, donc le lemme dont on cherche à définir les relations ;
- Rel : représente la relation entre la tête et le dépendant ;
- Dep : représente le dépendant, donc le lemme qui est en relation avec la tête.

À partir de l'exemple précédent nous pouvons extraire les triplets suivants :

- (green,$mod_<$,jacket)
- (big,$mod_<$,dog)
- (man,$with$,jacket)
- (see,$obj_>$,dog)
- (see,$obj_<$,man)

Nous pouvons constater dans cet exemple 5 types de relations qui sont :

- $mod_<$: qui est un modificateur à gauche pour $<$. Dans ce cas, on se retrouve avec une relation de dépendance entre un adjectif qui modifie un nom (head), un nom qui apporte une information sémantique au nom (head) ;
- $mod_>$: qui est un modificateur à droite pour $>$. Dans ce cas, on se retrouve avec une relation de dépendance entre un adjectif qui modifie le nom (head) ou un nom qui modifie le nom (head). L'adjectif (ou le nom) se positionne à droite du nom (head) ;
- $obj_<$: qui est une relation d'objet qui concerne le verbe. L'argument du verbe se trouve à sa gauche ;
- $obj_>$: qui est une relation d'objet qui concerne le verbe. L'argument du verbe se trouve à sa droite ;
- $with$: qui est une relation prépositionnelle, ici avec la préposition $with$, mais cela peut être n'importe quelle préposition. Il existe deux types de relations prépositionnelles, tels que dans le premier, la tête de la relation de dépendance (head) est le nom, et dans le deuxième type c'est le verbe.

Le vecteur de contexte représentant le mot à traduire sera donc une représentation des différentes relations de dépendance entre le mot à traduire et tous les mots reliés par ces relations.

Les expériences menées sur un corpus espagnol/galicien de 20 millions de mots ont montré des résultats probants. La précision au top10 passe de 32% pour l'*approche directe* à base de sac de mots à 74% en utilisant les relations de dépendance.

Une autre approche basée sur la même idée d'utiliser les relations de dépendance a été proposée par [Yu et ichi Tsujii, 2009]. Les auteurs y définissent des relations de dépendance hétérogènes où l'hypothèse consiste à dire que dans un corpus comparable, un mot et sa traduction partagent la même tête et les mêmes modificateurs. À la différence de [Gamallo, 2008a], les auteurs utilisent un raciniseur (Stemmer) et génèrent un couple de 4 relations de dépendance hétérogènes :
($H_{NMODHEAD}, H_{SUBHEAD}, H_{OBJHEAD}, H_{NMODMod}$) avec :

- NMODHEAD : représente la relation du modificateur d'un nom avec la tête HEAD ;
- SUBHEAD : représente la relation de sujet avec la tête HEAD ;
- OBJHEAD : représente la relation d'objet avec la tête HEAD ;

– NMODMod : représente la relation du modificateur d'un nom avec un modificateur Mod.

Les résultats obtenus sur un corpus anglais/chinois (1 132 492 phrases anglaises et 665 789 phrases chinoises) montrent un gain de 57% environ au top5 et de 28% environ au top10.

2.2.8 Points d'ancrage

Partant de l'observation que les résultats de l'*approche directe* chutaient fortement lors de l'utilisation de corpus de taille modeste, et pour pallier cela, [Prochasson et Morin, 2009] ont proposé une nouvelle contribution à l'*approche directe* qui est l'utilisation des points d'ancrage. Le principe consiste à s'appuyer sur le vocabulaire spécialisé du domaine pour renforcer la représentation du contexte des termes à traduire. Les points d'ancrage sont des éléments de confiance dont la présence ou l'absence est particulièrement discriminante pour caractériser un terme. [Prochasson et Morin, 2009] citent quelques propriétés que doivent respecter les points d'ancrage :

– Faciles à identifier ;
– Pertinents ;
– Peu polysémiques.

Les auteurs se sont intéressés aux translittérations [8] et aux composés savants [9] comme points d'ancrage. Les expériences menées sur les corpus spécialisés anglais/japonais et français/japonais traitant les thématiques de l'alimentation et du diabète ont montré un gain significatif en précision. Les résultats obtenus pour le corpus anglais/japonais montrent un gain de 18,2% au top1 en utilisant les translittérations et les composés savants comme points d'ancrage. Un gain de 8,2% au top10 pour les translittérations et un gain de 11,2% pour les composés savants sont obtenus. En revanche, les gains obtenus pour le corpus français/japonais sont moins important. Les résultats montrent un gain de 10% au top1 en utilisant les composés savants mais pas de gain pour les translittérations [10]. Concernant le top10, les résultats montrent un gain de 2,8% pour les translittérations et 5,6% pour les composés savants.

2.3 Approche par similarité interlangue

Introduite par [Déjean et Gaussier, 2002] cette approche lève le principal inconvénient de l'*approche directe*, qui est l'impossibilité dans certains cas de traduire les mots d'un vecteur de contexte. L'approche par similarité interlangue utilise le principe des k plus proches voisins du mot à traduire pour effectuer le transfert en langue cible. Elle suppose que deux mots sont traductions l'un de l'autre avec une forte probabilité si leurs similarités avec les entrées des ressources bilingues

8. Translittération : emprunt d'un mot d'une langue source vers une langue cible.
9. Composés savants : mots construits à partir de racines spécifiques.
10. Ceci peut s'expliquer par le fait que les translittérations sont plus rares.

disponibles sont proches. Ces k plus proches voisins, qui sont donc des mots du lexique de transfert, vont servir de pont entre les corpus source et cible.

Pour un mot candidat à la traduction, la première étape consiste à construire son vecteur de contexte ainsi que les vecteurs de contexte de tous les mots du lexique pivot. Puis, choisir à partir de ce lexique les k plus proches voisins du mot à traduire grâce à une mesure de similarité (Cosinus ou Jaccard ou autre). Ensuite, et à partir des traductions de ses k plus proches voisins, construire leurs vecteurs de contexte en langue cible et les comparer avec tous les vecteurs de contexte des mots de la langue cible, et ainsi en choisir les n premiers grâce à la combinaison linéaire qui suit :

$$P(w_2 \mid w_1) \quad = \quad \sum_C P(C \mid w_1) P(w_2 \mid C) \tag{2.11}$$

avec w_i un mot, C_i une entrée de la ressource bilingue en langue i, et $P(w_2 \mid w_1)$ la probabilité d'associer le mot w_2 au mot w_1.

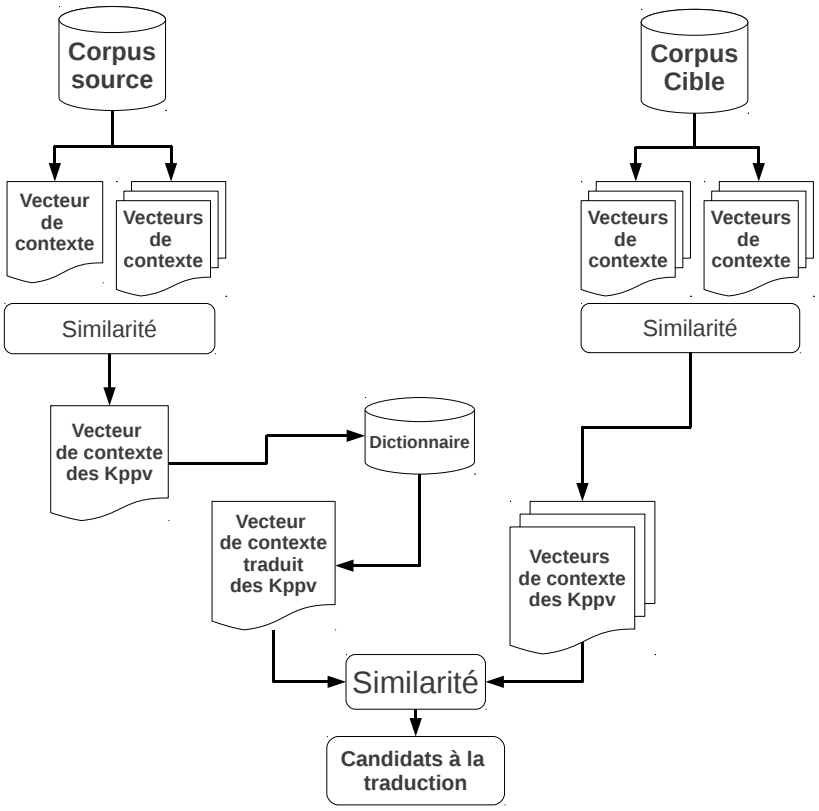


FIGURE 2.2 – Illustration de l'approche par similarité interlangue

[Déjean et Gaussier, 2002] ont utilisé 845 résumés d'articles scientifiques dans le domaine médical rédigés en anglais et en allemand ce qui revient à environ 100 000 mots pour chaque langue. Comme ressources bilingues, ils ont utilisé le dictionnaire ELRA anglais/allemand ainsi que le thésaurus anglais du domaine médical Mesh et sa version allemande DMD (environ 15 000 entrées), et un deuxième jeu de données de la tâche GIRT de CLEF 2002 et qui porte sur les sciences sociales. Le thésaurus

GIRT contient environ 10 000 entrées allemandes avec leurs traductions anglaises, le corpus est constitué d'environ 86 000 articles.

En utilisant le thésaurus Mesh, l'*approche directe* montre des scores de précisions de 44% pour les 10 meilleurs candidats (top10) et 57% pour les 20 meilleurs (top20), qui sont des résultats en deçà des évaluations fournies dans les articles de [Fung, 1998] 76% et [Rapp, 1999] 89%. [Déjean et Gaussier, 2002] pensent que la différence peut s'expliquer par la différence de taille entre leur corpus médical et les corpus utilisés précédemment. Pour la première version de leur approche par similarité interlangue qu'ils ont appelé la méthode plate, par opposition à celle utilisant la hiérarchie, les meilleurs résultats obtenus sont de 43% pour les 10 meilleurs candidats et 51 % pour les 20 meilleurs candidats, en utilisant 50 entrées, ce qui donne un léger avantage pour l'*approche directe*. Pour la méthode hiérarchique, ils obtiennent un score de 50% pour les 10 meilleurs candidats et 63% pour les 20 meilleurs candidats, pour un nombre d'entrées égal à 50, ce qui est meilleur que l'*approche directe* et la méthode plate mais cela reste en deçà des résultats fournis par [Fung, 1998] et [Rapp, 1999].

En utilisant le thésaurus GIRT et en comparant la méthode plate et l'*approche directe*, les résultats montrent une différence très significative des scores car [Déjean et Gaussier, 2002] obtiennent 79% (respectivement 84%) pour la méthode plate sur les 10 premiers candidats (respectivement les 20 premiers candidats), alors qu'ils obtiennent seulement 35% (respectivement 42 %) pour l'*approche directe* sur les 10 premiers candidats (respectivement les 20 premiers candidats). Une première explication à propos de cette différence de résultats donnée par [Déjean et Gaussier, 2002], pourrait être donnée par la taille des corpus, qui est en effet de 100 000 mots et 9 millions de mots. Une autre explication pourrait être la présence de multi-mots qui ne sont pas pris en compte par l'*approche directe* contrairement à la leur.

En utilisant une combinaison des approches directe et par similarité interlangue, [Déjean et Gaussier, 2002] montrent que cette fusion est des plus efficaces. En utilisant une simple combinaison linéaire entre les méthodes, les résultats sont très probants comme le montre la table 2.2.

Approche directe combinée avec :	10 meilleurs
Plate 1	79%
Plate 100	80%
Plate 200	83%
HIER 10	82%
HIER 20	84%
HIER 50	84%

TABLE 2.2 – Résultats de la combinaison des méthodes

Nous pouvons constater qu'avec la combinaison, la méthode hiérarchique est meilleure que la méthode plate et qu'il faut 200 classes pour la méthode plate pour qu'elle rivalise avec la méthode hiérarchique qui elle, n'a besoin que de 20 classes.

2.4 Approche géométrique

[Gaussier *et al.*, 2004] ont présenté dans leur article une vue géométrique concernant l'extraction lexicale bilingue à partir de corpus comparables. Ils donnent une interprétation géométrique de l'*approche directe* ainsi que des problèmes de synonymie et de polysémie dont elle souffre. Par la suite, ils présentent 3 nouvelles approches afin de résoudre ces problèmes qui sont : une extension de l'*approche directe*, la CCA (Canonical Correlation Analysis) et la Multilingual PLSA (Probabilistic Latent Sementic Analysis).

Selon [Gaussier *et al.*, 2004], l'utilisation d'un dictionnaire bilingue peut induire deux problèmes. Le premier concerne la couverture du dictionnaire. Si un nombre relativement faible de mots du corpus est présent dans le dictionnaire, ceci conduit à une mauvaise couverture. Un deuxième problème est celui de la polysémie et de la synonymie des mots. Les entrées du dictionnaire étant considérées comme des vecteurs orthogonaux, des problèmes pourraient survenir si plusieurs entrées ont le même sens ou si une entrée a plusieurs sens, surtout quand un seul sens est présent dans le corpus.

Pour résoudre les problèmes de synonymie et de polysémie, [Gaussier *et al.*, 2004] ont proposé dans un premier temps une extension de l'*approche directe*, en partant du principe que les synonymes peuvent être détectés à travers les similarités entre les vecteurs de contexte [Lewis *et al.*, 1967, Grefenstette, 1994b]. Ils proposent de remplacer la projection faite par l'*approche directe*, par une projection dans un sous-espace formé par les vecteurs de contexte des mots du dictionnaire. En faisant cela, ils construisent un espace vectoriel où les entrées du dictionnaire qui sont des synonymes seront proches les unes des autres.

[Gaussier *et al.*, 2004] ont proposé ensuite une autre approche basée sur la CCA et sa version à noyau. Les données sont représentées par une matrice où les lignes correspondent aux entrées du dictionnaire et les colonnes aux mots des corpus source et cible. Cette matrice permet d'avoir deux vues connectées grâce au dictionnaire bilingue. La CCA est utilisée pour identifier les directions qui sont fortement corrélées dans les langues source et cible. Intuitivement ces directions définissent la LSA (Latent Sementic Analysis) qui tente de modéliser les relations implicites entre les couples de traductions.

Pour la troisième approche qui est basée sur la Multilingual PLSA, les données seront représentées par la même matrice que pour la méthode CCA mais au lieu de représenter pour chaque mot du corpus son nombre de cooccurrences avec chaque entrée du dictionnaire, elle va mesurer la probabilité de cooccurrence de chaque mot du corpus avec une entrée du dictionnaire et pour ce faire cette méthode va s'appuyer sur la PLSA [Hofmann, 1999].

Les tests ont été conduits sur un corpus anglais/français dérivé de CLEF 2003 correspondant au journal Los Angeles Times pour l'anglais et le journal Le Monde pour le français avec une taille d'environ 34 966 mots anglais et 21 140 mots français. Le lexique de transfert est le dictionnaire ELRA de 13 500 entrées.

L'extension de l'*approche directe* montre une amélioration du rappel du système d'extraction lexicale. Contrairement à l'*approche directe*, tous les mots du diction-

naire, qu'ils soient présents ou non dans le contexte d'un mot, peuvent être utilisés pour le traduire. Ceci peut induire du bruit en créant des relations inattendues entre les mots et les entrées du dictionnaire. Pour remédier à ce problème, [Gaussier *et al.*, 2004] proposent de n'utiliser que les couples de traductions qui sont en réelle relation avec le mot à traduire.

La Multilingual PLSA n'améliore pas les performances du système, elle est même légèrement en dessous de l'*approche directe*. En la combinant avec l'extension de l'*approche directe*, [Gaussier *et al.*, 2004] obtiennent des résultats similaires à l'*approche directe*.

La CCA ne donne pas de bons résultats non plus. Celle-ci souffre du bruit engendré par le fait que chaque direction soit définie par une combinaison linéaire, qui peut impliquer différents mots du vocabulaire. La combinaison de ces méthodes améliore de 10 points la précision moyenne.

2.5 Corpus déséquilibré

[Morin, 2009] a introduit dans son article la notion de corpus comparable déséquilibrés, c'est-à-dire des corpus comparables n'étant pas de même taille ou n'ayant pas le même rapport en termes de volume. Il montre que l'on peut obtenir, sous certaines conditions, un gain significatif dans la qualité des lexiques extraits en exploitant ce type de corpus avec l'*approche directe*. Il fait remarquer que les chercheurs se sont naturellement ou implicitement dirigés vers des corpus comparables de même taille avec un ratio entre 1 et 1.2 en général. Et que le fait de travailler avec des corpus équilibrés conduit à ramener la quantité de données de la partie du corpus la plus importante à la plus faible, et que ceci, est préjudiciable pour des corpus comparables spécialisés.

[Morin, 2009] a construit un corpus comparable spécialisé français/anglais extrait du portail Elsevier concernant la thématique du cancer du sein. Ce corpus correspond à environ 530 000 mots en français et 7,4 millions de mots en anglais. Les documents anglais ont été divisés en 14 parties contenant chacune 530 000 mots. Le dictionnaire d'un domaine général, a été construit à partir de différentes ressources du Web, et comporte 22 300 mots pour le français avec une moyenne de 1,6 traductions par entrée. Le découpage du corpus est représenté dans la table 2.3.

Les résultats varient sensiblement selon les corpus utilisés individuellement. En choisissant le français comme langue source, les variations vont de 43.4% à 54.9%. Et en prenant l'anglais comme langue source, les variations vont de 49.2% à 56,6%. L'utilisation d'une combinaison de mesures d'associations a permis d'améliorer les résultats. [Morin, 2009] obtient par exemple une précision de 62% pour le corpus 14 dans la direction anglais/français.

La combinaison des résultats entre les différents corpus a aussi permis une amélioration de la précision. L'utilisation de la moyenne arithmétique des scores de similarité pour un mot à traduire et la moyenne harmonique des positions des traductions candidates ont permis d'atteindre une précision de 67,2% (direction français/anglais) respectivement 69,7% (direction anglais/français).

	# documents	# mots
Français		
Part 1	130	7,376
Anglais		
Partie 1	118	8,214
Partie 2	114	7,788
Partie 3	101	8,370
Partie 4	114	7,992
Partie 5	119	7,958
Partie 6	117	8,230
Partie 7	109	8,035
Partie 8	116	8,008
Partie 9	129	8,334
Partie 10	114	7,978
Partie 11	126	8,373
Partie 12	137	8,065
Partie 13	123	7,847
Partie 14	103	8,457

TABLE 2.3 – Représentation du découpage du corpus du cancer du sein en 14 parties

2.6 Méthode compositionnelle

Les principaux travaux en extraction de lexiques bilingues à partir de corpus comparables se sont focalisés sur les termes simples [11]. Or, une grande partie des termes issus des domaines de spécialité sont des termes complexes, c'est-à-dire des termes composés de plusieurs unités lexicales, aussi appelés termes polylexicaux. Si la traduction des termes simples constitue une étape fondamentale pour la traduction des termes complexes, les méthodes qui leurs sont dédiées sont pour la plupart inadaptées aux termes complexes. Ainsi, plusieurs chercheurs se sont penchés sur la question de l'extraction et de la traduction des termes complexes car elle constitue à elle seule un domaine de recherche bien défini.

La méthode la plus simple pour traduire un terme complexe suppose qu'une bonne traduction de celui-ci peut être obtenue par la traduction de chacun des éléments qui le compose. Une fois la traduction des éléments effectuée, toutes les combinaisons possibles des traductions sont générées et filtrées soit à l'aide d'une liste de termes complexes extraite du corpus cible [Morin et Daille, 2010], soit à l'aide du corpus cible [Robitaille *et al.*, 2006], soit à l'aide du web [Grefenstette, 1999]. Cette méthode n'est efficace que si les termes complexes en relation de traduction partagent la propriété compositionnelle : *la traduction du tout est fonction de la traduction de chacune des parties*. Il a été rapporté dans [Baldwin et Tanaka, 2004] que 48,7% des termes complexes anglais/japonais partagent cette propriété. Si les termes complexes sont moins polysémiques [Savary et Jacquemin, 2000] et plus

11. Un seul mot plein.

représentatifs [Nakagawa et Mori, 2003] des domaines de spécialité que les termes simples, trouver leurs traductions pose plusieurs problèmes.

D'après [Morin et Daille, 2010] il existe principalement quatre types de problèmes :

- Les variations lexicales où des termes sources et cibles contiennent des termes sémantiquement liés mais non équivalents. Exemple : *machine translation* –> *traduction automatique*
- Les variations morphosyntaxiques où les structures morphosyntaxiques des termes source et cible sont différentes. Exemple : anti-cancer –> anti-cancéreux
- La fertilité connue comme un problème de différence de longueur en nombre d'éléments qui composent les termes complexes source et cible. Comme par exemple le terme simple allemand *axilladissektion* (longueur = 1) traduit en anglais par le terme complexe *axillary dissection* (longueur = 2), ou encore le terme français *dépistage du cancer du sein* (longueur = 3) traduit en anglais par *breast screening* (longueur = 2)
- Les variations terminologiques où un terme source peut être traduit de différentes manières en langue cible comme par exemple : *oophorectomy* qui peut être traduit par *ovariectomie* ou *ablation des ovaires*.

Des solutions aux problèmes morphosyntaxiques, lexicaux et quelques variations terminologiques ont été proposées, notamment sous forme de thésaurus comme source de référence par [Robitaille *et al.*, 2006], de règles de dérivation morphologique [Morin et Daille, 2010], de dictionnaires de variation morphologique [Cartoni, 2009] ou de modèles morphosyntaxiques [Baldwin et Tanaka, 2004, Weller *et al.*, 2011]. Le problème de fertilité a été adressé par [Weller *et al.*, 2011] pour le cas de l'allemand, [Delpech *et al.*, 2012] pour l'anglais et le français, etc.

Selon l'état de l'art, les méthodes compositionnelles peuvent être divisées en deux familles : (i) les méthodes compositionnelles basées sur les unités lexicales, et (ii) les méthodes compositionnelles basées sur les unités morphologiques.

2.6.1 Méthode compositionnelle basée sur des unités lexicales

La mise en œuvre du principe de la traduction compositionnelle d'un terme complexe (MWT) se base sur la traduction de chacune de ses parties [Grefenstette, 1999, Tanaka, 2002, Robitaille *et al.*, 2006, Morin et Daille, 2010]. Chaque élément composant le MWT est traduit à l'aide d'un dictionnaire. La forme lexicale est examinée sans considérer la catégorie grammaticale. Ensuite, toutes les combinaisons possibles des traductions sont générées sans prendre en compte l'ordre des mots. En définitive, les traductions candidates sont ordonnées en fonction de leur fréquence et filtrées à l'aide d'un système d'extraction terminologique monolingue. Le nombre de traductions candidates peut être réduit en utilisant des règles d'association de catégories grammaticales pour termes complexes (MWT POS patterns) appliquées aux corpus source et cible. Pour filtrer les traductions candidates, [Baldwin et Tanaka, 2004] ont défini les règles suivantes : un terme complexe japonais composé de deux noms N_1 N_2 peut être traduit en anglais par un terme de la forme N_1 N_2 dans 33,2% des cas, Adj_1 N_2 dans 28,4% des cas et N_2 of (*the*) N_1 dans 4,4% des cas.

2.6.2 Méthode compositionnelle basée sur des unités morphologiques

La méthode compositionnelle à base d'unités lexicales est restrictive et comporte certains inconvénients, notamment lorsqu'une ou plusieurs unités d'un terme complexe ne peuvent être traduites. Pour tenter de résoudre ce problème, [Robitaille *et al.*, 2006] ont proposé une méthode de repli qui décompose un MWT de longueur n en combinaisons de MWT de longueurs inférieures ou égales à n. Cette méthode rend la traduction des sous-parties d'un MWT possible si celles-ci sont présentes dans le dictionnaire. La méthode compositionnelle à base d'unités morphologiques est une généralisation de la méthode de repli proposée par [Robitaille *et al.*, 2006] à l'échelle de la racine (stem). Si la traduction d'un terme n'est pas trouvée, l'idée est d'essayer de relier ce terme à un mot du dictionnaire en se basant sur des informations morphologiques, par exemple la dérivation qui est un processus productif pour les langues romanes où plusieurs types d'informations peuvent être marqués par des morphèmes qui forment de nouveaux termes.

Partant de l'observation que dans les langues de spécialité les dérivations nominales sont très productives, [Morin et Daille, 2010] ont proposé une méthode dont l'hypothèse est que la dérivation morphologique constitue un processus compositionnel qui doit faire partie du processus de traduction. Ils soulignent deux cas à considérer à savoir : (i) si le processus dérivationnel garde le sens compositionnel alors les deux formes sont sémantiquement liées, (ii) quand un terme faisant partie d'un MWT est une dérivation nominale ou adjectivale, le terme dérivé est transformé dans une forme neutre en utilisant des règles.

Partant du constat qu'une traduction correcte était trop souvent noyée dans une liste de traductions candidates, et que les traducteurs préféraient utiliser un lexique plus petit mais plus précis, [Delpech *et al.*, 2012] se sont orientés vers l'approche compositionnelle dédiée à la traduction assistée par ordinateur (TAO), plus adaptée aux besoins des traducteurs. La méthode proposée nommée : méthode *morpho-compositionnelle*, se base sur la traduction d'unités monolexicales où le terme source est découpé en morphèmes, traduits puis recomposés en un terme cible, ensuite, les candidats à la traduction sont restreint en sélectionnant uniquement les traductions attestées en langue cible. Ces dernières sont ordonnées selon des critères de plausibilité. Trois axes de recherche ont été explorés, à savoir la génération de traductions fertiles (cas où le terme cible contient plus de mots lexicaux que le terme source), l'indépendance aux structures morphologiques et l'ordonnancement des traductions candidates. [Delpech *et al.*, 2012] obtiennent une précision moyenne de 91% au top1 sur deux corpus spécialisés (anglais/français et anglais/allemand).

Des travaux connexes spécifiques à certain types de termes complexes ayant des caractéristiques particulières ont été menés, comme par exemple les composés savants et les termes composés dont l'un des constituants est un adjectif relationnel [Harastani *et al.*, 2012, Harastani *et al.*, 2013]. Partant de l'hypothèse qu'un composé savant dans une langue source peut être traduit compositionnellement par un composé savant dans une langue cible, [Harastani *et al.*, 2012] se sont intéressés à traduire automatiquement les composés savants à partir des corpus comparables. Ils

considèrent qu'un terme est composé savant s'il contient au moins une racine gréco-latine. Par exemple, *aérogénérateur* est un composé savant parce qu'il comprend la racine gréco-latine *aéro*. Considérant que la formation néoclassique des mots dans différentes langues suit le modèle des langues grecque et latine pour former des termes, [Harastani *et al.*, 2012] définissent le modèle gréco-latin pour un terme XY qui se compose de deux éléments X et Y comme suit : [déterminé déterminant]. Selon ce modèle, *cardiologie* se compose de *logie* (étude), le déterminé (identifie la classe dont le composé savant est une sorte) et *cardio* (coeur), le déterminant (donne le trait distinctif). Des résultats expérimentaux sur quatre couples de langues ont montré une précision de 96%.

Concernant les termes complexes constitués d'adjectifs relationnels, [Harastani *et al.*, 2013] alignent automatiquement les adjectifs avec les noms dont ils sont dérivés en utilisant un corpus monolingue. Les alignements adjectif-nom seront ensuite utilisés dans la traduction compositionelle des termes complexes de la forme [N AdjR] à partir d'un corpus comparable. Un nouveau terme [N N′] (ex. cancer du poumon) sera obtenu en remplaçant l'adjectif relationnel $AdjR$ (ex. pulmonaire) dans [N AdjR] (ex. cancer pulmonaire) par le nom N' (ex. poumon) avec lequel il est aligné. Si aucune traduction n'est proposée pour [N AdjR], [Harastani *et al.*, 2013] considèrent que ses traduction(s) sont équivalentes à celle(s) de sa paraphrase [N N′]. Les résultats expérimentaux sur un corpus comparable dans le domaine de cancer du sein montrent une précision de 86 %.

2.7 Comparabilité des corpus comparables

L'utilisation des corpus dits *comparables*, introduit de manière intrinsèque la notion de mesure de comparabilité. Cette comparabilité peut être appliquée au niveau des corpus, des documents ou des sous-parties des documents. Si la plupart des définitions de la comparabilité stipulent une similitude des textes selon des critères qualitatifs (auteur, période, thème, genre, type de discours...) et/ou quantitatifs (mesures de fréquences,...) [Bowker et Pearson, 2002, Déjean et Gaussier, 2002], la notion de comparabilité reste néanmoins subjective et sujette à interprétation. De plus, il n'y a pas d'accord établi concernant le degré de similarité ou le seuil minimum que devraient respecter des corpus ou des documents comparables. Plusieurs travaux analysent la comparabilité en évaluant la composition des corpus [Maia, 2003, McEnery et Xiao, 2007a, Sharoff, 2007], en se basant sur des critères structurels (format, taille...) ou linguistiques (thème, domaine, genre...). Dans un contexte monolingue, des auteurs comme [Kilgarriff, 2001] et [Rose *et al.*, 1997] introduisent les notions de similarité et d'homogénéité entre corpus. Motivé par une démarche de comparaison plus objective, *kilgarriff* par exemple, essaye de déterminer les mots qui sont les plus caractéristiques du corpus. Il considère différentes mesures de similarité et établit que la mesure du χ^2 était la plus performante. Dans un contexte bilingue, certains auteurs préfèrent utiliser la notion de degré de comparabilité [Goeuriot, 2009] permettant de quantifier les ressemblances entre les textes d'un corpus bilingue [Saralegi *et al.*, 2008, Li et Gaussier, 2010, Su et Babych, 2012], etc. [Saralegi *et al.*, 2008] par exemple, mesurent la comparabilité d'un corpus (anglais/basque) en se basant sur la distribution des thèmes et des dates de publication des documents. [Munteanu et Marcu, 2005, Munteanu et Marcu, 2006] quant

à eux, collectent des corpus comparables à la manière des systèmes de recherche d'informations en utilisant le toolkit Lemur [12]. Les paires de documents extraits servent ensuite à la tâche d'extraction de phrases parallèles et semi-parallèles. [Smith *et al.*, 2010] considèrent Wikipedia comme un corpus comparable et utilisent les liens "interwiki" pour aligner des paires de documents comparables pour la tâche d'extraction de phrases parallèles. [Li et Gaussier, 2010] introduisent une mesure de comparabilité qui se base sur un dictionnaire bilingue, l'hypothèse étant que deux corpus sont comparables s'ils partagent une partie non négligeable du vocabulaire en commun. De plus, ils proposent d'améliorer la comparabilité des corpus comparables comme préalable à l'extraction lexicale bilingue. Nous revenons dans ce qui suit sur les principales méthodes mesurant la comparabilité des corpus comparables.

2.7.1 Similarité entre corpus

Dans un contexte monolingue, nous retrouvons plusieurs travaux qui s'intéressent aux mesures de similarité entre corpus [Kilgarriff, 2001]. [Johansson et Hofland, 1989] par exemple, qui dans le but de trouver les traits grammaticaux (genres) qui se ressemblent le plus dans le corpus LOB, ont sélectionné les 89 mots les plus communs du corpus, trouvé leur rang selon chaque genre et calculé la corrélation statistique de *Spearman* entre les mots. [Rose *et al.*, 1997] quant à eux, mesurent les performances d'un système de reconnaissance de la parole en fonction de la taille et des spécificités des corpus d'entraînement, utilisés pour construire leur modèle de langage. Partant d'un corpus de petite taille, ils augmentent la taille de ce dernier en rajoutant des textes de même type. Les textes similaires ont été évalués en utilisant deux mesures de similarité (spearman et le taux de vraisemblance). En analyse en dépendance syntaxique, [Sekine, 1997] explore des textes de différents genres. Il comptabilise le nombre d'occurrences de chaque sous-arbre de dépendance de profondeur 1. Ceci lui permet ensuite de comparer les corpus en s'appuyant sur les fréquences des sous-arbres de dépendance de chaque corpus. [Kilgarriff, 2001] propose une démarche selon laquelle il faut disposer au préalable de corpus dont les similarités sont connues (Known-Similarity Corpus (KSC)). Pour ce faire, il choisit d'utiliser au départ deux textes distincts A et B. Ensuite, il construit un premier corpus contenant 100% du texte A ; un deuxième corpus contenant 90% de A et 10% de B ; un troisième corpus contenant 80% de A et 20% de B et ainsi de suite. À la fin, il obtient un ensemble de corpus basé sur la mesure KSC. Il est ainsi possible d'évaluer différentes mesures de similarité. Dans ses expériences, il montre que la mesure du χ^2 donne les meilleurs résultats.

2.7.2 Mesures de comparabilité au niveau du corpus

[Li et Gaussier, 2010] proposent d'améliorer la qualité des corpus comparables afin d'améliorer la qualité des lexiques bilingues extraits. Pour ce faire, ils introduisent une mesure de comparabilité et une stratégie basée sur un processus itératif de construction de corpus comparables. La mesure de comparabilité se base sur l'espérance de trouver la traduction de chaque mot du corpus. Étant donné un corpus

12. `http://www.lemurproject.org/`

C, avec C_a sa partie anglaise et C_f sa partie française, la mesure de comparabilité M_{af} peut être définie sur la base de l'espérance de trouver pour chaque mot w_a du vocabulaire C_a^v sa traduction dans le vocabulaire C_f^v. Soit σ la fonction indiquant si une traduction de l'ensemble des traductions T_w existe dans le vocabulaire C^v du corpus C.

$$\sigma(w, C^v) = \begin{cases} 1 & \text{ssi } \mathrm{Tw} \cap \mathrm{C^v} \neq 0 \\ 0 & \text{sinon} \end{cases} \tag{2.12}$$

La mesure de comparabilité est définie comme suit :

$$M_{af}(C_a, C_f) = \frac{1}{C_a^v \cap D_a^v} \sum_{w \in C_a^v \cap D_a^v} \sigma(w, C_f^v) \tag{2.13}$$

L'équation 2.13 représente la proportion des mots anglais possédants une traduction. De la même manière, dans la direction inverse, l'équation M_{fa} est représentée comme suit :

$$M_{fa}(C_f, C_a) = \frac{1}{C_f^v \cap D_f^v} \sum_{w \in C_f^v \cap D_f^v} \sigma(w, C_a^v) \tag{2.14}$$

Une version symétrique de ces mesures considère la proportion de mots anglais et français pour lesquels une traduction existe. Cette mesure symétrique est représentée comme suit :

$$M(C_a, C_f) = \frac{\sum_{w \in C_a^v \cap D_a^v} \sigma(w, C_f^v) + \sum_{w \in C_f^v \cap D_f^v} \sigma(w, C_a^v)}{|C_a^v \cap D_a^v| + |C_f^v \cap D_f^v|} \tag{2.15}$$

Pour tester et valider la mesure de comparabilité, les auteurs développent des scores de comparabilité de référence à partir des corpus parallèles *Europarl* et *AP*. Ils dégradent graduellement la comparabilité en rajoutant des documents, soit d'*Europarl* ou du corpus *AP*. Les résultats des expériences menées sur trois groupes de corpus comparables montrent de très bonnes performances de la mesure de comparabilité symétrique M. Par la suite, les auteurs utilisent cette mesure pour améliorer la comparabilité des corpus. L'idée consiste à extraire à partir d'un corpus comparable de base C, sa sous-partie la plus comparable, notée C_H et ensuite, à enrichir le reste du corpus C, sa partie donc la moins comparable, notée C_L, par des textes puisés dans d'autres ressources. Là encore, une validation rigoureuse avec différentes ressources externes est employée. Les auteurs utilisent deux corpus de base : *GH95* et *SDA95* et deux ressources externes pour l'enrichissement des corpus qui sont : (i) *LAT94*, *MON94* et *SDA94* ; (ii) *Wiki-En* et *Wiki-Fr*. Les résultats obtenus montrent que les corpus enrichis sont plus comparables que les corpus de départ et donnent de meilleures performances sur la tâche d'extraction de lexiques bilingues.

[Liu et Zhang, 2013] s'intéressent aux domaines de spécialité et proposent une mesure de comparabilité nommée *Termhood*. Cette mesure se base non pas sur la fréquence mais sur la qualité de la terminologie. Cette qualité est mesurée en utilisant la spécificité (*Termhood*). Chaque corpus est représenté par un vecteur contenant

les termes spécifiques du domaine accompagnés de leur score de spécificité. Une fois la spécificité calculée à l'aide de l'équation 2.16, la mesure du cosinus est appliquée pour mesurer la comparabilité des corpus.

$$Termhood(w) = \frac{r_D(w)}{|V_D|} - \frac{r_B(w)}{|V_B|} \tag{2.16}$$

avec D qui représente un corpus du domaine spécialisé et son vocabulaire V_D, et B un corpus de langue générale et son vocabulaire V_B. $r(w)$ représente le rang du mot w. Les expériences menées sur un corpus anglais/chinois traitant du domaine des sciences de l'information et de la libraire, ont montré de meilleurs résultats et une meilleure stabilité que les mesures de comparabilité basées sur la fréquence.

2.7.3 Mesures de comparabilité au niveau du document

[Su et Babych, 2012] définissent la comparabilité des corpus comme étant liée à leur capacité à améliorer la performance des systèmes de traduction automatique. Ainsi, meilleure est la qualité des traductions parallèles ou quasi-parallèles extraites (mots, phrases, paragraphes, etc.) meilleure sera la comparabilité des corpus utilisés. Ils proposent d'évaluer trois mesures de similarité permettant d'extraire des documents comparables. Une première mesure basée sur l'alignement lexical consiste à mesurer la similarité entre deux documents bilingues en utilisant la mesure du cosinus. Les documents seront préalablement représentés sous forme de vecteurs contenants les mots des documents appartenant au dictionnaire bilingue. Dans le cas où il est difficile d'obtenir un dictionnaire bilingue, les auteurs s'appuient sur l'outil d'alignement statistique GIZA++ [Och et Ney, 2000] pour construire automatiquement cette ressource. Une deuxième mesure basée sur les mots-clés consiste à utiliser ces derniers pour représenter les documents à comparer. L'hypothèse étant de dire que plus il y a de mots-clés en commun entre deux documents, plus ils sont comparables. Les mots-clés sont extraits en utilisant le TF-IDF et la similarité entre deux documents se calcule aussi en utilisant le cosinus. La troisième mesure est basée sur la traduction automatique. L'idée étant d'utiliser un système de traduction automatique (Microsoft API) pour traduire en anglais les textes des langues sous représentées, comme par exemple le lituanien ou le slovène. Puis, d'explorer plusieurs traits. La mesure de comparabilité est en fait une combinaison linéaire pondérée de quatre mesures, chacune basée sur un trait particulier. Ces traits peuvent être des entités nommées, des mots-clés, des traits lexicaux (représentation des documents par sac de mots) ou des structures particulières (le nombre d'adjectifs, d'adverbes, de noms, de verbes et de noms propres ainsi que le nombre de phrases dans chaque document). [Su et Babych, 2012] montrent que les trois mesures proposées sont fiables et permettent de dire si des documents sont peu comparables, très comparables ou parallèles.

2.8 Bilan

Nous avons présenté dans cette partie des travaux répartis sur deux chapitres. Le premier introduit les notions de corpus multilingues parallèles et comparables avec leurs avantages et inconvénients. Le second chapitre présente les principales méthodes dédiées à l'extraction de lexiques bilingues à partir de corpus comparables ainsi que la notion de comparabilité des corpus comparable. Cet état de l'art non exhaustif couvre néanmoins les principales idées et hypothèses qui permettent de mettre en relation des couples de traductions. Nous retiendrons ainsi le cas des termes simples traités par les méthodes distributionnelles et le cas des termes complexes traités par les méthodes compositionnelles. Nous retiendrons aussi le fait qu'il est difficile de comparer les différents travaux sus-décrits, principalement à cause de la différence des ressources utilisées en termes de type ou de taille. Néanmoins, se confronter à l'*approche directe* qui fait office de référence permet de donner une idée sur les améliorations apportées par chaque nouvelle approche proposée.

II

Unités lexicales et contexte

3

Étude du contexte

Introduction

La plupart des méthodes distributionnelles utilisées pour l'alignement bilingue à partir de corpus comparables se basent sur trois actants principaux qui sont : (1) le corpus bilingue, (2) le dictionnaire bilingue et (3) la caractérisation du contexte des mots issus du corpus comparable. Nous développerons dans ce chapitre le troisième point qui constitue la principale partie sur laquelle nous nous concentrerons dans notre travail.

Le choix du contexte qui caractérise le mot à traduire et les candidats à la traduction joue sans nul doute un rôle déterminant dans la performance des méthodes d'alignement basées sur les corpus comparables. Il arrive souvent en linguistique computationnelle de classer les mots en fonction de leurs cooccurrences avec d'autres mots (selon la tradition Firthienne [1]). Cette manière de faire, bien qu'efficace surtout quand il s'agit de mots fréquents, reste néanmoins approximative et peut dans certains cas considérer des mots qui ne devraient pas l'être. Ces mots, à défaut d'apporter une information utile à la traduction, pourraient au contraire ajouter une information négative (bruit) qui nuirait à un bon alignement. L'une des questions qui reste à résoudre est : comment améliorer la représentation du contexte pour qu'il soit le plus discriminant possible ? Ceci dit, est-ce qu'une bonne représentation du contexte suffit à assurer une extraction terminologique bilingue de qualité ?

Dans ce chapitre, nous nous efforcerons d'analyser le contexte sous ses différentes formes et essayerons de déterminer ou du moins d'apporter des éléments de réponse en ce qui concerne le choix du contexte. Celui-ci n'étant pas trivial, nous verrons si les intuitions qui guident tel ou tel choix génèrent des règles applicables dans notre problématique d'extraction terminologique bilingue.

1. You shall know a word by the company it keeps.

3.1 Retour sur les notions de mots et de termes

Différencier les mots des termes n'est pas toujours une tâche aisée : «*...the differenciation of terms from words is not straightforward, since the relationship between general language and sublanguages...is an interdependent one. (Pointer final report 1996, Section 4, p.17)*».

Selon les terminologues, les mots deviennent des termes, c'est-à-dire acquièrent un statut particulier, lorsqu'ils sont utilisés dans des domaines de spécialité. Les chercheurs en traitement du langage naturel (TALN) quant à eux, soutiennent que le lexique et la grammaire utilisés dans certains domaines ou dans un certain type de textes sont restreints. La manière dont un terme est défini varie sans doute d'une communauté à une autre, mais la conviction que le terme est différent d'un mot reste la même pour tout le monde. Nous retrouvons dans la littérature deux types de définitions d'un terme, à savoir la définition dite ***traditionnelle*** et celle dite ***pragmatique*** [Pearson, 1998].

3.1.1 Définition traditionnelle d'un terme

Selon [Rondeau, 1984], le terme est un signe linguistique au sens Saussurien[2], c'est-à-dire qu'il a un signifiant et un signifié. Alors que [Wuster, 1968] utilise le mot ***terme*** pour désigner une étiquette qui elle-même désigne un concept[3], [Rondeau, 1984] quant à lui, donne le nom de ***dénomination*** à l'étiquette et le nom de ***notion*** au concept et utilise le mot ***terme*** pour décrire la combinaison des deux. D'après Rondeau et Wuster, les terminologues doivent d'abord définir et décrire le concept pour ensuite décider de l'étiquette la plus appropriée à lui octroyer.

[Sager, 1990] présente trois assertions pour distinguer les ***termes*** des ***mots***. Il stipule que le lexique d'un langage spécifique tend à fournir autant d'unités lexicales que de concepts établis de manière conventionnelle. Il ajoute que ce même lexique comporte deux classes d'unités lexicales, à savoir les unités lexicales avec des références spécifiques qui sont les ***termes*** et les unités lexicales avec des références générales qui sont les ***mots***. Ces derniers ne sont spécifiques à aucune discipline et leurs références sont uniformes, vagues ou générales.

[Felber, 1984] définit aussi trois types de symboles linguistiques qui sont le ***mot***, le ***terme*** et le ***mot thésaurus***. Il décrit le mot comme ayant une multiplicité de sens non définie et pouvant être utilisé pour nommer des objets. Il ajoute que le sens concret d'un mot est donné par son contexte. Concernant le terme, Felber le définit comme étant un symbole linguistique assigné à un ou plusieurs concepts. Le sens du terme qui est le concept est dépendant de la position de ce concept par rapport aux autres concepts concernés. Il définit le mot thésaurus comme pouvant être un terme dans la plupart des cas, ou un nom utilisé pour indexer et extraire de l'information dans des systèmes d'information.

2. Ferdinand de Saussure (1857-1913) : Linguiste Suisse. Fondateur de la linguistique moderne et des bases de la sémiologie (science des signes).

3. Concept : Construction mentale représentant les objets et phénomènes perçus dans le monde réel.

Pour finir, si l'on s'appuie sur le standard international (ISO 1087), Selon [ISO 1087-1:2000, 2000] un terme [4] est une désignation verbale d'un concept général dans un domaine spécifique. Selon [ISO 1087-1990:5, 1990] un terme [5] est une désignation d'un concept défini dans un langage spécifique par une expression linguistique. Tandis qu'un mot [6] est la plus petite unité lexicale ayant un sens spécifique et pouvant exister comme une unité séparée dans une phrase.

En résumé, pour les terminologues traditionnels, la notion de terme s'apparente aux unités lexicales faisant une référence spécifique à un domaine bien défini et bien restreint (Sager) ; le terme peut être une étiquette ou un symbole linguistique désignant un concept (ISO,Felber,Wuster) ; c'est équivalent au signe linguistique de De Saussure (Rondeau) ; une distinction est faite entre les termes techniques qui sont utilisés dans des domaines spécifiques, et les mots de langue générale qui sont utilisés dans plusieurs domaines (Sager).

3.1.2 Définition pragmatique d'un terme

À la différence des approches traditionnelles concernant la terminologie définie précédemment, nous allons dans ce paragraphe examiner des approches plus pragmatiques dans la définition du terme.

[Hoffmann, 1985] suggère trois différentes manières de définir un terme. La première est une vision restreinte et limitée qui consiste à dire que seule la terminologie propre à un sujet ou domaine spécifique doit être considérée comme terme, et que toute autre unité lexicale devra appartenir au vocabulaire de langue générale. La deuxième consiste à dire que toutes les unités lexicales utilisées dans un domaine spécifique peuvent être considérées comme des termes. Et la troisième approche (qui est la plus commune des trois) considère qu'à l'intérieur d'un vocabulaire spécialisé il existe trois catégories de termes : (i) des termes monosémiques qui sont seulement utilisés dans un seul domaine [7], (ii) des termes spécialisés qui peuvent être utilisés dans plusieurs domaines [8] et (iii) les mots de langue générale qui n'ont pas de références spécifiques à un quelconque domaine spécialisé [9]. Ainsi Hoffmann distingue bien les termes spécifiques faisant référence à un domaine particulier de ceux qui peuvent référencer plusieurs domaines de spécialité et des mots ordinaires qui ne sont donc pas des termes. Il reconnaît néanmoins qu'il est difficile en pratique de décider d'une manière systématique de la classe à laquelle appartient une unité lexicale.

[Trimble et Trimble, 1978] quant à eux, ne définissent pas le terme en lui-même mais distinguent bien la différence entre trois catégories de termes : (i) les termes très techniques (highly technical terms) donc spécifiques à un domaine particulier, (ii) une banque de termes techniques pouvant être utilisés dans plusieurs domaines

4. term : Verbal designation of a general concept in a specific subject field.
5. term : Designation of a defined concept in a special language by a linguistique expression.
6. word : Smallest linguistic unit conveying a specific meaning and capable of existing as a separate unit in a sentence.
7. Fachwortschatz : Vocabulaire scientifique spécialisé.
8. Allgemeinwissenschaftlicher Wortschatz : Vocabulaire scientifique général.
9. Allgemeiner Wortschatz : Vocabulaire général.

([Trimble et Trimble, 1978] rejoignent donc [Hoffmann, 1985] pour les points (i) et (ii)) et (iii) les termes techniques à un degré moindre. Par exemple : *control, operation, current, ground, sense, positive, contact, lead, folder, flux*, etc. Ces termes sont en fait des mots de la langue générale qui ont pris un sens particulier dans un domaine de spécialité.

[Herbert, 1965] divise les termes en deux catégories : (i) les termes très techniques (highly technical terms), ceci rejoint [Trimble et Trimble, 1978] et (ii) les termes semi-techniques ou semi-scientifiques qui sont des mots ayant un large spectre de sens et qui sont utilisés d'une manière idiomatique, par exemple : *work, plant, load, feed, force*, etc. Le point (ii) proposé par [Herbert, 1965] suggère qu'il y a des mots de langue générale qui, lorsqu'ils sont utilisés dans un domaine de spécialité, prennent un sens différent de celui qu'ils ont habituellement dans la langue générale. [Goldman et Payne, 1981] font eux aussi la distinction entre les termes techniques et les termes non techniques. Ils considèrent les termes techniques comme étant ceux où il y a eu congruence des concepts entre les scientifiques et ceci quelle que soit la langue utilisée. Ils divisent les termes non techniques en deux catégories : (i) les termes de la langue générale, par exemple les termes logiques comme les coordinateurs, les subordonnées, les déterminants, les quantificateurs, les compléments, etc., et (ii) les termes basiques qui peuvent être utilisés en science, par exemple : *study, assumption, inference, evidence, similarity, distance*. [Yang, 1986] défend la distinction entre les termes spécifiques à un domaine et les termes de degré moindre comme par exemple : *absolute, accuracy, electrical, fact, factor, result, feature*, etc., qui sont des termes identifiés dans ses corpus scientifiques anglais.

En observant les différentes définitions pragmatiques du terme, nous pouvons constater que la notion de terme semble être plus ou moins claire, et que la plupart des définitions se focalisent sur la distinction entre les différentes catégories de termes. La question qui n'est jamais abordée dans ces définitions est : comment reconnaître un terme d'un mot ? Notre problématique n'étant pas de reconnaître automatiquement les termes, nous renvoyons le lecteur vers [L'Homme, 2004] qui dresse un très bon état de l'art sur les méthodes d'extraction automatique de termes.

3.1.3 Synthèse

Il est bien sûr intéressant de creuser la définition de ce qu'est un terme, et de connaître les différentes visions des chercheurs mais en pratique, et surtout dans notre problématique de construction de contexte à des fins d'alignement, nous retiendrons deux définitions de [L'Homme, 2004] qui nous semblent être les plus claires :

1. *... Les termes sont des unités lexicales dont le sens est envisagé par rapport à un domaine de spécialité, c'est-à-dire un domaine de connaissance humaine, souvent associé à une activité socio-professionnelle...* ;
2. *... La particularité du terme, par rapport aux autres unités lexicales d'une langue, est d'avoir un sens spécialisé, c'est à dire un sens qui peut être mis en rapport avec un domaine de spécialité....*

Pour éviter la confusion entre les notions de mot et de terme, nous choisissons d'utiliser le terme *unité lexicale* qui englobe ces deux notions. Nous retiendrons

des définitions traditionnelles et pragmatiques citées plus haut, trois éléments qui paraissent être les plus importants concernant notre problématique de l'extraction terminologique bilingue, à savoir :

1. Différencier les mots des termes ;
2. L'existence de termes très techniques qui sont propres à un seul domaine ;
3. L'existence de termes pluridisciplinaires.

De ces observations découlent naturellement des questions liées d'une manière directe ou indirecte à la problématique de l'extraction terminologique bilingue. Est-ce que les mots et les termes jouent le même rôle dans la représentation du contexte d'une unité lexicale à traduire ? Doit-on les considérer de la même manière ou plutôt les différencier ? Nous gardons en tête ces questions pour y revenir plus tard. Pour le moment, nous allons aborder dans la section suivante le contexte proprement dit.

3.2 Contexte d'une unité lexicale

Intuitivement, la plus simple des définitions que l'on puisse donner au terme : *contexte*, et que nous retrouvons dans la plupart des dictionnaires et ouvrages, est que le contexte d'une unité lexicale peut être vu comme l'ensemble des éléments d'un texte qui l'entoure et qui apporte un éclairage sur le sens de celle-ci. Selon cette définition, la construction du contexte d'une unité lexicale i reviendrait à rechercher des relations ou associations de i avec d'autres unités lexicales qui apparaissent dans son environnement. Ceci étant dit, cette définition reste une définition de surface et donc incomplète.

D'après [de Saussure, 1916], il existe deux types de relations fondamentales entre les unités lexicales, qui correspondent aux opérations basiques du cerveau humain et qui sont les associations dites *syntagmatiques* et les associations dites *paradigmatiques*. Grefenstette quant à lui [Grefenstette, 1994a] présente trois niveaux de relations entre les unités lexicales qui sont les relations du premier ordre (relations *syntagmatiques*), les relations du second ordre (relations *paradigmatiques*), et les relations du troisième ordre. La question qui s'impose et se pose de manière intrinsèque est : comment regrouper les unités lexicales entre elles pour représenter le contexte d'une unité lexicale donnée ? Des éléments de réponse semblent apparaître dans les paragraphes suivants.

3.2.1 Associations syntagmatiques ou affinités du premier ordre

Selon [Rapp, 2002], il existe une relation syntagmatique entre deux unités lexicales si elles cooccurrent plus fréquemment que par chance et si elles ont un rôle grammatical dans la phrase où elles apparaissent. À titre d'exemple, nous pouvons prendre des mots comme *(professeur et école)*, *(boire et café)*, *(chaleur et soleil)*, etc. Une méthode simple pour extraire ce type de relation d'un corpus consiste à sélectionner les couples de mots dont les cooccurrences sont significativement grandes. Pour ce faire, le test du $\chi 2$ peut être utilisé, sauf si le degré de dispersion des

données est grand, dans ce cas, il est préférable d'utiliser la mesure du rapport de vraisemblance [Dunning, 1993].

Extraire les affinités du premier ordre[10] revient à faire une représentation locale du contexte d'une unité lexicale [Grefenstette, 1994a]. Nous pouvons extraire plusieurs types d'affinités du premier ordre [Grefenstette, 1994a]. Par exemple, [Church et Hanks, 1990] ont montré que l'utilisation de la mesure de l'information mutuelle [Fano, 1961] entre les unités lexicales, en faisant passer une fenêtre contextuelle de taille fixe sur un grand corpus, permettait de reconnaître des affinités entre paires de mots comme *(doctor et nurse)*, *(save et from)*, etc. Pour l'extraction de collocations [Smajda, 1993] a proposé une technique pour la reconnaissance d'expressions, de patrons fixes et de collocations. [Manning, 1993] quant à lui, suivant les travaux de [Brent, 1991], dérive des sous-catégorisations d'un texte à l'aide d'un étiquetage (tagging) stochastique et d'une analyse (parsing) robuste, en faisant une évaluation statistique des phrases apparaissant autour d'un verbe donné. [Grefenstette, 1993] extrait des familles de mots propres à un domaine particulier (variantes morphologiques) en s'appuyant sur leur contexte défini dans les documents où ils apparaissent, mais aussi à l'aide de techniques de comparaison de chaînes de caractères.

3.2.2 Associations paradigmatiques ou affinités du second ordre

Il existe une relation paradigmatique entre deux unités lexicales i et j si elles ont une forte similarité sémantique [Rapp, 2002], c'est-à-dire si en substituant i à j ceci n'affecte pas le sens grammatical de la phrase. Cette similarité sémantique peut être déterminée par l'extraction des plus proches voisins d'une unité lexicale. Des exemples typiques de ce type d'association sont les synonymes et les antonymes, comme par exemple : *rapide* - *vite*, *boire* - *manger*, *rouge* - *bleu*, etc. La relation paradigmatique entre les unités lexicales *rouge et bleu* peut être déterminée en observant que ces deux unités lexicales apparaissent avec des unités lexicales comme *fleurs, robe, couleur*, etc.

Selon [Grefenstette, 1994a] les affinités du second ordre montrent des unités lexicales qui partagent les mêmes environnements et qui n'ont pas besoin d'apparaître ensemble. Les techniques utilisant ces relations ont montré des résultats intéressants. [Brown *et al.*, 1992] par exemple ont montré que les techniques de comparaison de chaînes (string matching technique) utilisant une fenêtre autour de chaque unité lexicale à comparer fournissent assez d'informations pour extraire des unités lexicales en relation sémantique.

[Deerwester *et al.*, 1990] quant à eux représentent les relations entre les unités lexicales et les documents au travers d'une matrice dans laquelle les lignes correspondent aux unités lexicales et les colonnes aux documents dans lesquelles elles apparaissent. Ils utilisent la LSA (Latent Semantic Analysis) pour représenter dans un nouvel espace sémantique les relations entre les unités lexicales. Une première étape consiste à réduire la matrice de représentation en utilisant la décomposition en valeurs singulières. Une fois le nouvel espace vectoriel construit, chaque unité lexicale

10. First-order term affinities describe what other words are likely to be found in the immediate vicinity of a given word [Grefenstette, 1994a].

est projetée dans ce nouvel espace, et la relation sémantique entre chaque couple est calculée à l'aide d'une mesure de distance. [Deerwester *et al.*, 1990] ont montré que cette technique améliorait l'extraction d'information mais qu'elle souffrait de la complexité relative à la réduction matricielle. [Schütze, 1993] utilise une technique semblable nommée CCA (Canonical Correlation Analysis) pour construire un nouvel espace vectoriel sémantique utilisant les cooccurrences entre unités lexicales, à l'aide d'une fenêtre contextuelle de 1000 caractères. Cette méthode souffre aussi de la complexité de calcul.

[Hindle, 1990] quant à lui représente l'information sémantique d'une manière plus précise, en combinant les paires nom-verbe. Il extrait d'un grand corpus (6 millions de mots) des triplets Sujet-Verbe-Objet pour ensuite calculer l'information mutuelle entre les paires Verbe-Nom. Pour extraire les affinités du second ordre, il calcule la similarité entre les noms en considérant la quantité d'information mutuelle partagée avec tous les verbes du corpus. [Hindle, 1990] a réussi à obtenir des résultats intéressants. Par exemple, avec sa méthode et pour le mot *boat*, il a extrait des unités lexicales en relation sémantique comme : *ship, plane, bus, vessel, truck, car, helicopter, ferry*, etc.

3.2.3 Affinités du troisième ordre

Extraire des affinités du troisième ordre consiste à partir des affinités du second ordre et à les reclasser d'une manière plus pertinente. Les techniques d'extraction d'affinités du troisième ordre prennent en entré une liste d'unités lexicales produite par des techniques d'extraction des relations du second ordre, pour ensuite en dériver des sous-groupes d'unités lexicales fortement similaires. [Schütze et Pedersen, 1993] par exemple extraient les contextes à droite d'une unité lexicale i et déterminent ainsi une liste de relations de second ordre pour i, puis ils examinent les contextes à gauche de i avec cette même liste pour extraire des sous-groupes d'unités lexicales similaires.

3.3 Construction du contexte

3.3.1 Contexte par sac de mots

Comme son nom l'indique le contexte par ***sac de mots*** est tout simplement une collecte des mots entourant un mot donné, sans règles précises, si ce n'est fixer un nombre de mots à gauche et à droite d'un mot dont nous cherchons à construire le contexte, appelé aussi **fenêtre contextuelle**.

Exemple : Soit la phrase suivante :

> ...Pour les cas **traités** pour **danger** ***ostéoporotique*** les **densitométries osseuses** comparatives ont montré une amélioration sous THS...

Pour le terme : ***ostéoporotique***, si nous choisissons une fenêtre contextuelle de taille 5, c'est-à-dire deux mots à gauche et deux mots à droite de celui-ci, le contexte

de ***ostéoporotique*** sera : **traités**, **danger**, **densitométries** et **osseuses**.

Ce processus est répété autant de fois que le terme ***ostéoporotique*** apparaît dans un corpus donnée. Cette manière de faire, bien que discutable, a montré son efficacité surtout s'agissant de mots très fréquents. Intuitivement, nous pouvons penser que tous les mots entourant un mot donné n'ont pas la même importance, et qu'il serait parfois utile de ne pas tous les considérer de la même manière. Ceci dit, toute la difficulté réside dans la prise de décision concernant tel ou tel mot. Une méthode qui vise à pallier cette difficulté consiste en l'utilisation des relations de dépendance syntaxique entre les mots que nous définissons dans la section suivante.

3.3.2 Relations de dépendance syntaxique

Dans un souci de mieux représenter le contexte d'un mot, plusieurs travaux se sont tournés vers les relations de dépendance syntaxique, notamment [Gamallo, 2008a] et [Garera *et al.*, 2009] où l'idée consiste à décrire un mot par les relations de dépendances qu'il entretient avec les mots avoisinants.

Une relation de dépendance est une relation binaire asymétrique entre un mot appelé tête ou parent (head or parent) et un modificateur ou dépendant (modifier or dependant). Les relations de dépendance forment un arbre qui interconnecte tous les mots d'une phrase. Un mot dans une phrase peut avoir plusieurs modificateurs mais chaque mot ne peut modifier au plus qu'un seul mot [Lin, 1998b]. La racine de l'arbre de dépendance aussi appelée Head ne modifie aucun mot de la phrase.

Exemple : *I have a brown dog* (1)

Une liste de tuples est utilisée pour représenter un arbre de dépendances : (mot catégorie [tête] [relation])

- mot : est le mot représenté dans le nœud de l'arbre ;
- catégorie : constitue la catégorie lexicale du mot ;
- tête : spécifie quel mot est modifié par mot ;
- relation : est une étiquette attribuée à la relation de dépendance (subj pour subject, spec pour specifier, etc.).

Le signe « < » signifie prédécesseur et « > » signifie successeur . L'arbre de dépendance de la phrase (1) est le suivant :

mot	catégorie	tête	relation
I	N	< have	subj
have	V	–	–
a	Det	< dog	spec
brown	Adj	< dog	adjn
dog	N	> have	comp

TABLE 3.1 – Exemple d'arbre de dépendance

Pour plus de détails concernant les dépendances syntaxiques [Gamallo, 2008b] aborde ce sujet en détail dans son article « The meaning of Syntactic Dependencies », plus particulièrement pour les tâches de désambiguïsation de mots et la résolution des dépendances (attachement resolution). Gamallo aborde dans son article 3 notions élémentaires de dénotation :

- Les mots lexicaux (Mary, ran, fast, nice, etc.) ;
- Les dépendances syntaxiques (subject, direct object, prepositional relation between two nouns, prepositional relation between a verb and a noun,etc.) ;
- Modèle lexico-syntaxic qui consiste à combiner les mots et leurs catégories syntaxiques en terme de dépendance. (Noun+ subj + Verb).

D'après Gamallo, les mots lexicaux représentent des ensembles de propriétés N, V, Adj, Adv ... alors que les dépendances et les modèles lexicaux syntaxiques sont définis comme opérations sur ces ensembles. Une dépendance est une relation binaire qui prend comme entrée deux ensembles de propriétés et donne en sortie un ensemble plus restreint qui est l'intersection des ensembles d'entrées.

On retrouve sept types de relations de dépendances [Gamallo, 2008a] résumées dans la table suivante :

Relation	Type	Exemple
Lmod	modificateur gauche si relation Adj - Noun	local - recurrence
Rmod	modificateur droite si relation Noun - Adj	number - insufficient
modN	modificateur de Nom si relation Noun - Noun	breast - cancer
Lobj	objet à gauche si relation Noun - Verb	study - demonstrate
Robj	objet à droite si relation Verb - Noun	have - effect
PRP	si relation prépositionnelle Noun-PRP-Noun	malignancy - in - woman
iobj	si relation objet indirecte Verb-PRP-Noun	occur - in - portion

TABLE 3.2 – Relations de dépendance syntaxique

Pour le mot : ***recurrence*** par exemple, il existe une relation Lmod avec l'adjectif **local**. Donc, dans le processus de construction du contexte de ***recurrence***, nous comptabiliserons le nombre de fois où l'adjectif **local** apparaît à gauche de ***recurrence*** dans le corpus. Nous ferons de même pour les autres relations de dépendance syntaxique.

Synthèse

Nous venons de voir deux manières de représenter le contexte, à savoir une représentation graphique (par sac de mots) et une représentation syntaxique (par relation de dépendance). L'intérêt de passer d'une représentation graphique à une représentation syntaxique des mots peut être vu selon deux aspects. Le premier est de se dire que l'information véhiculée par une représentation graphique n'est en fin de compte qu'une information quantitative, très variable et fortement dépendante des corpus utilisés, d'où l'idée d'abandonner ce type de représentation pour passer à une représentation syntaxique qui serait donc, porteuse d'information qualitatives,

dans l'idéal indépendante de la taille des corpus. Le deuxième aspect serait de dire que malgré tout, la représentation graphique a quand même un grand intérêt même si elle comporte pas mal de points faibles, et au lieu de s'en écarter, il vaudrait peut-être mieux la fusionner avec la représentation syntaxique afin de tirer le meilleur des deux. Une étude empirique traitant ces deux aspects sera présentée dans le chapitre 5.

3.4 Mesures d'association

L'une des étapes les plus importantes dans la construction du contexte d'une unité lexicale à traduire est l'affectation d'un poids ou d'un score aux unités lexicales appartenant au contexte de cette unité lexicale. Ce score est un indicateur de l'importance d'une unité lexicale vis-à-vis d'une unité lexicale à traduire, plus connu sous le nom de score ou mesure d'association. Pour ce faire, plusieurs mesures d'association ont été introduites. Dans ce qui suit, nous présentons les principales mesures utilisées.

3.4.1 Score de cooccurrence

Il s'agit là d'un simple comptage du nombre de fois où deux unités lexicales i et j apparaissent ensemble dans un corpus, souvent noté $cooc(i, j)$. Il est probable que cette mesure d'association ne peut suffire à indiquer l'importance d'une unité lexicale par rapport à une autre.

3.4.2 Le tf x idf

Le $tf \times idf$ (term frequency x inverse document frequency) est souvent utilisé en recherche d'information pour désigner l'importance d'un terme par rapport à un document en prenant en compte sa fréquence dans un document (tf) mais aussi l'inverse du nombre de documents où il apparaît (idf).

$tf_{i,j}$ désigne la fréquence d'un terme i dans un document j représentée dans l'équation suivante :

$$tf_{i,j} = \frac{f_{i,j}}{max_{l,j}} \tag{3.1}$$

- $f_{i,j}$: le nombre total de documents dans le corpus ;
- $max_{l,j}$: la fréquence maximale des termes dans le document j.

idf_i désigne le pouvoir discriminant d'un terme i, il est représenté comme suit :

$$idf_i = \log \frac{N}{n_i} \tag{3.2}$$

avec :

- N : le nombre total de documents dans le corpus ;
- n_i : le nombre de documents où le terme i apparaît.

L'application de la mesure du $tf \times idf$ au cas des distributions de mots dans des contextes conduit à remplacer le document j par le mot j cooccurrant avec le mot i. La reformulation du $tf \times idf$ devient alors :

$$tf_{i,j} = \frac{cooc_{i,j}}{max_{cooc_j}} \tag{3.3}$$

avec :
- $cooc_{i,j}$: le nombre de cooccurrence de i et j ;
- max_{cooc_j} : la fréquence maximale de cooccurrence du mot j.

et :

$$idf_i = 1 + \log \frac{max_{cooc}}{cooc_i} \tag{3.4}$$

avec :
- $cooc_i$: le nombre total de mots cooccurrent avec i ;
- max_{cooc} : le nombre maximal de mots de contexte dans le corpus.

Finalement, le poids d'un mot de contexte i pour un mot j est :

$$p_{i,j} = tf_{i,j} \times idf_i \tag{3.5}$$

3.4.3 Information mutuelle

Introduite par Fano [Fano, 1961], l'information mutuelle, comme son nom l'indique représente la quantité d'information partagée par deux variables i et j, notée $IM(i,j)$.

$$IM(i,j) = \log_2 \frac{P(i \mid j)}{P(i)} \tag{3.6}$$

Sachant que d'après la formule de Bayes $P(i \mid j)$ s'écrit $P(i,j)/P(j)$ où $P(i,j)$ correspond à la probabilité d'observer i et j simultanément. L'information mutuelle devient alors :

$$IM(i,j) = \log_2 \frac{P(i,j)}{P(i)P(j)} \tag{3.7}$$

avec :
- $P(i)$ (respectivement $P(j)$) est estimée en fonction de sa fréquence ($f(i)$) dans le corpus (nombre d'occurrences normalisé par le nombre total de mots dans le corpus, soit N ;

– $P(i,j)$ est estimée en fonction de la fréquence de cooccurrence de i et j ($f(i,j)$) (nombre de cooccurrences normalisé par le nombre total de mots dans le corpus, soit N).

$$IM(i,j) = \log_2 N \times \frac{f(i,j)}{f(i) \times f(j)} \tag{3.8}$$

L'information mutuelle a tendance à favoriser les liens exclusifs, notamment entre les mots de faibles fréquences en surestimant leurs cooccurrences. Pour pallier cela, deux variantes de l'information mutuelle ont été proposées qui sont l'information mutuelle locale et l'information mutuelle cubique (information mutuelle de puissance k ou heuristique).

3.4.4 Information mutuelle locale

Le principe de l'information mutuelle locale (IML) est de prendre en considération la fréquence de cooccurrence du couple. Ceci se fait simplement en pondérant l'information mutuelle par l'observation $f(i,j)$. L'information mutuelle locale est représentée comme suit :

$$IML(i,j) = f(i,j) \times (\log_2 N \times \frac{f(i,j)}{f(i) \times f(j)}) \tag{3.9}$$

3.4.5 Information mutuelle heuristique

L'une des variantes de l'information mutuelle est l'IM^2. Cette mesure heuristique a pour but d'augmenter l'influence de la fréquence de cooccurrence dans le numérateur pour amortir l'effet de surestimation des couples de faible fréquence. L'IM^2 est représentée comme suit :

$$IM^2(i,j) = \log_2 N \times \frac{f(i,j)^2}{f(i) \times f(j)} \tag{3.10}$$

Une autre variante de l'information mutuelle est l'IM^3 qui utilise un exposant de rang 3 pour booster encore plus le score d'association des couples de haute fréquence. [Daille, 1994] a testé plusieurs autres variantes de l'information mutuelle heuristique notée IM^k avec $k = 2, ..., 10$. Elle a trouvé que les meilleurs résultats étaient obtenus avec un $k = 3$.

$$IM^3(i,j) = \log_2 N \times \frac{f(i,j)^3}{f(i) \times f(j)} \tag{3.11}$$

3.4.6 La mesure du χ^2

La statistique du χ^2 mesure le degré de dépendance entre deux mots. Elle est calculée en fonction de la table de contingence représentée ci dessous :

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

TABLE 3.3 – Table de contingence pour la dépendance de deux unités i et j

avec :

- a : le nombre de contextes dans lesquels i et j apparaissent ensemble ;
- b : le nombre de contextes *où* j est présent mais pas i ;
- c : le nombre de contextes *où* i est présent mais pas j ;
- d : le nombre de contextes *où* i et j sont absents.

La mesure d'association χ^2 est alors définie comme suit :

$$\chi^2 = \frac{N(ad - cb)^2}{j_1 i_1 i_0 j_0} \tag{3.12}$$

Il est à noter que deux mots indépendants ont un χ^2 nul. Dans le cas où le degré de dispersion des données est grand, il est préférable d'utiliser la mesure du rapport de vraisemblance présentée ci-dessous.

3.4.7 Le rapport (taux) de vraisemblance

Introduit par Duning [Dunning, 1993] comme mesure d'association, le rapport de vraisemblance entre deux termes i et j noté par $ll(i, j)$ (ll : log likelihood) est défini comme suit :

$$ll(i, j) = 2 \sum O_{ij} \log \frac{O_{ij}}{E_{ij}} \tag{3.13}$$

où O_{ij} est le nombre d'occurrences observé entre i et j et E_{ij} est le nombre d'occurrences attendu entre i et j.

3.4.8 La mesure du odds-ratio

En s'appuyant sur la table de contingence présentée plus haut, la mesure du odds-ratio telle que définie par [Evert, 2005] est :

$$odds - ratio = \log \frac{a \times d}{b \times c} \tag{3.14}$$

Le problème de cette mesure est qu'elle suppose une valeur infinie dans le cas où les valeurs observées sont égales à zéro ($-\infty$ pour $a = 0$ ou $d = 0$, $+\infty$ pour $b = 0$ ou $c = 0$). Pour pallier cela, plusieurs applications utilisent une variante appelée *discounted odds-ratio* représentée dans l'équation suivante :

$$odds - ratio_{disc} = \log \frac{(a + \frac{1}{2}) \times (d + \frac{1}{2})}{(b + \frac{1}{2}) \times (c + \frac{1}{2})} \tag{3.15}$$

Cette dernière a montré de bonnes performances dans plusieurs études et applications [Evert, 2005, Laroche et Langlais, 2010].

3.5 Synthèse

Les mesures d'association présentées ci-dessus ont pour but d'indiquer la force de la relation entre deux unités lexicales. Cette information qui est observée à partir d'un ensemble de documents peut parfois être non fiable d'un point de vue statistique. Le choix de la mesure d'association la plus appropriée dépend fortement du contexte dans lequel elle est utilisée. [Evert, 2005] dresse dans ses travaux de thèse une étude comparative approfondie de plusieurs mesures d'association dont celles présentées dans ce chapitre. Une étude comparative concernant cette fois la tâche d'extraction terminologique a été effectuée par [Laroche et Langlais, 2010]. Ces derniers ont constaté que la mesure du *discounted odds-ratio* combinée à la mesure du cosinus permettait d'obtenir les meilleurs scores en termes de rappel et de précision sur un corpus de petite taille.

3.6 Bilan

Nous avons abordé dans ce chapitre plusieurs indices liés à la caractérisation du contexte des unités lexicales. Nous sommes revenus sur les notions de mot et de terme ainsi que sur les différentes représentations contextuelles. À la question : comment améliorer la représentation du contexte pour qu'il soit le plus discriminant

possible ? Une première réponse émane de l'état de l'art, en s'appuyant notamment sur les travaux de [Ismail et Manandhar, 2010] sur les termes du domaine (In-domain terms) et les travaux de [Prochasson et Morin, 2009] sur les points d'ancrage. L'idée étant de considérer différemment les termes important du domaine spécialisé par rapport aux autres mots qui seraient moins discriminants. Une autre piste pourrait venir des deux manières de représenter le contexte (graphique et syntaxique). Ceci nous conduit à envisager la combinaison des deux représentations. La question est de savoir comment exploiter conjointement ces informations ? Nous répondrons à cette question dans le chapitre 5. À la question : est-ce qu'une bonne représentation contextuelle suffit à assurer une extraction terminologique bilingue de qualité ? Il est difficile de répondre à cela notamment à cause du caractère subjectif de cette question. Il serait plus pertinent de dire qu'une bonne représentation contextuelle augmenterait les chances de trouver la bonne traduction d'un terme donné. Sachant que les méthodes distributionnelles dépendent fortement de la caractérisation du contexte, nous faisons de celui-ci notre principal objet d'étude et proposons dans le prochain chapitre une évaluation détaillée de l'*approche directe* faisant intervenir les notions abordées ici.

III

Étude de l'approche directe

4

Mise en œuvre de l'approche directe

Introduction

Si l'*approche directe* reste à ce jour l'une des méthodes les plus efficaces concernant l'extraction bilingue de termes simples à partir de corpus comparables, les résultats obtenus dans les différents travaux présentés dans l'état de l'art sont loin d'égaler les méthodes se basant sur les corpus parallèles. La multitude de paramètres de l'*approche directe* joue sans doute un rôle important dans ses performances. Afin de mieux comprendre l'impact de chaque paramètre, nous présentons dans ce chapitre une étude détaillée de cette approche de référence du domaine. Rappelons que l'*approche directe* se compose de trois étapes fondamentales qui sont : (i) la caractérisation du mot à traduire par un vecteur représentant de son contexte, (ii) la traduction de ce vecteur de contexte à l'aide d'un dictionnaire bilingue et (iii) la comparaison du vecteur de contexte traduit avec tous les vecteurs de contexte des mots de la langue cible, pour en extraire ensuite les n plus proches comme traductions candidates. Notre but est d'analyser chacune de ces étapes afin de relever les différents manques et points à améliorer.

4.1 Méthode et ressources

Dans cette section, nous commençons par présenter l'approche fondatrice en extraction de lexiques bilingues à partir de corpus comparables. Nous décrivons ensuite les différentes ressources mobilisées pour ce travail.

4.1.1 Approche directe

Les principaux travaux en extraction de lexiques bilingues à partir de corpus comparables reposent sur la simple observation qu'un mot et sa traduction ont tendance à apparaître dans les mêmes environnements lexicaux. C'est l'idée du linguiste J. R.

Firth (1957) selon laquelle « *On reconnaît un mot à ses fréquentations.* [1] ». La mise en œuvre de cette observation s'appuie sur la caractérisation du contexte des mots pour faire émerger un ensemble de traits singuliers dénotant de l'usage des mots en contexte. Cette caractérisation, qui s'inscrit dans l'hypothèse distributionnelle de Harris (1971), revient à identifier des *affinités du premier ordre* : « *Les affinités du premier ordre décrivent les mots qui sont susceptibles d'être trouvés dans le voisinage immédiat d'un mot donné.* [2] » [Grefenstette, 1994a, p. 279].

Les contextes lexicaux sont observés d'un point de vue monolingue à travers le prisme d'une fenêtre plus ou moins étroite qui peut aller de quelques mots [Rapp, 1999, Otero, 2007] à quelques phrases [Déjean *et al.*, 2002, Daille et Morin, 2005]. Quelle que soit la représentation mise en œuvre (graphique comme syntaxique), les traits associés à un mot à caractériser sont d'autres mots. Dans le cadre d'une représentation graphique, un mot est caractérisé par les mots avec lesquels il cooccurre. À ce niveau, les mots agrammaticaux (qui ne sont pas considérés comme porteur de sens) ne sont généralement pas pris en compte. En outre, les mots sont le plus souvent lemmatisés afin de renforcer la caractérisation des contextes. Dans le cadre d'une représentation syntaxique, un mot sera quant à lui caractérisé par les relations de dépendance syntaxique qu'il entretient avec ses voisins.

Un simple dénombrement permet de déterminer la force de la relation qu'entretient un mot à caractériser avec d'autres mots du corpus. Une mesure de récurrence contextuelle est en générale préférée à un dénombrement pour limiter la portée des mots fréquents. Enfin, la traduction d'un mot est obtenue en comparant son contexte préalablement transféré en langue cible à l'aide d'un dictionnaire bilingue à l'ensemble des contextes de la langue cible.

L'implémentation que nous faisons de l'*approche directe* se décompose de la manière suivante [Rapp, 1995, Fung et Mckeown, 1997] :

Identification des contextes lexicaux

Pour chaque partie du corpus comparable, le contexte de chaque mot w_i est extrait en repérant les mots qui apparaissent autour de lui selon une caractérisation graphique [3] ou syntaxique [4]. Afin d'identifier les mots caractéristiques des contextes lexicaux et de supprimer l'effet induit par la fréquence des mots, nous normalisons l'association entre les mots sur la base d'une mesure de récurrence contextuelle comme l'*Information Mutuelle* - IM [Fano, 1961], le *Taux de vraisemblance* - TV [Dunning, 1993] ou le *discounted odds-ratio* - DOR [Evert, 2005] (cf. les équations 4.1 à 4.3 et le tableau 4.1). Après normalisation, à chaque élément w_j du vecteur de contexte du mot w_i nous attachons le taux d'association $assoc(w_i, w_j)$.

Transfert d'un mot à traduire

Le transfert d'un mot w_k à traduire de la langue source à la langue cible repose sur la traduction de chacun des éléments de son vecteur de contexte au

1. « *You shall know a word by the company it keeps.* »

2. « *First-order affinities describe what other words are likely to be found in the immediate vicinity of a given word.* »

3. Dans le cas d'une représentation graphique, les mots sont extraits dans une fenêtre contextuelle de n mots autour d'un mot w_i à caractériser et les mots agrammaticaux ne sont pas considérés.

4. Dans le cas d'une représentation syntaxique, ce sont les mots en relations de dépendance syntaxique avec un mot w_i qui sont sélectionnés pour faire partie de son contexte lexical.

moyen d'un dictionnaire bilingue. Si le dictionnaire propose plusieurs traductions pour un élément, nous ajoutons au vecteur de contexte de w_k l'ensemble des traductions proposées [5] (lesquelles sont pondérées par la fréquence de la traduction en langue cible). Dans le cas où l'élément n'est pas présent dans le dictionnaire, il ne sera pas exploité dans le processus de traduction.

Identification des vecteurs proches du mot à traduire

Le vecteur de contexte v_i du mot w_i ainsi traduit est ensuite comparé à l'ensemble des vecteurs de la langue cible en s'appuyant sur une mesure de distance vectorielle comme le *Cosinus* - COS [Salton et Lesk, 1968] ou le *Jaccard pondéré* - JAC [Grefenstette, 1994b] (cf. les équations 4.4 et 4.5).

Obtention des traductions candidates

En fonction des précédentes valeurs de similarité, nous obtenons une liste ordonnée de traductions candidates pour le mot w_i.

	w_j	$\neg w_j$
w_i	$a = occ(w_i, w_j)$	$b = occ(w_i, \neg w_j)$
$\neg w_i$	$c = occ(\neg w_i, w_j)$	$d = occ(\neg w_i, \neg w_j)$

TABLE 4.1 – Table de contingence

$$\mathrm{IM}(w_i, w_j) = log\frac{a}{(a+b)(a+c)} \tag{4.1}$$

avec $occ(w_i, w_j)$ qui correspond à la valeur de cooccurrence entre les mots w_i et w_j.

$$\begin{aligned}\mathrm{TV}(w_i, w_j) = \; & a\log(a) + b\log(b) + c\log(c) + d\log(d) \\ & +(a+b+c+d)\log(a+b+c+d) - (a+b)\log(a+b) \\ & -(a+c)\log(a+c) - (b+d)\log(b+d) - (c+d)\log(c+d)\end{aligned} \tag{4.2}$$

$$\mathrm{DOR}(w_i, w_j) = \log\frac{(a+\frac{1}{2}) \times (d+\frac{1}{2})}{(b+\frac{1}{2}) \times (c+\frac{1}{2})} \tag{4.3}$$

$$\mathrm{COS}_{v_i}^{v_k} = \frac{\sum_t assoc(w_i, w_t)\, assoc(w_k, w_t)}{\sqrt{\sum_t assoc(w_i, w_t)^2}\sqrt{\sum_t assoc(w_k, w_t)^2}} \tag{4.4}$$

$$\mathrm{JAC}_{v_i}^{v_k} = \frac{\sum_t min(assoc(w_i, w_t), assoc(w_k, w_t))}{\sum_t max(assoc(w_i, w_t), assoc(w_k, w_t))} \tag{4.5}$$

4.1.2 Ressources

Dans nos expérimentations, nous avons besoin de trois ressources linguistiques, à savoir : i) un corpus comparable, ii) un dictionnaire bilingue et iii) une liste d'évaluation. Nous commençons par présenter ces trois ressources, puis nous décrivons la ressource exploitée pour identifier les relations de dépendance syntaxique.

5. Cela correspond à une approche classique lorsque les traductions ne sont pas ordonnées dans le dictionnaire bilingue. D'autres techniques ont été proposées par Bouamor *et al.* (2013).

Corpus comparables

Nous avons utilisé des corpus comparables français/anglais dont trois de langues spécialisées et un de langue générale :

Corpus du cancer du sein Le corpus du cancer du sein a été construit à partir de documents extraits du portail Elsevier[6]. L'ensemble des documents collectés relève du domaine médical restreint à la thématique du « cancer du sein » et comporte, dans le titre ou dans les mots clés, le terme *cancer du sein* en français et *breast cancer* en anglais pour la période de 2001 à 2008. Le corpus comparable spécialisé obtenu est composé de 130 documents pour le français (7 376 mots distincts) et 103 documents pour l'anglais (8 457 mots distincts) pour une taille d'environ 1 million de mots.

Corpus des énergies renouvelables Le corpus des énergies renouvelables a été construit à partir du web à l'aide du crawler *Babouk* [Groc, 2011]. Pour la recherche et l'extraction des différents documents du corpus, le crawler s'est appuyé sur des mots clés du domaine tels que *wind, energy, rotor...* en anglais et *vent, énergies, éolien, renouvelables...* en français. Nous disposons ainsi d'un corpus comparable spécialisé d'environ 600 000 mots avec 5 606 mots distincts pour le français et 6 081 mots distincts pour l'anglais.

Corpus de vulcanologie Le corpus de vulcanologie a été construit manuellement par Amélie Josselin-Leray du Laboratoire CLLE-ERSS. Il contient des documents extraits du web, des manuels universitaires, des ouvrages de vulgarisation scientifique, des quotidiens généralistes, des magazines de vulgarisation et de semi-vulgarisation ainsi que des magazines de voyages/découverte et des glossaires. La taille du corpus est d'environ 800 000 mots avec 9 142 mots distincts pour le français et 8 623 mots distincts pour l'anglais.

Corpus journalistique Le corpus journalistique a été construit à partir des publications électroniques du journal français *Le Monde* et du journal américain *Los Angeles Times* de l'année 1994. Dans le but de construire des modèles d'apprentissage[7] , nous avons extrait deux versions de ce corpus, une première version de petite taille d'environ 500 000 mots et une seconde version de grande taille d'environ 10 millions de mots.

Pour chaque corpus, les documents sont nettoyés et normalisés à travers les traitements suivants : segmentation en occurrences de formes, étiquetage morpho-syntaxique et lemmatisation.

Dictionnaire bilingue

Nous avons sélectionné deux dictionnaires français-anglais. Il s'agit du dictionnaire ELRA-M0033[8] de 200 000 entrées et du dictionnaire wordreference[9] (WDREF) de 22 300 entrées environ. Nous comparons ces deux dictionnaires en donnant quelques statistiques sur leurs couples de traduction. Nous illustrons dans la table 4.2

6. www.elsevier.com
7. Ce corpus est utilisé dans le chapitre 6 pour la ré-estimation des cooccurrences des mots.
8. http://www.elra.info/
9. http://www.wordreference.com/

la fréquence des couples de traduction en fonction de leurs catégories grammaticales. Dans la table 4.3 nous illustrons les mêmes informations mais cette fois après projection des dictionnaires sur les corpus comparables.

POS(FR-EN)	# WDREF	# ELRA
N N	27 656	127 236
N V	0	3
N J	9	29
V V	20 054	52 697
V N	0	2
V J	1	0
J J	12 632	47 827
J N	147	429
J V	0	10

TABLE 4.2 – Comparaison des couples de traduction des dictionnaires WDREF et ELRA

La table 4.2 montre qu'en termes de nombre d'entrées, le dictionnaire ELRA est largement supérieur au dictionnaire WDREF et ce, quelles que soient les catégories grammaticales des couples de traduction. On remarquera aussi une prédominance des couples Nom/Nom (N N) pour les deux dictionnaires.

	Cancer du sein		Énergies renouvelables		Vulcanologie	
(FR->EN)	# WDREF	# ELRA	# WDREF	# ELRA	# WDREF	# ELRA
N N	4435	5881	4911	6216	8328	10770
N V	0	0	0	0	0	0
N J	1	1	2	2	2	5
V V	3405	3940	3721	4200	6393	7675
V N	0	0	0	0	0	0
V J	0	0	2	0	0	0
J J	1673	2358	1596	2149	2986	4229
J N	30	35	33	45	59	64
J V	0	0	0	0	0	0

TABLE 4.3 – Comparaison des couples de traduction des dictionnaires WDREF et ELRA après projection sur les corpus comparables

La table 4.3 montre que la différence en termes de couples de traduction est nettement moins prononcée après projection des deux dictionnaires sur les trois corpus comparables. On remarquera que le rapport de prédominance des couples N N reste prépondérant. On remarquera aussi la faible proportion de couples de traduction adjectif/Nom (J N). Parmi les trois corpus comparables, le corpus de vulcanologie contient le plus de couples de traduction distincts. Les corpus du cancer du sein et des énergies renouvelables sont plus ou moins équivalents de ce point de vue.

Listes d'évaluation

Pour construire les listes de couples de traduction nécessaires à l'évaluation des approches mises en œuvre, nous nous appuyons sur des nomenclatures attestées des termes du domaine. Cette méthode de création d'une liste de référence est différente de celle proposée par Déjean *et al.* (2002) qui construisent leur liste à partir d'un sous-ensemble du dictionnaire bilingue. Nous pensons que cette approche, plus fiable d'un point de vue statistique, ne correspond pas aux véritables difficultés rencontrées avec des corpus spécialisés. En domaine spécialisé, les termes qui représentent une difficulté de traduction n'appartiennent que rarement au dictionnaire de langue générale. Nous sélectionnons les couples de traduction pour lesquels le mot français apparaît au moins cinq fois dans la partie française et sa traduction au moins cinq fois dans la partie anglaise du corpus comparable. Ce choix est motivé par la nécessité d'avoir un minimum de contexte nécessaire pour utiliser l'*approche directe*, ce qui est clairement difficile voire impossible pour des termes ayant une fréquence très faible.

Pour le corpus du cancer du sein, nous obtenons 321 couples de termes simples français-anglais à partir du meta-thesaurus UMLS[10] et du *Grand dictionnaire terminologique*[11]. Nous obtenons 150 couples pour le corpus des énergies renouvelables et 158 pour celui de vulcanologie à l'aide de lexiques et de glossaires de chaque domaine.

Il n'est pas aisé de juger de la difficulté de la tâche d'extraction de termes bilingues à partir de corpus comparables. Le fait d'utiliser plusieurs ressources linguistiques signifie que la difficulté peut émaner de chacune d'elle voire de toutes les ressources. Il n'existe pas, à notre connaissance, une manière de mesurer la difficulté de notre tâche. Néanmoins, nous pouvons supposer que si le corpus ou le dictionnaire bilingue est de mauvaise qualité alors la tâche est difficile. Nous pouvons ajouter que si la liste d'évaluation contient beaucoup de termes peu fréquents ou des couples de traduction ayant un rapport de fréquence (spécificité) très élevé, alors la tâche est plus difficile. Il est souvent plus facile de traduire des termes très fréquents. Or, en domaine de spécialité, nous sommes principalement confrontés à des termes moyennement voire peu fréquents. Pour avoir une idée de la fréquence des termes à traduire et de la fréquence leur bonne traduction, nous présentons dans le tableau 4.4 une représentation des trois listes d'évaluation selon plusieurs plages de fréquences.

	Cancer du sein			**Énergies renouvelables**			**Vulcanologie**		
Plage	#En	#Fr	#En/Fr	#En	#Fr	#En/Fr	#En	#Fr	#En/Fr
[5,10]	57	67	25	10	15	3	5	9	0
]10,50]	132	137	71	38	48	20	60	60	38
]50,100]	50	40	9	41	29	15	24	26	9
]100,500]	73	64	32	47	47	22	54	50	30
]500,1000]	4	7	1	8	6	1	8	9	4
]1000,5000]	5	6	5	6	5	3	7	4	4

TABLE 4.4 – Comparaison des listes d'évaluation par plage de fréquences

10. www.nlm.nih.gov/research/umls
11. www.granddictionnaire.com

Ce tableau indique la distribution fréquentielle des termes source et cible des listes d'évaluation ainsi que des couples de traduction selon différentes plages de valeurs. Nous pouvons remarquer qu'une bonne partie des couples de traduction ne partage pas les mêmes plages de valeurs. Si nous prenons par exemple la plage [5,10] de la liste d'évaluation du cancer du sein, seulement 25 couples de traduction appartiennent à cette plage. Si l'appartenance des couples de traduction aux mêmes plages de valeurs ne peut en aucun cas suffire à mesurer le degré de difficulté d'une liste d'évaluation, celle-ci peut néanmoins indiquer une tendance dans des cas spécifiques. Par exemple, si un terme source appartient à la plage]10,50] et sa traduction à la plage]1000,5000], dans ce cas, la taille du vecteur de contexte du terme source serait au maximum de 300 mots et celle de sa traduction de 6 000 mots au minimum, et ceci pour une taille de fenêtre égale à 7 (3 mots avant et 3 mots après le terme à caractériser). La différence de taille des vecteurs de contexte engendrée par une trop grande différence de fréquence pourrait justifier de la difficulté à traiter des couples de traduction se trouvant dans cette configuration.

Identification des relations de dépendance syntaxique

Concernant l'extraction des relations de dépendance syntaxique, nous avons utilisé l'outil fourni par [Gamallo, 2008a][12]. Ci-dessous un exemple de l'application de cette analyseur.

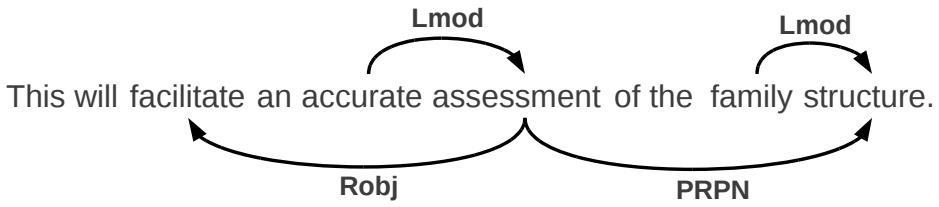


FIGURE 4.1 – Illustration des relations de dépendance syntaxique fournies par l'-analyseur syntaxique appliqué sur une phrase anglaise extraite du corpus du cancer du sein.

Le résultat de l'analyseur syntaxique sur l'exemple illustré par la figure 4.1 montre qu'il existe une relation Lmod entre *accurate* et *assessment*, Robj entre *facilitate* et *assessment*, Lmod entre *family* et *structure* et PRPN entre *assessment* et *structure*. Ainsi, le vecteur de contexte du mot *assessment* par exemple, contiendra le mot *accurate* avec une étiquette Lmod ($accurate_{Lmod}$) et une valeur de cooccurrence équivalente au nombre de fois que *accurate* apparaît comme modificateur gauche de *assessment* dans le corpus.

12. `http://gramatica.usc.es/pln/tools/deppattern.html`

4.2 Évaluation

L'évaluation de l'*approche directe* est effectuée en utilisant les différentes ressources présentées dans les précédentes sections. Nous étudions les paramètres suivants :

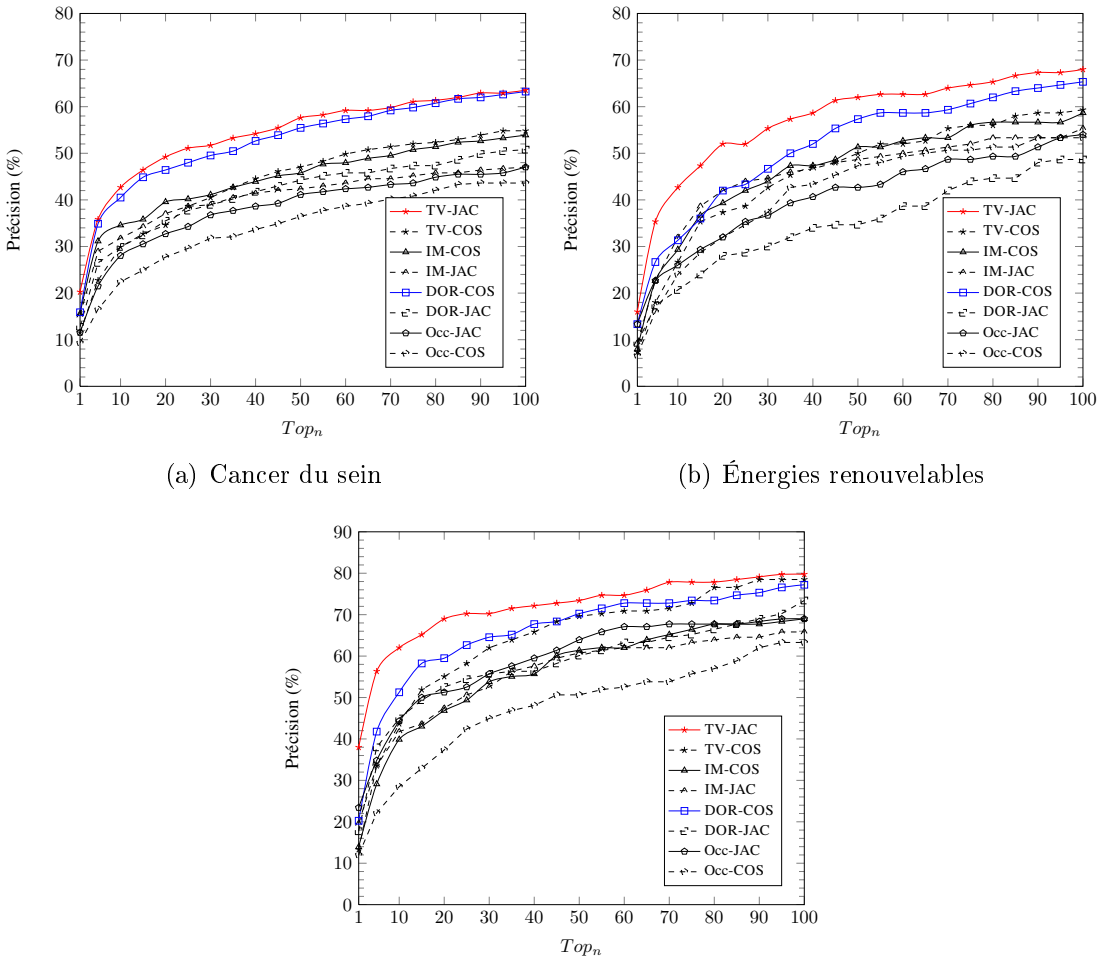
1. Les mesures d'association et de similarité ;
2. La taille des fenêtres contextuelles ;
3. L'impact du dictionnaire bilingue utilisé ;
4. La taille des vecteurs de contexte.

4.2.1 Variation des mesures d'association et de similarité

Dans cette première expérience nous explorons différentes combinaisons de mesures d'association et de similarité. Nous choisissons d'utiliser les trois mesures d'association qui reviennent le plus dans l'état de l'art, à savoir l'*Information Mutuelle* (IM), le *Taux de vraisemblance* (TV) et le *discounted odds-ratio* (DOR). À titre comparatif, nous ajoutons une quatrième mesure qui est simplement le nombre de cooccurrences des mots (Occ). Nous sélectionnons aussi comme mesures de similarité le *cosinus* (COS) et le *Jaccard* (JAC). Il est à noter que d'autres mesures d'association et de similarité ont été étudiées sans résultats satisfaisants.

La figure 4.2 montre que c'est la combinaison TV-JAC qui donne les meilleurs résultats sur les trois corpus comparables. Concernant le corpus du cancer du sein, la courbe de la configuration DOR-COS est très proche de celle de la configuration TV-JAC. La configuration IM-COS, quant à elle, arrive en troisième position dans les premiers tops, elle est dépassée ensuite par TV-COS à partir du top 40. Les plus mauvais résultats sont obtenus par les configurations Occ-COS et Occ-JAC. Pour les deux autres corpus, la supériorité de la configuration Log-JAC est beaucoup plus nette et plus particulièrement dans les premiers tops. La configuration DOR-COS arrive en deuxième position. Elle se fait toutefois rejoindre par la configuration TV-COS dans les derniers tops pour le corpus de vulcanologie. Il est à noter que la configuration Occ-COS obtient toujours les moins bons résultats. Nous remarquerons aussi que la configuration Occ-JAC dépasse la configuration IM-COS sur le corpus de vulcanologie.

Nous retiendrons de cette expérience trois configurations. Celle du TV-JAC qui donne les meilleurs résultats sur ce jeu d'expériences. La configuration DOR-COS qui montre aussi des résultats intéressants. Et la configuration IM-COS, cette dernière, bien qu'elle soit en deçà des deux autres configurations avec des résultats moyens, est souvent utilisée. Nous souhaitons donc l'étudier dans d'autres situations.



(a) Cancer du sein

(b) Énergies renouvelables

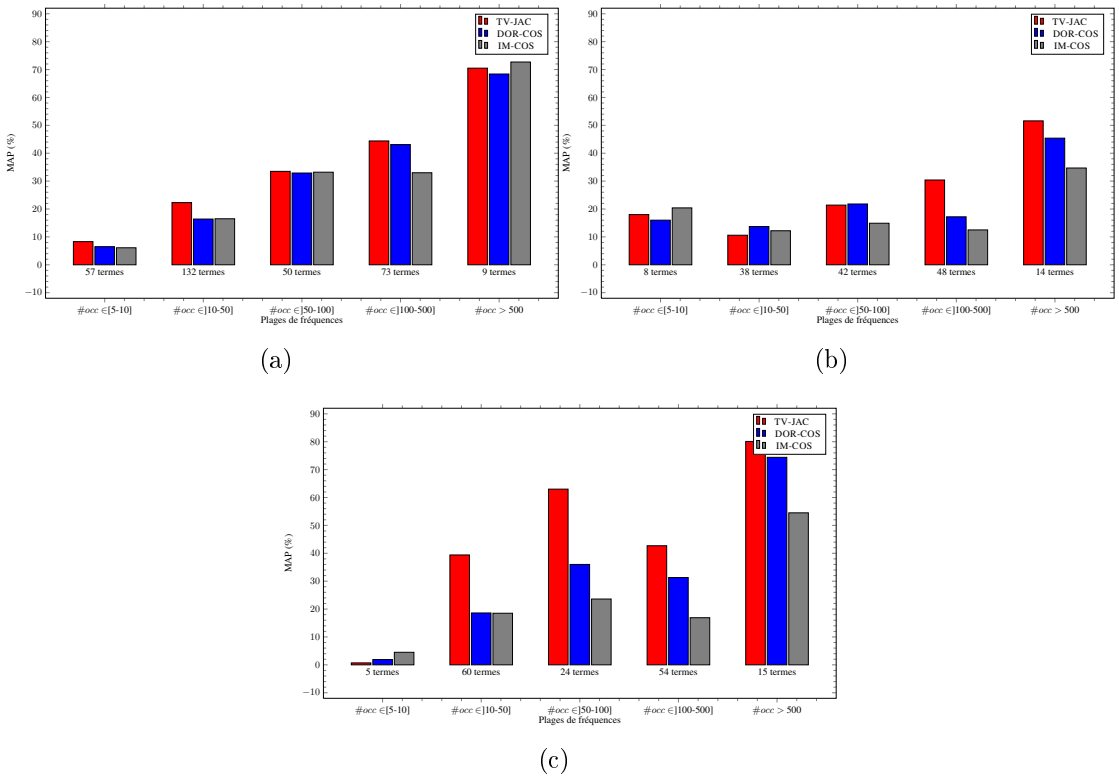
(c) Vulcanologie

FIGURE 4.2 – Approche directe : Comparaison des mesures d'association et de similarité

Comparaison des mesures d'association et de similarité par plages de fréquences

Le but de cette deuxième expérience est de comparer les résultats des trois configurations retenues (TV-JAC, DOR-COS et IM-COS) par plages de fréquence des termes à traduire. Ainsi $\#occ \in [5-10]$ par exemple veut dire qu'un terme de la liste de référence apparaît entre 5 et 10 fois dans la partie anglaise du corpus comparable.

La figure 4.3 illustre cette comparaison en fonction de la MAP (%). Nous constatons que les résultats de la configuration TV-JAC sont globalement supérieurs à ceux des autres configurations sur pratiquement toutes les plages de fréquences. Cette supériorité est d'autant plus remarquable sur le corpus de vulcanologie (mis à part la plage $\#occ \in [5-10]$). Au vu de ces résultats il est difficile de statuer sur les performances d'une configuration par rapport à une plage de fréquences donnée. Si nous prenons par exemple la configuration IM-COS, nous remarquons que ses performances sont variables d'un corpus à l'autre pour une même plage de fréquences (plage $\#occ \in [5-10]$ par exemple). Nous pouvons aussi noter l'équivalence des performances des configurations IM-COS et DOR-COS pour les plages de faibles fréquences, et plus particulièrement pour la plage $\#occ \in [10-50]$.



(a) (b)

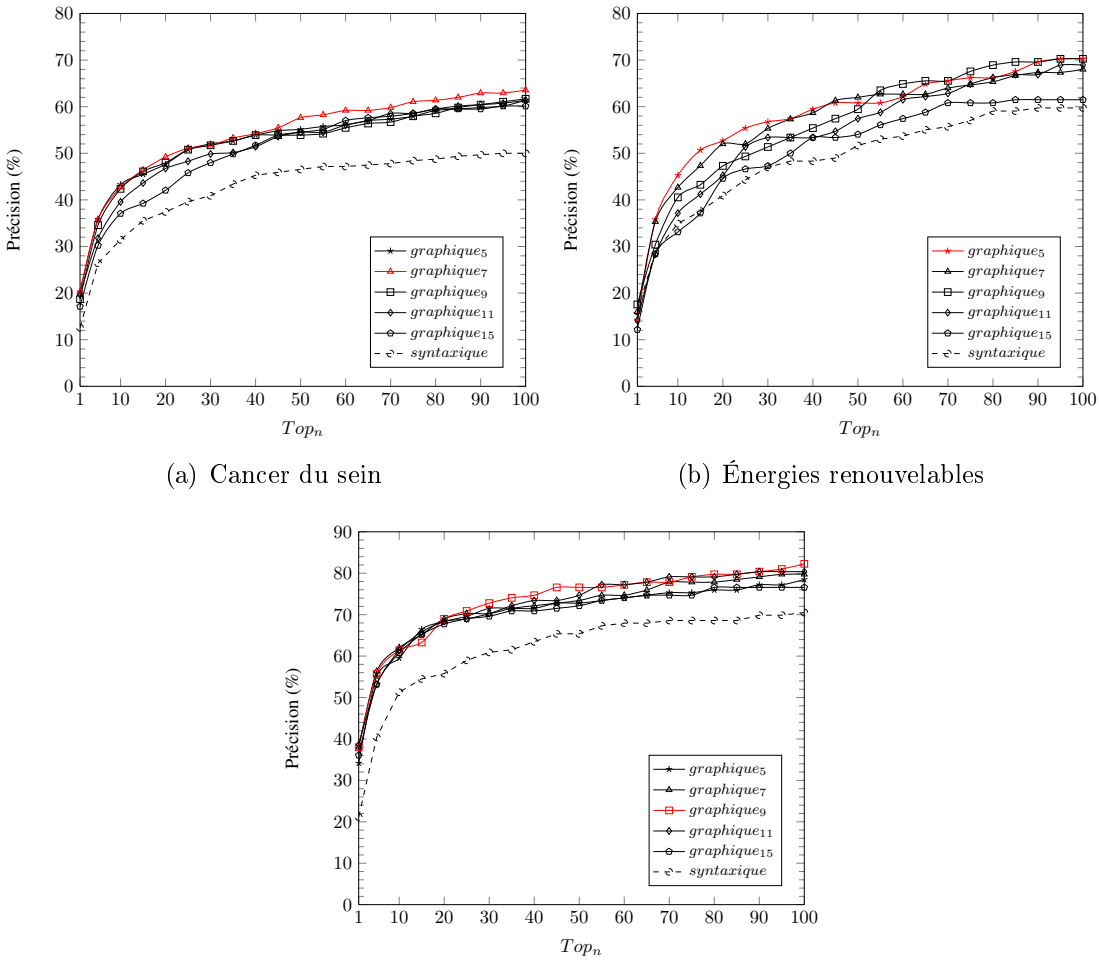
(c)

FIGURE 4.3 – Approche directe : Comparaison des mesures d'association et de similarité par plages de fréquences

4.2.2 Variation de la taille des fenêtres contextuelles

Dans cette expérience nous comparons, d'une part plusieurs tailles de fenêtres contextuelles entre elles, et d'autre part, ces dernières avec la représentation par relations de dépendance syntaxique (*syntaxique*). Une taille de fenêtre représentée par $graphique_5$ par exemple, signifie que nous choisissons les deux mots se trouvant avant et les deux mots se trouvant après le mot à caractériser. Nous utilisons dans cette expérience la configuration TV-JAC.

La figure 4.4 montre que ce sont les fenêtres de petite taille qui donnent les meilleurs résultats et ce pour les corpus du cancer du sein et des énergies renouvelables. Ce constat est moins clair pour le corpus de vulcanologie. C'est une fenêtre de taille 7 qui donne globalement les meilleurs résultats pour le corpus du cancer du sein. Concernant celui des énergies renouvelables, une fenêtre de taille 5 offre les meilleures performances, même si elle est rejointe et dépassée par la courbe $graphique_9$ dans les derniers tops. Pour le corpus de vulcanologie, et bien que les résultats des fenêtres soient beaucoup plus rapprochés que les deux autres corpus, nous pouvons néanmoins remarquer que la fenêtre de taille 9 montre les meilleures performances. Enfin, concernant la courbe représentative de la configuration par relations de dépendance syntaxique, elle est en deçà de celles par fenêtres contextuelles sur les trois corpus.



(a) Cancer du sein

(b) Énergies renouvelables

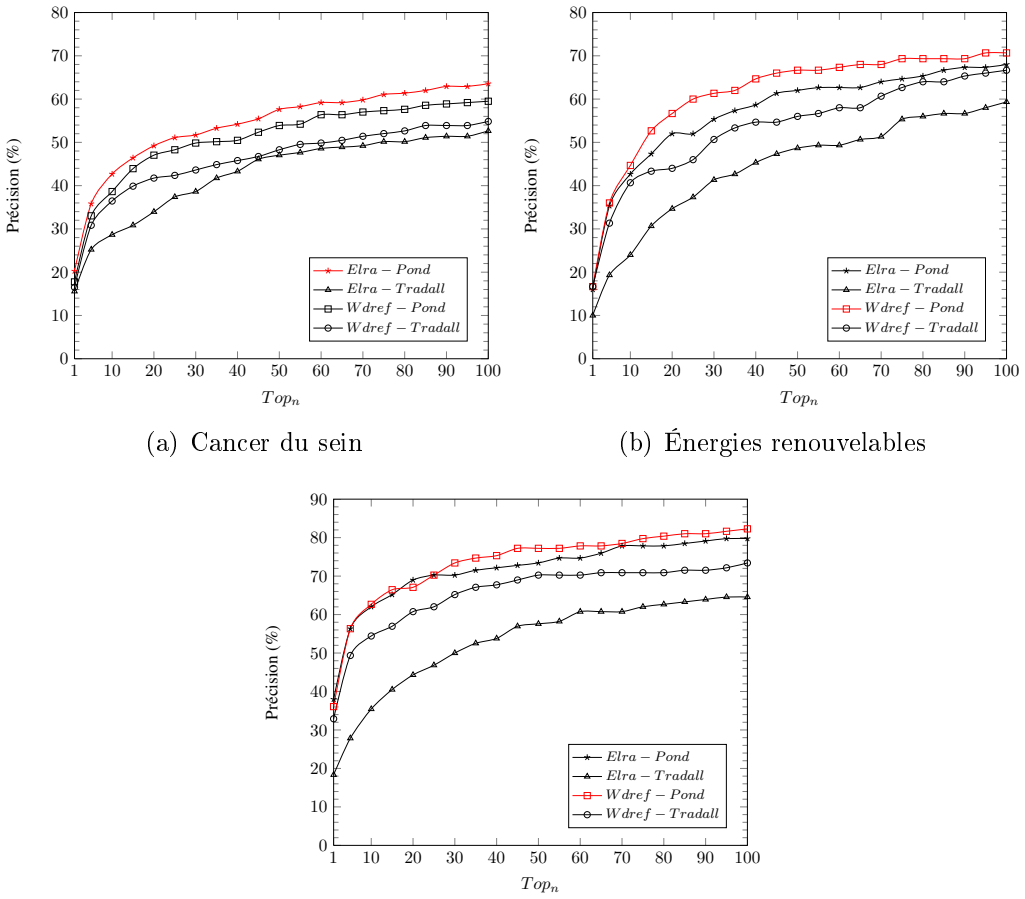
(c) Vulcanologie

FIGURE 4.4 – Approche directe : Comparaison des représentations contextuelles graphique et syntaxique

4.2.3 Dictionnaires bilingues

Sachant que le dictionnaire bilingue joue un rôle important dans l'*approche directe*, que ce soit en termes de qualité ou de couverture, nous nous proposons dans cette expérience de comparer deux dictionnaires bilingues, à savoir ELRA et WDREF. Lors de la phase de traduction des vecteurs de contexte, et pour chaque mot source ayant plusieurs traductions, nous choisissons deux stratégies. La première consiste à prendre toutes les traductions et à leur attribuer le même score que celui du mot source. La deuxième stratégie consiste à pondérer ce score par la fréquence en langue cible de chaque traduction.

Le premier constat que nous pouvons dégager de la figure 4.5 est que la stratégie de pondération des traductions est nettement meilleure que celle qui consiste à attribuer le même score d'association à chaque traduction et ce pour les deux dictionnaires bilingues et les trois corpus comparables. Le deuxième constat concerne la comparaison des performances des deux dictionnaires. Pour le corpus du cancer du sein, le dictionnaire ELRA donne de meilleurs résultats que WDREF pour la stratégie de pondération alors que pour l'autre stratégie, c'est WDREF qui donne



(a) Cancer du sein

(b) Énergies renouvelables

(c) Vulcanologie

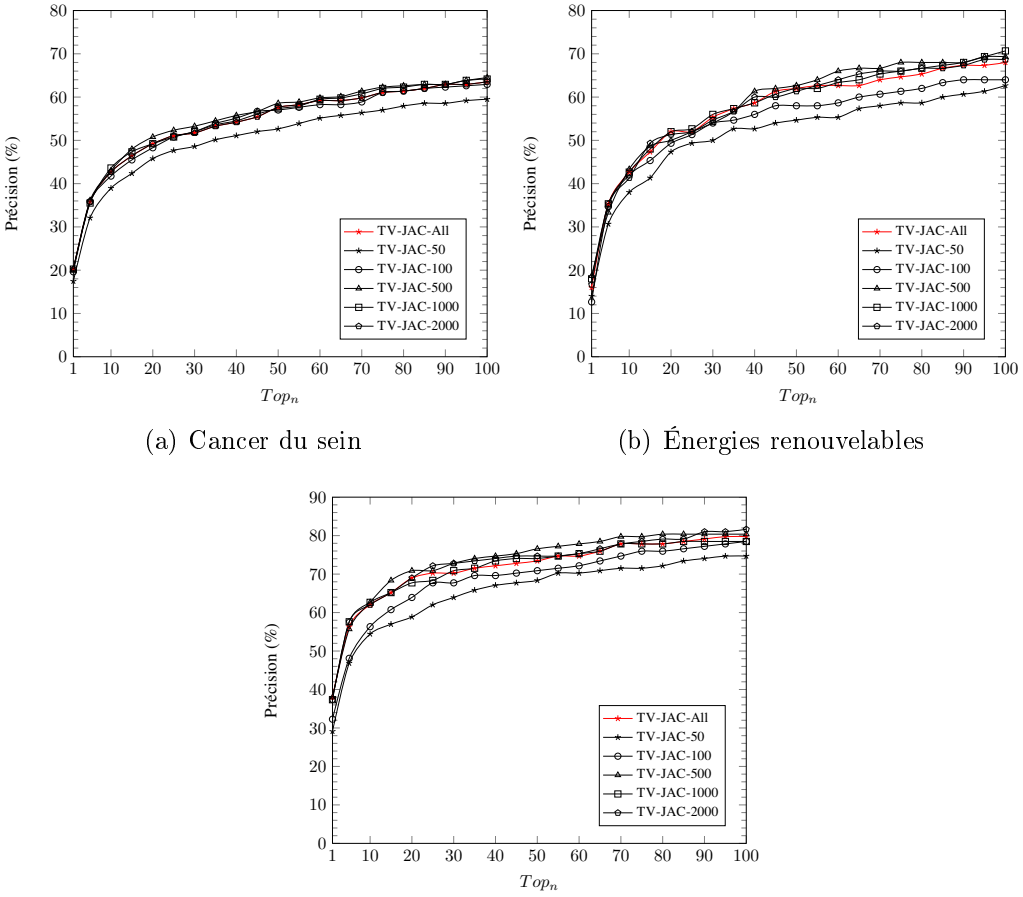
FIGURE 4.5 – Approche directe : Comparaison des dictionnaires ELRA et WDREF

les meilleurs résultats. En revanche, pour le corpus des énergies renouvelables, c'est WDREF qui donne les meilleurs résultats si on compare les deux dictionnaires par stratégie. Enfin, pour le corpus de vulcanologie et concernant la stratégie de traduction par pondération, les courbes des deux dictionnaires sont très proches dans les premiers tops, par la suite nous remarquons de meilleurs résultats pour WDREF à partir du top 25. Pour l'autre stratégie, WDREF obtient largement de meilleurs résultats que ELRA.

4.2.4 Variation de la taille des vecteurs de contexte

Dans cette dernière expérience nous faisons varier la taille des vecteurs de contexte. Le but est de voir l'impact de la variation de ce paramètre sur les performances de l'*approche directe* et ce, pour les trois configurations que nous avons retenu dans la première expérience. Nous fixons la taille de la fenêtre contextuelle à 7. La courbe en rouge représente l'*approche directe* en gardant toutes les entrées des vecteurs de contexte des corpus source et cible.

La figure 4.6 montre que le nombre d'entrées des vecteurs de contexte influe sur les résultats de l'*approche directe*. Globalement, le comportement de l'*approche directe*



(a) Cancer du sein

(b) Énergies renouvelables

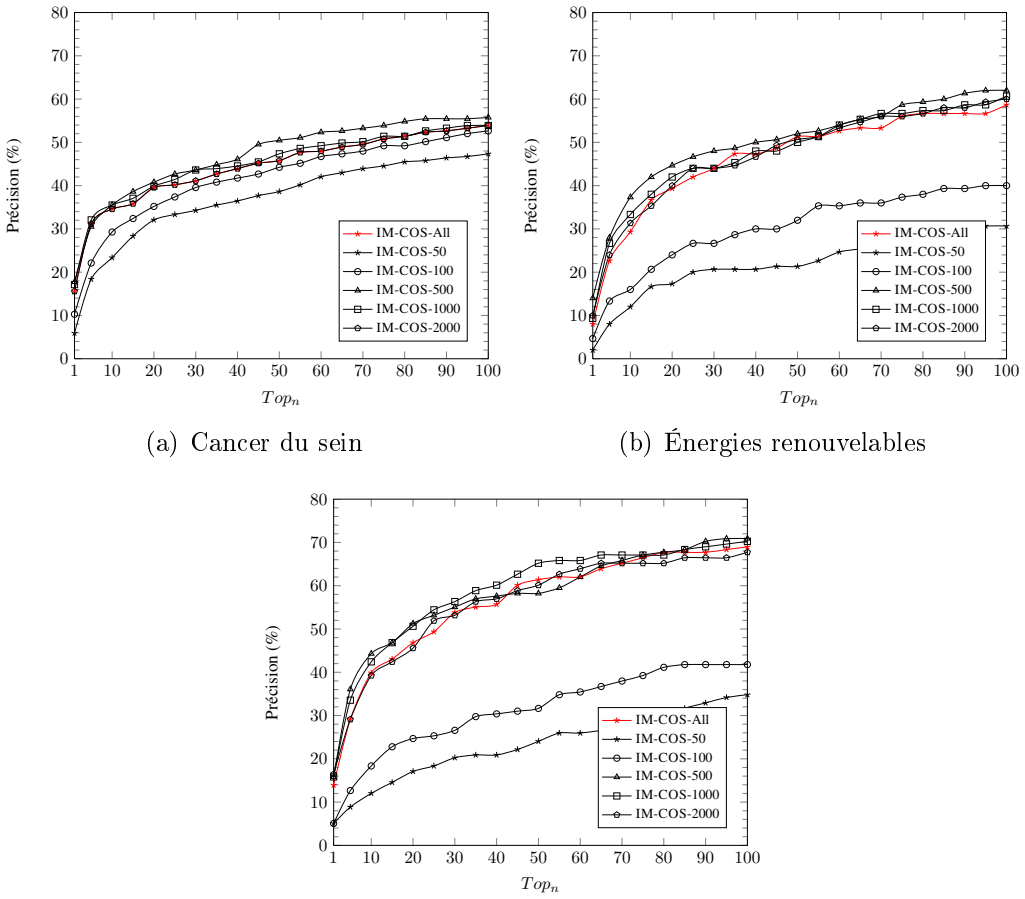
(c) Vulcanologie

FIGURE 4.6 – Approche directe : Variation de la taille des vecteurs de contexte pour la combinaison TV-JAC

reste le même pour les trois corpus comparables. En effet, d'un côté l'utilisation des vecteurs de contexte de petite taille donne de mauvais résultats, et d'un autre côté, l'utilisation des vecteurs de contexte dans leur globalité n'est pas la solution optimale. Nous remarquons que pour les trois corpus, une taille de 500 à 1 000 entrées donne les meilleurs résultats pour la configuration TV-JAC.

La figure 4.7 montre que la configuration IM-COS est plus sensible à la taille des vecteurs de contexte. Cette constatation est très remarquable pour les corpus des énergies renouvelables et de vulcanologie. De même que pour la configuration TV-JAC, nous remarquons qu'une taille de 500 à 1 000 entrées donne les meilleurs résultats pour la configuration IM-COS.

À la différence des deux précédentes figures, la figure 4.8 montre que l'utilisation de la totalité des entrées des vecteurs de contexte ne dégrade pas ou peu les résultats pour la configuration DOR-COS. Globalement, les résultats sont équivalents pour une taille de vecteur à partir de 1 000 entrées. Il est à noter par ailleurs que des vecteurs de petite taille donnent de très mauvais résultats. De par ces constatations il semble que la configuration DOR-COS soit beaucoup plus sensible à des vecteurs de petite taille.



(a) Cancer du sein

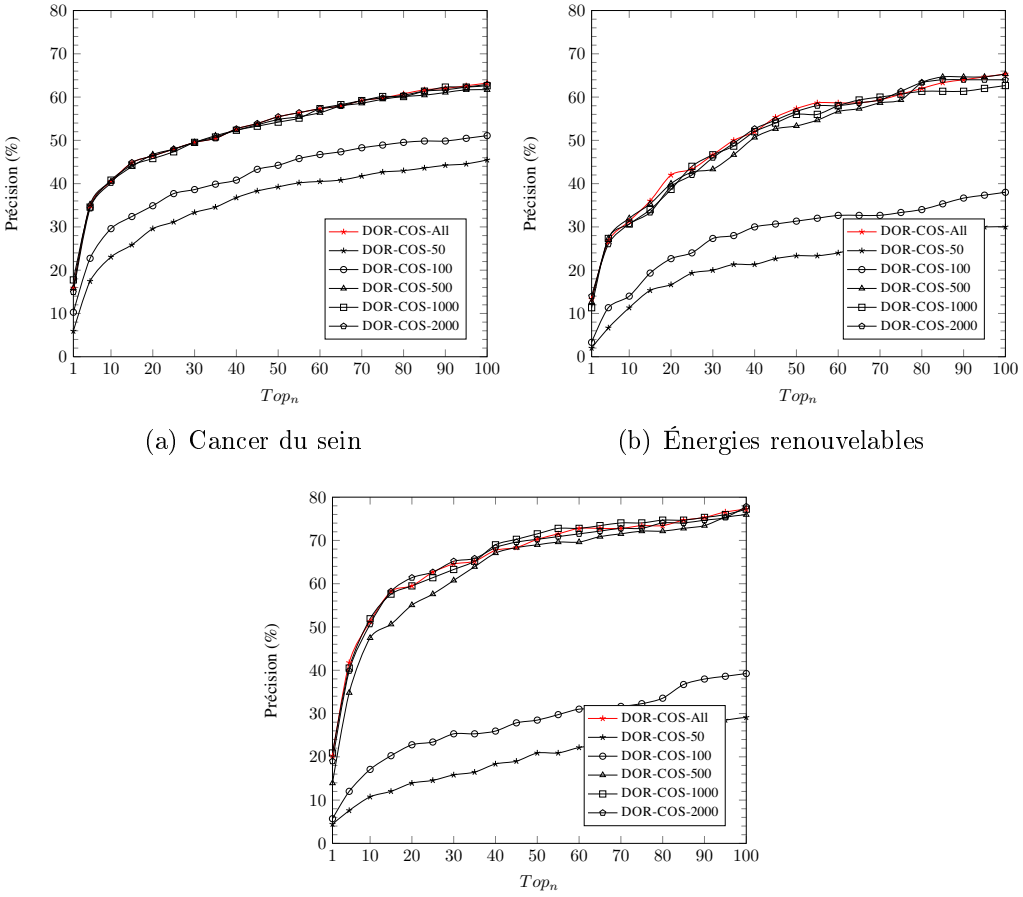
(b) Énergies renouvelables

(c) Vulcanologie

FIGURE 4.7 – Approche directe : Variation de la taille des vecteurs de contexte pour la combinaison IM-COS

4.3 Discussion

Les quatre expériences confirment la nécessité d'une bonne paramétrisation de l'*approche directe* et de l'utilisation de ressources de bonne qualité pour obtenir de bons résultats. La première expérience a montré que le choix de la mesure d'association et de similarité était primordial. Nous avons pu constater que l'association du *taux de vraisemblance* et du *Jaccard* (TV-JAC) était la meilleure configuration sur les trois corpus utilisés. Rappelons que le principe des mesures d'association est d'indiquer la force de la relation entre deux mots. En se basant sur ces résultats expérimentaux nous pouvons dire que la mesure du *taux de vraisemblance* est sans doute la plus adéquate dans ce sens, suivie par la mesure du *discounted odds-ratio*. Le point commun entre ces deux mesures est qu'elles se basent toutes les deux sur l'ensemble de la table de contingence. L'information capturée prend en compte, en plus de l'information conjointe entre deux mots, des informations disjointes qui peuvent expliquer les meilleures performances de ces mesures en comparaison avec l'*information mutuelle* par exemple, qui ne considère que l'information conjointe entre deux mots. Ceci est sans doute l'une des raisons des moins bons résultats de la mesure IM. De plus, l'*information mutuelle* a tendance à surestimer les cooccurrences de faible fréquence. Cela peut être plus problématique quand il s'agit de



(a) Cancer du sein

(b) Énergies renouvelables

(c) Vulcanologie

FIGURE 4.8 – Approche directe : Variation de la taille des vecteurs de contexte pour la combinaison DOR-COS

corpus de petite taille comme les corpus de langue de spécialité. Ceci étant dit, l'inférence statistique à partir d'une petite quantité de données est un problème bien connu. L'*information mutuelle*, souvent utilisée en recherche d'information sur de grandes quantités de données, semble être inadaptée pour des corpus de petite taille. [Dunning, 1993] a montré l'applicabilité de la mesure du *taux de vraisemblance* sur des corpus de petite taille. Ceci a été confirmé par les résultats empiriques que nous avons obtenu. Enfin, nous avons pu constater dans les travaux de [Laroche et Langlais, 2010] que les meilleures performances étaient obtenues en utilisant le *discounted odds-ratio* sur des corpus de petite taille. La divergence de nos résultats avec ceux montrés dans les travaux de [Laroche et Langlais, 2010] nous renvoie vers la complexité de la tâche qui dépend de plusieurs paramètres rendant difficile une conclusion formelle sur la mesure d'association la plus adaptée. Nous parlerons donc dans ce cas, d'une tendance favorable dans un contexte particulier pour les mesures du *taux de vraisemblance* et du *discounted odds-ratio* dépendant des ressources utilisées (corpus, dictionnaire et liste d'évaluation).

La mesure de similarité joue aussi un rôle important, comme nous avons également pu le constater dans nos expériences. En effet, une mesure d'association combinée à différentes mesures de similarité peut donner des résultats très différents. C'est le cas par exemple pour les configurations DOR-COS et DOR-JAC ou encore IM-COS ou IM-JAC. Ici encore, bien choisir ce paramètre semble être aussi

important que le choix de la mesure d'association. Cependant, il est difficile de statuer sur la meilleure mesure de similarité puisque celle-ci dépend de la mesure d'association utilisée. Il serait sans doute intéressant d'étudier plus en détail les particularités des mesures d'association et de similarité dans ce contexte afin de mieux comprendre l'origine des affinités ou de la non compatibilité des unes avec les autres, mais ceci dépasse l'objet de cette étude qui vise principalement à trouver les meilleures configurations. Par ailleurs, Evert [Evert, 2005] traite dans ses travaux de thèse une multitude de mesures d'association. Enfin, nous retiendrons de la première expérience de ce chapitre les combinaisons DOR-COS et TV-JAC comme les plus adaptées à notre tâche.

La deuxième expérience a montré l'importance du choix de la taille de la fenêtre contextuelle lors de l'utilisation d'une représentation graphique du contexte. La variabilité des résultats selon la taille de la fenêtre suggère que l'utilisation d'une taille fixe pour tous les mots du corpus n'est pas la meilleure solution. [Prochasson, 2010] a traité ce problème dans ses travaux en montrant que les mots très fréquents étaient mieux caractérisés par des fenêtres de petite taille (< 3) et inversement, les mots moins fréquents voire rares étaient mieux caractérisés par des fenêtres de grande taille (> 15). Ceci dit, le choix de la taille de la fenêtre en fonction de la fréquence du mot reste une démarche empirique et l'utilisation d'une représentation syntaxique par relations de dépendance nous dispenserait d'un choix arbitraire de cette taille comme l'a montré [Gamallo, 2008a] dans ses travaux. Cependant, nous avons constaté le contraire dans nos expériences. En effet, l'approche par représentation syntaxique a montré de moins bons résultats que celle par représentation graphique. Nous expliquons cela en partie par la différence de taille des corpus utilisés, celle-ci se mesurant en millions de mots dans les expériences de [Gamallo, 2008a]. Une autre explication pourrait venir des erreurs qui peuvent être engendrées par l'analyseur syntaxique utilisé. Aussi, une étude de différents analyseurs syntaxiques serait sans doute nécessaire pour enrichir ce travail.

Nous avons utilisé dans la troisième expérience deux dictionnaires bilingues de taille très différente. Le dictionnaire ELRA de 200 000 mots et WDREF de 22 300 mots. Nous avons pu remarquer que les meilleurs résultats étaient globalement obtenus en utilisant WDREF sur deux des trois corpus utilisés. Nous déduisons de ce constat qu'un dictionnaire de très grande taille n'est pas forcement plus favorable pour la tâche de l'extraction de lexiques bilingues à partir de corpus comparables. Ceci peut s'expliquer d'une part, par la polysémie des mots qui fait que parfois la traduction se retrouve hors contexte et peut donc constituer un bruit et d'autre part, par la manière de pondérer les traductions des entrées du vecteur à traduire. Nous avons pu constater la différence de résultats obtenus pour les deux stratégies de traduction. L'*approche directe* est très sensible au dictionnaire utilisé et au choix des traductions de chaque entrée du vecteur de contexte. Une attention plus particulière doit être portée à cette phase pour optimiser au mieux l'*approche directe*. Par ailleurs, il est à noter que récemment [Bouamor *et al.*, 2013] se sont penchés sur ce problème et ont proposé une stratégie qui traite les mots polysémiques dans les vecteurs de contexte par l'introduction d'un processus de désambiguïsation sémantique basé sur WordNet.

Dans la quatrième et dernière expérience, nous avons pu observer l'impact du changement de taille des vecteurs de contexte sur l'alignement bilingue. Bien que les résultats aient montré une tendance à choisir une taille de vecteur allant de 500 à

1 000, ce résultat reste empirique et spécifique à nos expériences. Nous ne pouvons en aucun cas en déduire une paramétrisation optimale. Par ailleurs, sachant que la taille du vecteur de contexte dépend en partie de la taille de la fenêtre contextuelle dans le cas d'une représentation graphique, utiliser un seuil pour filtrer les mots ayant un score d'association faible ou non significatif pourrait constituer une alternative. Là encore, se pose le problème du choix du seuil adéquat. Enfin, nous pouvons considérer les différents paramètres de l'*approche directe* comme un problème d'optimisation à plusieurs paramètres, mais partir dans cette direction voudrait dire occulter la partie inexorable qu'est la linguistique. Un compromis entre les deux serait sans doute une démarche plus pertinente.

Finalement, un modèle distributionnel construit à partir d'un corpus comparable induit intrinsèquement une dépendance du modèle à ce corpus. De ce fait, un corpus de moins bonne qualité devrait produire un modèle de moins bonne qualité. Nous avons pu remarquer que les résultats obtenus sur le corpus de vulcanologie étaient supérieurs à ceux des corpus du cancer du sein et des énergies renouvelables. La meilleure qualité du corpus de vulcanologie n'est sans doute pas étrangère à ces meilleurs résultats. S'appuyer sur un corpus de bonne qualité est aussi nécessaire voire plus importante que les méthodes d'extraction en elles mêmes. Dans ce cadre, nous pouvons citer les travaux de [Li et Gaussier, 2010] qui s'intéressent à l'amélioration de la qualité des corpus comparables pour améliorer la qualité des lexiques extraits.

4.4 Bilan

Nous avons présenté dans ce chapitre une étude détaillée de l'*approche directe* ainsi que plusieurs résultats expérimentaux sur trois corpus de spécialité. La méthodologie proposée repose principalement sur une analyse distributionnelle qui ne fait pas ou peu intervenir une analyse linguistique sans doute fondamentale. Si les résultats présentés dans cette étude ont permis de dégager certaines caractéristiques de l'*approche directe* et certaines spécificités liées aux dictionnaires, aux listes d'évaluation et aux corpus comparables, des efforts restent à faire et une analyse linguistique serait sans doute nécessaire pour comprendre les raisons de l'échec de l'*approche directe* quant à la traduction de certains termes.

IV

Contributions à l'extraction terminologique bilingue à partir de corpus comparables

5

Combinaison de contextes

Introduction

La plupart des travaux utilisant les corpus comparables en extraction de lexiques bilingues ont comme base commune le contexte. Celui-ci représente le cœur de l'extraction lexicale bilingue. Selon l'état de l'art, le contexte d'un mot w est habituellement représenté par les mots faisant partie de son environnement, c'est-à-dire les mots qui l'entourent. Ces mots sont extraits soit à l'aide d'une fenêtre contextuelle [Rapp, 1999, Déjean et Gaussier, 2002] (représentation graphique), soit à l'aide de relations de dépendance syntaxique [Gamallo, 2008a] (représentation syntaxique). Chacun de ces mots se voit attribuer un score qui représente son degré d'association par rapport à w. Le contexte d'un mot w est ainsi défini par : (i) la manière de choisir les mots avec lesquels il cooccurre, et (ii) la manière de lui attribuer un score d'association. Nous nous intéressons dans ce qui suit à chacun de ces points.

L'un des problèmes sous-jacents au contexte extrait à l'aide de fenêtres contextuelles est le choix de leur taille. Celle-ci est habituellement fixée empiriquement, et bien que différentes études aient montré une tendance à choisir des fenêtres de petite taille quand il s'agit de caractériser des mots fréquents, et des fenêtres de grande taille quand il s'agit de caractériser des mots peu fréquents [Prochasson et Morin, 2009], cela reste empirique car il n'y a pas à notre connaissance de méthode dite optimale pour le choix de la taille de la fenêtre contextuelle. Quant aux relations de dépendance syntaxique, leur efficacité est très sensible à la taille des corpus et à la qualité des relations de dépendance. Bien que cette représentation soit plus intéressante d'un point de vue sémantique, elle atteint ses limites lorsqu'il s'agit de traiter des corpus de petite taille (comme l'ont montré les résultats des expériences du chapitre 4). S'agissant des scores d'association, il en existe une large palette, chacune spécifique à une ou plusieurs tâches et avec ses points forts et ses points faibles. Dans la suite de notre étude, nous continuons à tester les trois mesures d'association retenues pour l'*approche directe*, à savoir l'*information mutuelle*, le *discounted odds-ratio* ainsi que le *taux de vraisemblance* [Morin, 2009, Laroche et Langlais, 2010, Gamallo, 2008a].

Nous nous positionnons pour cette première contribution dans le cadre de l'amélioration de l'*approche directe* décrite dans plusieurs travaux dont [Fung, 1998, Rapp, 1999, Morin, 2009, Laroche et Langlais, 2010, Gamallo, 2008a], etc. Notre démarche vise à montrer que l'exploitation conjointe des deux principales représentations contextuelles a un intérêt particulier pour la tâche de construction de lexiques bilingues. Nous proposons donc deux manières de combiner les contextes (graphique et syntaxique), que nous appellerons la combinaison *a posteriori* des contextes et la combinaison *a priori* des contextes.

5.1 Combinaison de contextes

Une première manière de combiner les deux représentations contextuelles est une combinaison *a posteriori*, c'est-à-dire la combinaison des scores retournés pour chaque représentation de l'approche directe. Une seconde manière consiste à mettre en œuvre une combinaison *a priori* qui intègre les deux informations contextuelles dans le même vecteur pour ensuite appliquer l'*approche directe* sur l'ensemble du corpus.

5.1.1 Combinaison *a posteriori* des contextes

Dans le domaine de la recherche d'information, la combinaison de plusieurs listes renvoyées par différents moteurs de recherche est souvent utilisée pour améliorer les performances d'un système de question/réponse [Aslam et Montague, 2001]. Nous partons du principe que chaque représentation du contexte correspond à une approche bien définie. Nous nous retrouvons donc dans le cas d'une combinaison de deux approches bien distinctes. La première correspond à l'*approche directe* basée sur une représentation graphique et la seconde correspond à l'*approche directe* basée sur une représentation syntaxique. Une manière classique de fusionner ces deux approches est de prendre comme entrée, la sortie de chaque approche individuelle. Dans notre cas, pour chaque mot à traduire, nous prenons comme entrée une liste de scores retournée par chacune des deux approches, puis nous fusionnons les deux listes par une simple combinaison arithmétique des scores. Ceci nous donne une nouvelle liste de mots ordonnés (les scores fusionnés sont compatibles dans la mesure où nous utilisons la même mesure de similarité pour les deux approches). En utilisant les scores comme critère de fusion, nous calculons le score de similarité d'un candidat à la traduction comme la somme des scores renvoyés par chacune des deux approches comme suit :

$$\mathrm{S}_{comb}(w) = \mathrm{S}_{gra}(w) + \mathrm{S}_{syn}(w) \tag{5.1}$$

où $\mathrm{S}_{comb}(w)$ est le score final du mot w, $\mathrm{S}_{gra}(w)$ est le score retourné par l'*approche directe* basée sur une représentation graphique et $\mathrm{S}_{syn}(w)$ est le score retourné par l'*approche directe* basée sur une représentation syntaxique.

Cette équation peut aussi s'écrire comme suit :

$$\mathrm{S}_{comb}(w) = \lambda \times \mathrm{S}_{gra}(w) + (1 - \lambda) \times \mathrm{S}_{syn}(w) \tag{5.2}$$

avec λ comme indice de confiance donné à chaque méthode ($\lambda \in [0;1]$). Dans notre cas, $\lambda = 0,6$, notre but n'étant pas de trouver la valeur optimale de λ pour obtenir les meilleurs résultats. Différentes expériences que nous avons réalisé, indiquent que les meilleurs résultats sont globalement obtenus avec un *lambda* $\in [0,55; 0,65]$. Par ailleurs, d'autres méthodes de combinaisons de scores ont été testées comme la combinaison harmonique des rangs et des scores [Zweigenbaum et Habert, 2006, Morin, 2009], mais la méthode que nous avons choisi (combinaison arithmétique des scores) est celle qui donne les meilleures performances dans nos expériences.

5.1.2 Combinaison *a priori* des contextes

Le vecteur de contexte a pour but d'enregistrer un ensemble d'informations sur le contexte d'un mot w donné. Dans le cas de la représentation graphique, ces informations sont les mots qui cooccurrent avec le mot w. Dans le cas d'une représentation syntaxique, ce sont les mots en relations de dépendance syntaxique avec w qui sont sélectionnés pour faire partie de son vecteur de contexte. Dans un cadre plus générique, nous pourrions imaginer plusieurs autres sources d'informations à exploiter. Cependant, si chaque nouvelle information engendre un nouveau vecteur de contexte, nous pourrions vite être dépassés par le nombre de sources à fusionner. Pour remédier à cela, une autre manière serait de représenter dans un seul vecteur de contexte toutes les informations concernant le mot w. C'est la position adoptée avec la combinaison *a priori* des contextes.

Contexte graphique	**Contexte syntaxique**	**Combinaison**
$regional_{13}$	$regional_{Lmod_2}$	$regional_{13}$, $regional_{Lmod_2}$
$local_5$	$local_{Lmod_1}$	$local_5$, $local_{Lmod_1}$
$oestrogen_1$	-	$oestrogen_1$
$rate_{32}$	$rate_{modN_{29}}$,$rate_{PRPV_3}$	$rate_{32}$, $rate_{modN_{29}}$,$rate_{PRPV_3}$

TABLE 5.1 – Exemple de représentation du contexte du mot *recurrence* et du nombre de ses cooccurrences, en fonction des représentations graphique et syntaxique ainsi que de leur combinaison

Dans cette technique de combinaison, nous considérons le vecteur de contexte d'un mot comme un descripteur qui contient plusieurs informations pour chaque entrée du vecteur. Dans notre cas, nous avons deux types d'informations : (i) une information de cooccurrence globale fournie par la représentation graphique et (ii) une information plus spécifique fournie par la représentation syntaxique. Si nous prenons par exemple le mot *regional* (représenté dans le tableau 5.1), nous pouvons voir qu'il apparaît 13 fois avec le mot *recurrence* selon la représentation graphique et 2 fois comme modificateur gauche (Lmod) selon la représentation syntaxique. La combinaison prend en compte les deux informations, en considérant que le mot *regional* apparaît 13 fois avec *recurrence*, dont 2 fois en tant que modificateur gauche. Une information importante à souligner est que l'*approche directe* se basant sur les relations de dépendance syntaxique considère $rate_{modN_{29}}$ et $rate_{PRPV_3}$ par exemple, comme étant deux mots distincts. L'un des avantages de la combinaison *a priori* est que si l'une des méthodes manque une information (un mot), comme nous pouvons le constater avec le mot *oestrogen* par exemple, la fusion permet de pallier ce manque

(grâce ici à la représentation graphique). Nous considérons les deux représentations contextuelles comme étant complémentaires. Le but de la combinaison *a priori* est de préserver le classement et de renforcer les scores des entrées des vecteurs de contexte afin de lisser les contextes et corriger certaines erreurs qui peuvent apparaître.

Dans les tableaux 5.2, 5.3 et 5.4, nous illustrons les 10 premières entrées du vecteur de contexte du mot *recurrence* extrait du corpus du cancer du sein, en fonction de trois mesures d'association (TV, DOR et IM). La notation (+/−) indique l'apport positif ou négatif de la combinaison *a priori*. L'indice '+' indique qu'un mot classé dans les 10 premières entrées du vecteur de contexte de la représentation graphique ou syntaxique, conserve son classement dans les 10 premières entrées après combinaison. Le signe '−' en revanche, indique l'apparition d'un mot non classé dans les 10 premières entrées du vecteur de contexte. Nous notons par $graphique_k$ la représentation graphique avec une fenêtre de taille k et par *syntaxique* la représentation syntaxique.

graphique$_5$		**syntaxique**		**a_priori$_5$**		+/−
local	818,98	$local_{Lmod}$	618,17	$local_{Lmod}$	936,05	+
rate	119,71	$risk_{PRPN}$	96,02	local	791,15	+
distant	72,62	$rate_{modN}$	68,34	$risk_{PRPN}$	153,14	+
risk	61,00	$tumor_{modN}$	62,82	rate	113,96	+
salvage	39,15	$rate_{PRPN}$	40,18	$rate_{modN}$	110,28	+
year	39,08	$time_{PRPN}$	32,85	$tumor_{modN}$	104,71	+
time	31,84	$disease_{modN}$	28,76	distant	70,23	+
tumor	31,04	$isolated_{Lmod}$	24,29	$rate_{PRPN}$	64,69	+
isolate	30,15	$distant_{Lmod}$	24,28	risk	54,89	+
inoperable	28,16	$patient_{PRPN}$	23,64	$time_{PRPN}$	53,13	+

TABLE 5.2 – Illustration des 10 premières entrées du vecteur de contexte du mot *recurrence* en fonction du taux de vraisemblance pour les représentations graphique et syntaxique ainsi que par la combinaison *a priori*

Le tableau 5.2 montre que la combinaison *a priori* a un apport positif car elle engendre un vecteur de contexte qui respecte le classement des méthodes $graphique_5$ et *syntaxique* en utilisant TV. Le tableau 5.3 indique aussi que la combinaison *a priori* a un apport positif en utilisant DOR. Nous remarquons néanmoins que la combinaison a avantagé la méthode *syntaxique*, car il n'y a que ses entrées qui sont présentes dans les 10 premières entrées du vecteur de contexte de la méthode de combinaison *a priori*. Enfin, le tableau 5.4 montre que la combinaison *a priori* a un apport négatif pour au moins 5 mots. Ces mots qui n'étaient pas classés dans les 10 premières entrées des méthodes $graphique_5$ et *syntaxique* le sont maintenant avec la combinaison *a priori*. Il semble donc que IM n'est pas appropriée dans ce cas, car elle ne préserve pas le classement des entrées de $graphique_5$ et *syntaxique*. Elle affecte des scores élevés à des mots qui avaient des scores faibles comme pour *rate* ou *cancer* par exemple, qui passent respectivement de $5,59$ à $14,39$ et de $2,28$ à $14,04$.

graphique$_5$		**syntaxique**		**a_priori$_5$**		+/−
isolated	5,10	$freedom_{PRPN}$	7,83	$freedom_{PRPN}$	8,12	+
geographic	4,62	$heat_{Robj}$	6,72	fat_{PRPN}	7,02	+
adjudication	4,44	$operable_{Rmod}$	6,72	$threat_{PRPN}$	7,02	+
conspicuous	4,44	fat_{PRPN}	6,72	$operable_{Rmod}$	7,02	+
liberate	4,44	$threat_{PRPN}$	6,72	$heat_{Robj}$	7,02	+
evade	4,44	$local_{Lmod}$	5,89	$local_{Lmod}$	6,02	+
inoperable	4,38	$fear_{PRPN}$	5,63	$fear_{PRPN}$	5,93	+
quarter	4,29	$suspicion_{PRPN}$	5,63	$suspicion_{PRPN}$	5,93	+
local	4,28	$inoperable_{Lmod}$	5,63	$inoperable_{Lmod}$	5,93	+

TABLE 5.3 – Illustration des 10 premières entrées du vecteur de contexte du mot *recurrence* en fonction du *discounted odds-ratio* pour les représentations graphique et syntaxique ainsi que par la combinaison *a priori*

graphique$_5$		**syntaxique**		**a_priori$_5$**		+/−
isolated	8,73	$local_{Lmod}$	14,77	local	16,17	+
geographic	8,15	$tumor_{modN}$	13,84	$local_{Lmod}$	15,83	+
inoperable	8,00	$risk_{PRPN}$	12,84	breast	14,64	−
local	7,82	$time_{PRPN}$	12,44	rate	14,39	−
adjudication	7,73	$distant_{Lmod}$	12,09	tumor	14,15	−
conspicuous	7,73	$rate_{modN}$	11,91	cancer	14,04	−
reconcile	7,73	$year_{modN}$	11,80	$risk_{PRPN}$	13,90	+
liberate	7,73	$rate_{PRPN}$	11,63	patient	13,75	−
quarter	7,73	$tumour_{modN}$	11,63	$cancer_{modN}$	13,15	−
	⋮		⋮		⋮	⋮
rate	5,59	$cancer_{modN}$	10,51			
survival	4,12		⋮			
tumor	3,69					
patient	3,21					
breast	2,92					
cancer	2,28					

TABLE 5.4 – Illustration des 10 premières entrées du vecteur de contexte du mot *recurrence* en fonction de l'information mutuelle pour les représentations graphique et syntaxique ainsi que par la combinaison *a priori*

5.2 Évaluation

Nous évaluons ici i) l'*approche directe* basée sur une représentation graphique notée $graphique_k$ où k qui prend les valeurs 5, 7, 9, 11 et 15 correspond à la taille de la fenêtre contextuelle, ii) l'*approche directe* basée sur une représentation syntaxique notée *syntaxique*, iii) la première nouvelle approche notée $a\_posteriori_k$ qui combine donc les scores de $graphique_k$ et de *syntaxique a posteriori* et iv) la seconde nouvelle approche notée a_priori_k qui exploite les contextes fournis par une fenêtre contextuelle $graphique_k$ et les relations de dépendances *syntaxique* dans un même vecteur pour ensuite appliquer l'*approche directe*. L'évaluation est réalisée sur les 3 couples de mesures les plus utilisées dans l'état de l'art : TV-JAC [Morin, 2009], DOR-COS [Laroche et Langlais, 2010] et IM-COS [Gamallo, 2008a].

5.2.1 Résultats en MAP (%) des approches *graphique*, *syntaxique*, *a priori* et *a posteriori*

Le tableau 5.5 montre les résultats des expériences pour les approches *graphique*, *syntaxique*, *a priori* et *a posteriori* sur les 3 corpus comparables et pour les 3 combinaisons de mesures d'association et de similarité. Pour la configuration TV-JAC, l'approche par combinaison *a priori* des contextes est supérieure à l'*approche directe* de base ($graphique_k$) et ce pour les 3 corpus de spécialités. Les meilleurs résultats sont obtenus en combinant *syntaxique* avec des fenêtres de taille 5, 9 et 11. Concernant la configuration IM-COS, la combinaison *a priori* n'est pas efficace et dégrade même les résultats dans certains cas. Les résultats obtenus par la combinaison *a priori* associée à la configuration DOR-COS suivent le même comportement que les résultats de la configuration TV-JAC avec une nette amélioration des résultats. En ce qui concerne la combinaison *a posteriori*, le tableau 5.5 montre une nette amélioration des résultats pour les trois configurations.

5.2.2 Résultats en précision (%) des approches *graphique*, *syntaxique*, *a priori* et *a posteriori*

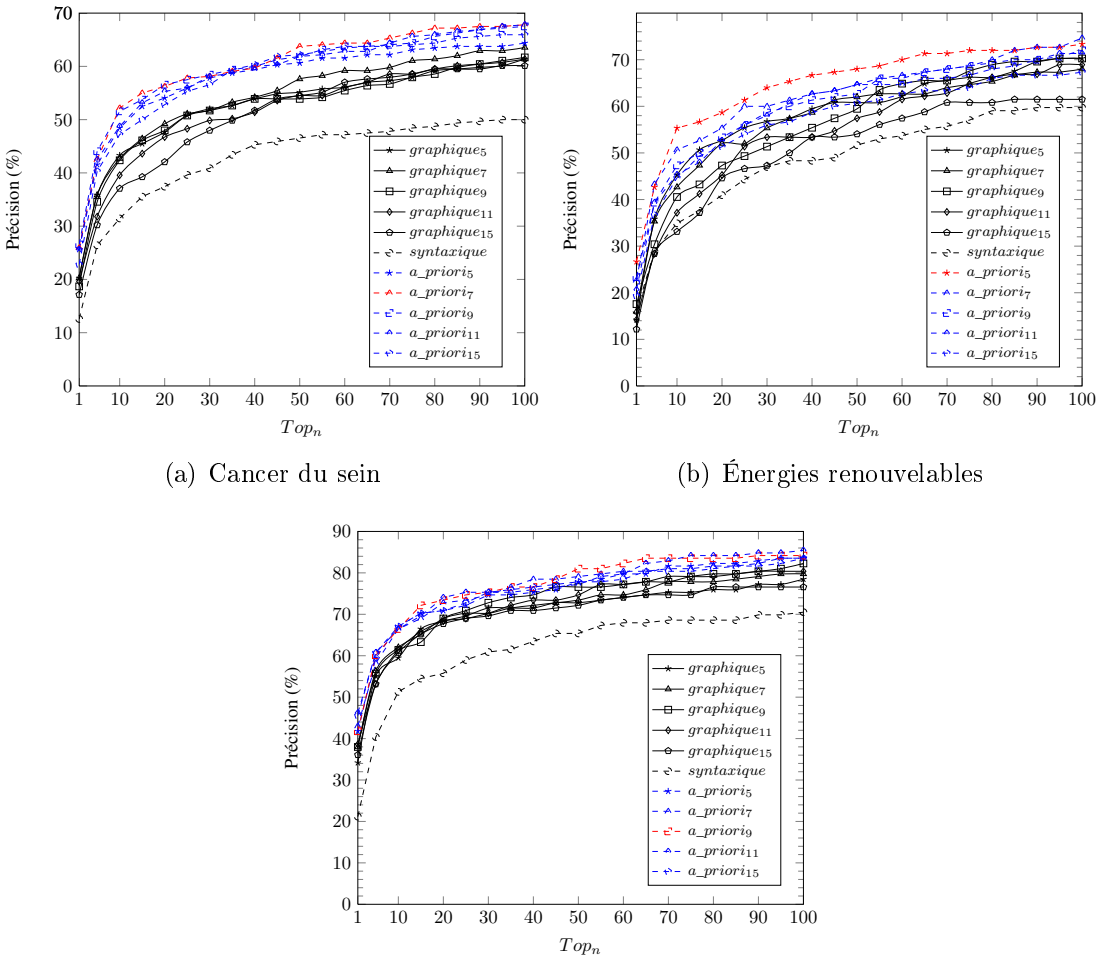
Nous présentons ici les résultats des différentes approches en fonction du top n. Les figures 5.1, 5.2 et 5.3 illustrent les résultats de l'*approche directe* en utilisant les représentations graphique, syntaxique ainsi que la combinaison *a priori* des contextes pour les configurations TV-JAC, IM-COS et DOR-COS.

Les figures 5.4, 5.5 et 5.6 illustrent les résultats de l'*approche directe* en utilisant les représentations graphique, syntaxique ainsi que la combinaison *a posteriori* des contextes pour les mêmes configurations.

La figure 5.7 montre les résultats des meilleures configurations des approches *a priori* et *a posteriori*. D'une manière globale, les deux méthodes de combinaison proposées obtiennent de meilleurs résultats que les deux représentations contextuelles prises séparément, avec un léger avantage pour la méthode *a priori*.

	IM-COS	DOR-COS	TV-JAC	
$graphique_5$	25,3	24,5	27,7	**Cancer du sein**
$graphique_7$	22,6	24,8	27,9	
$graphique_9$	17,3	23,9	26,6	
$graphique_{11}$	15,1	22,9	26,2	
$graphique_{15}$	11,1	20,4	23,7	
$syntaxique$	18,9	14,6	19,2	
a_priori_5	20,0	31,5	33,3	
a_priori_7	17,9	32,8	34,2	
a_priori_9	13,1	33,0	**34,5**	
a_priori_{11}	11,7	**34,0**	33,5	
a_priori_{15}	09,8	31,8	31,3	
$a\_posteriori_5$	**30,3**	31,8	32,0	
$a\_posteriori_7$	29,4	32,0	32,8	
$a\_posteriori_9$	27,2	33,5	33,4	
$a\_posteriori_{11}$	25,9	32,6	33,9	
$a\_posteriori_{15}$	23,2	32,5	32,0	
$graphique_5$	19,9	18,6	23,9	**Énergies renouvelables**
$graphique_7$	15,6	20,2	24,2	
$graphique_9$	12,5	18,2	24,3	
$graphique_{11}$	11,6	17,7	21,9	
$graphique_{15}$	09,1	14,9	20,0	
$syntaxique$	16,5	13,7	23,0	
a_priori_5	14,9	27,8	**34,1**	
a_priori_7	13,8	29,6	32,2	
a_priori_9	12,1	28,9	31,4	
a_priori_{11}	10,5	28,5	29,8	
a_priori_{15}	09,9	28,4	27,4	
$a\_posteriori_5$	**30,0**	29,1	30,6	
$a\_posteriori_7$	26,4	31,3	31,0	
$a\_posteriori_9$	23,7	30,3	32,2	
$a\_posteriori_{11}$	21,2	31,5	29,5	
$a\_posteriori_{15}$	19,4	**31,6**	28,7	
$graphique_5$	29,3	33,3	43,5	**Vulcanologie**
$graphique_7$	21,7	30,3	46,8	
$graphique_9$	20,5	34,7	46,1	
$graphique_{11}$	17,7	33,0	45,8	
$graphique_{15}$	13,3	28,3	44,2	
$syntaxique$	23,3	18,4	30,2	
a_priori_5	20,6	44,9	49,8	
a_priori_7	18,3	44,8	50,6	
a_priori_9	15,2	48,1	50,5	
a_priori_{11}	14,5	**50,2**	**53,2**	
a_priori_{15}	13,1	49,2	50,1	
$a\_posteriori_5$	**43,3**	45,7	49,1	
$a\_posteriori_7$	39,6	45,9	51,0	
$a\_posteriori_9$	37,4	48,4	51,9	
$a\_posteriori_{11}$	35,1	48,2	52,6	
$a\_posteriori_{15}$	33,1	49,0	52,5	

TABLE 5.5 – Résultats en MAP (%) des expériences pour les approches *graphique*, *syntaxique*,*a priori* et *a posteriori*

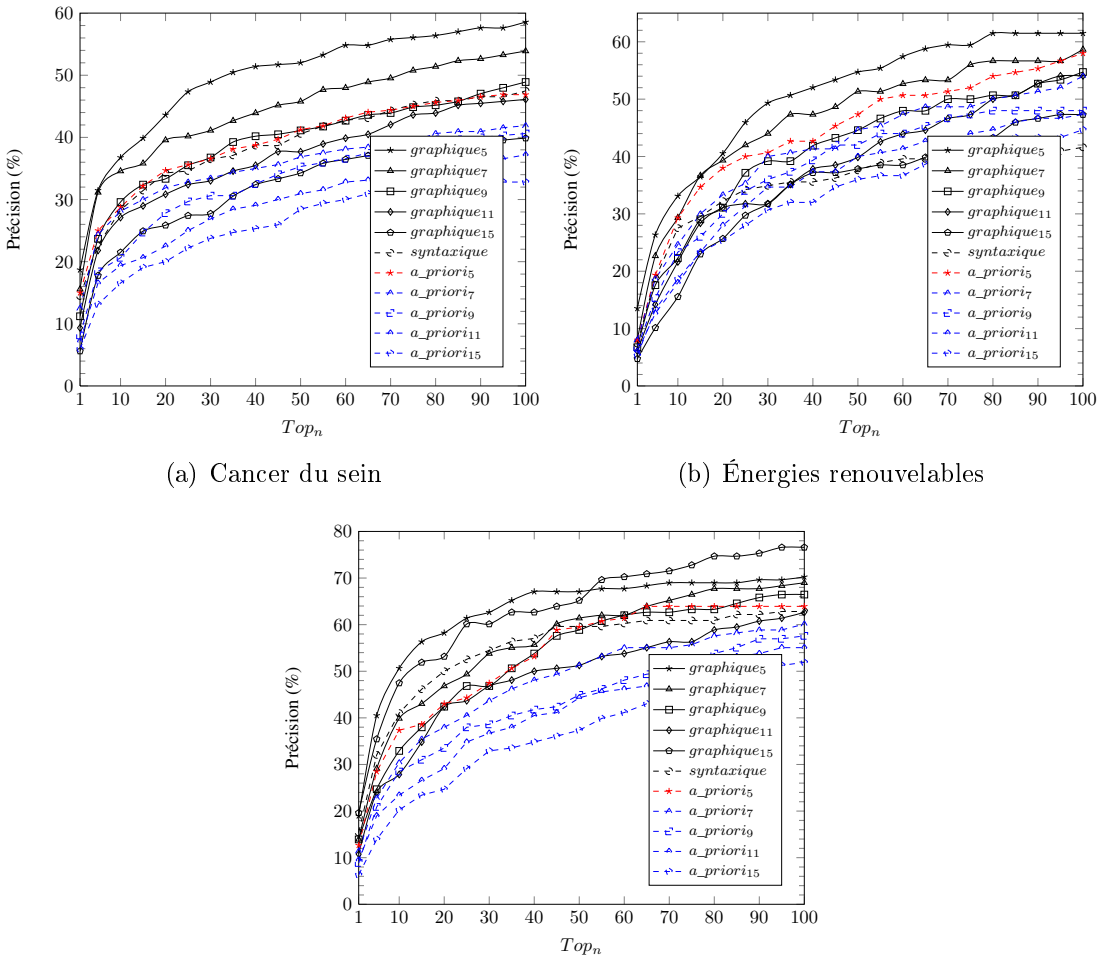


(a) Cancer du sein

(b) Énergies renouvelables

(c) Vulcanologie

FIGURE 5.1 – Comparaison de l'approche directe en utilisant les représentations graphique, syntaxique ainsi que la combinaison *a priori* des contextes pour la configuration TV-JAC

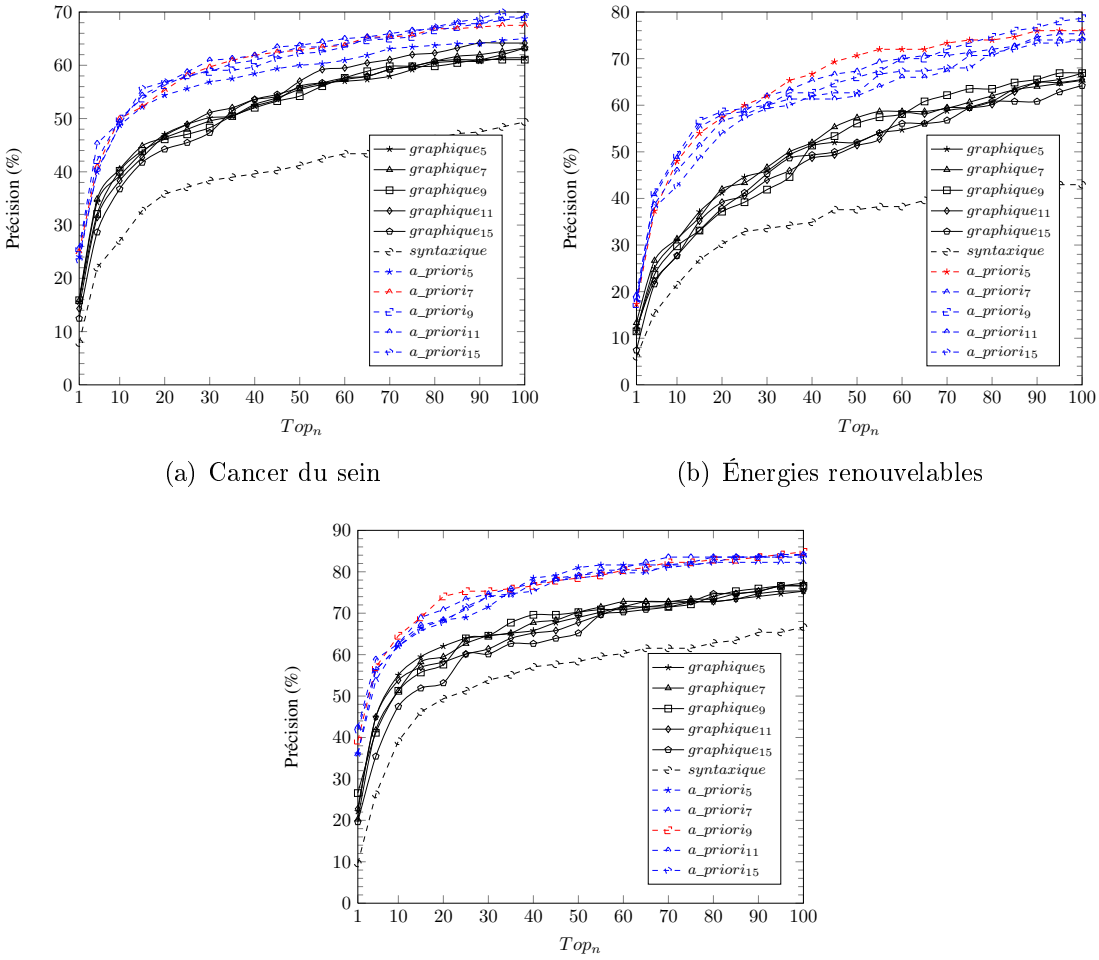


(a) Cancer du sein

(b) Énergies renouvelables

(c) Vulcanologie

FIGURE 5.2 – Comparaison de l'approche directe en utilisant les représentations graphique, syntaxique ainsi que la combinaison *a priori* des contextes pour la configuration IM-COS

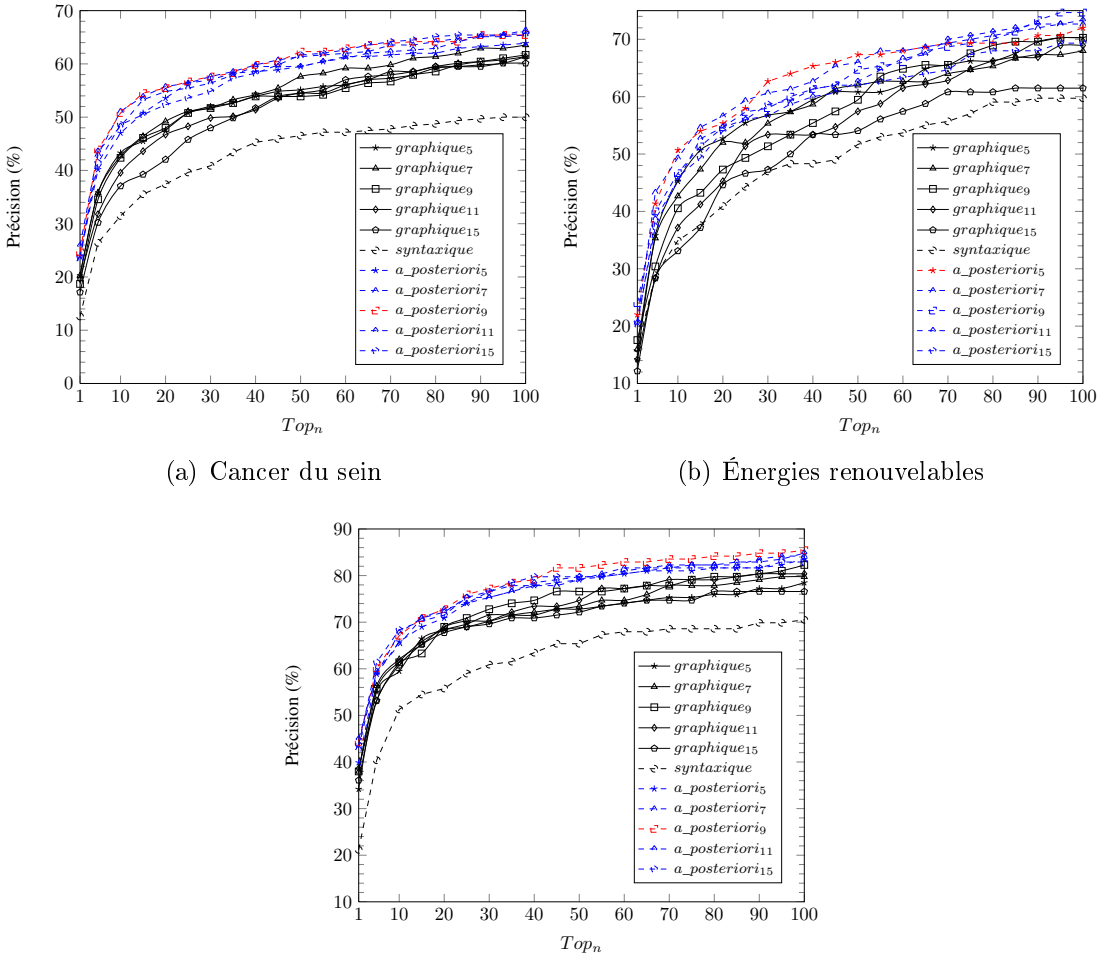


(a) Cancer du sein

(b) Énergies renouvelables

(c) Vulcanologie

FIGURE 5.3 – Comparaison de l'approche directe en utilisant les représentations graphique, syntaxique ainsi que la combinaison *a priori* des contextes pour la configuration DOR-COS

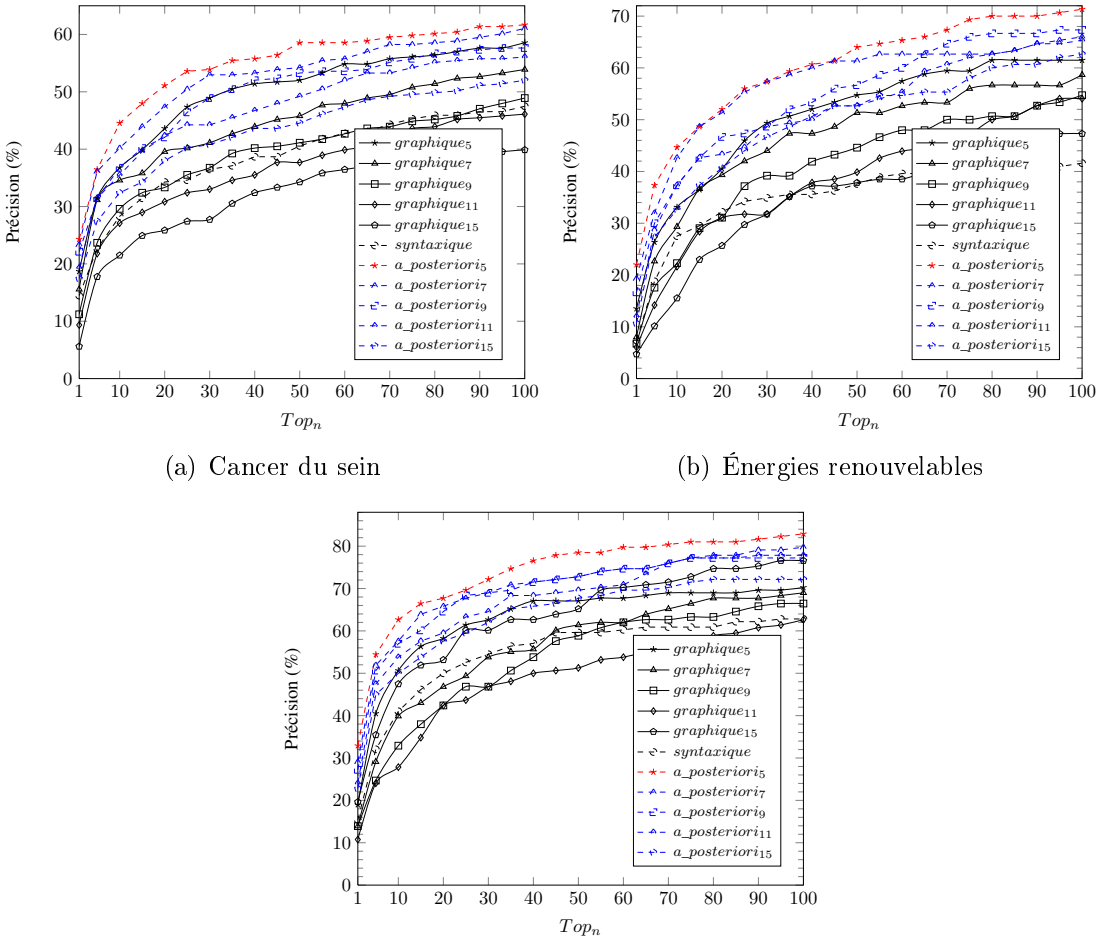


(a) Cancer du sein

(b) Énergies renouvelables

(c) Vulcanologie

FIGURE 5.4 – Comparaison de l'approche directe en utilisant les représentations graphique, syntaxique ainsi que la combinaison *a posteriori* des contextes pour la configuration TV-JAC

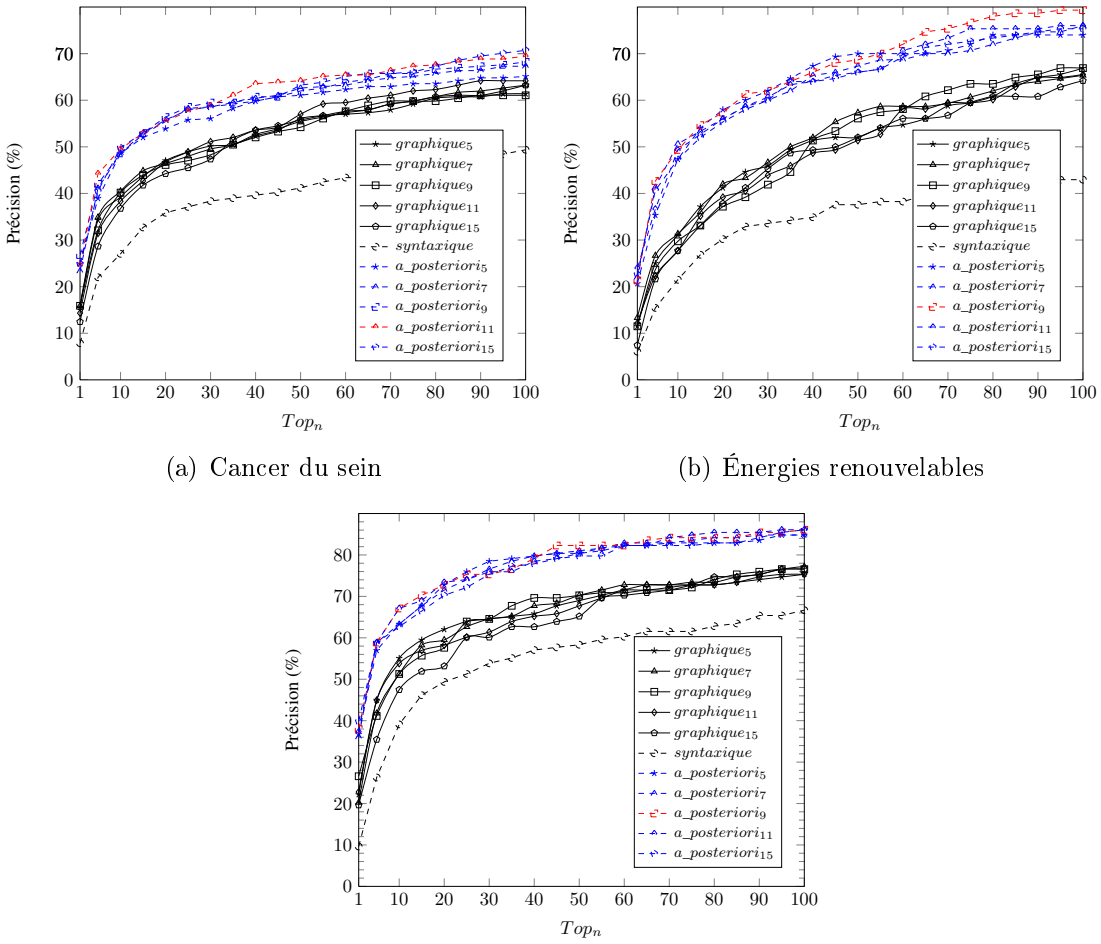


(a) Cancer du sein

(b) Énergies renouvelables

(c) Vulcanologie

FIGURE 5.5 – Comparaison de l'approche directe en utilisant les représentations graphique, syntaxique ainsi que la combinaison *a posteriori* des contextes pour la configuration IM-COS

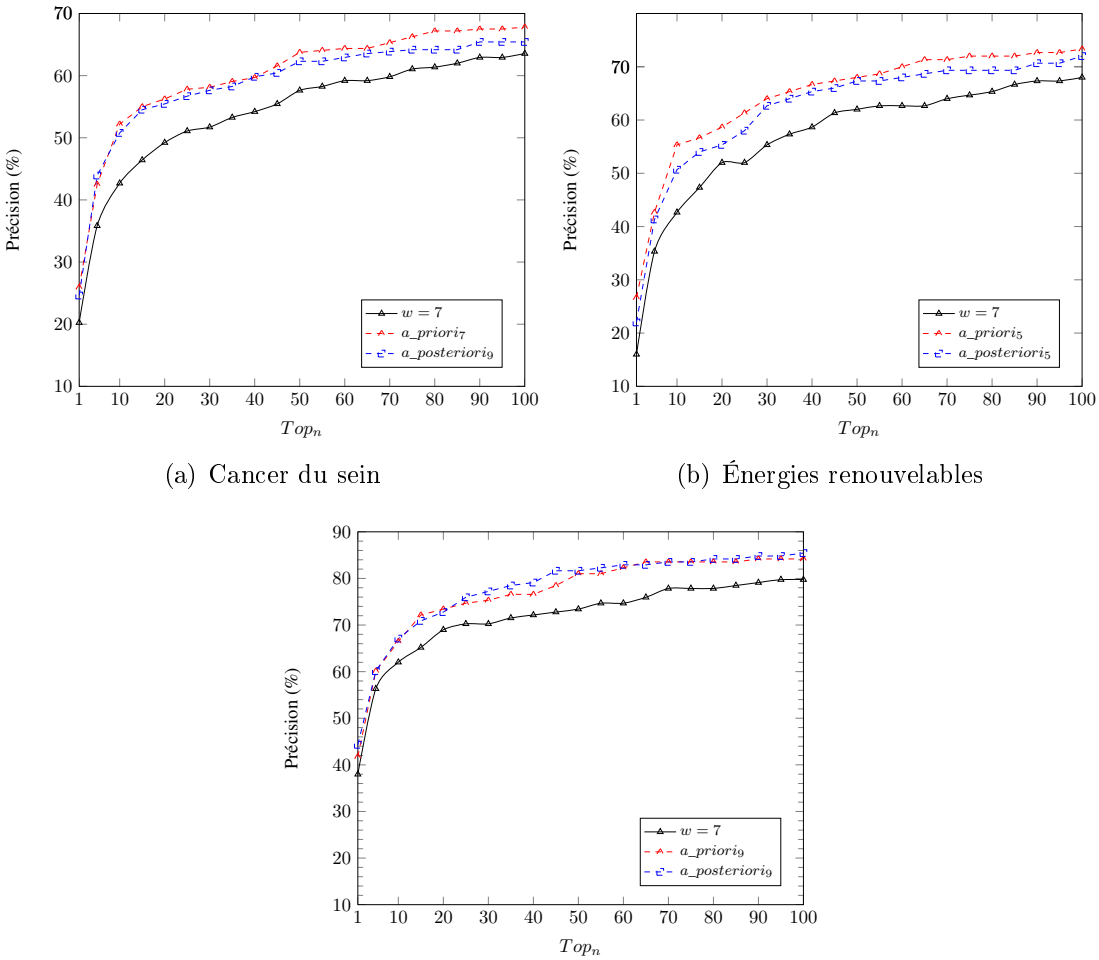


(a) Cancer du sein

(b) Énergies renouvelables

(c) Vulcanologie

FIGURE 5.6 – Comparaison de l'approche directe en utilisant les représentations graphique, syntaxique ainsi que la combinaison *a posteriori* des contextes pour la configuration DOR-COS



(a) Cancer du sein

(b) Énergies renouvelables

(c) Vulcanologie

FIGURE 5.7 – Comparaison de l'approche *a posteriori* et de l'approche *a priori* en fonction de leur meilleure configuration (TV-JAC)

5.3 Discussion

Dans ce travail, nous cherchions dans un premier temps à comparer les deux principales représentations contextuelles utilisées avec l'approche directe, puis dans un second temps à proposer deux nouvelles manières de les combiner pour en augmenter les performances. La première remarque concerne l'utilisation de la représentation graphique ($graphique_k$). Il est évident que le choix de la taille de la fenêtre joue un rôle important, comme nous avons pu le constater dans les différentes expériences. Dans la plupart des cas, ce sont des fenêtres de taille 5, 7 et 9 qui donnent les meilleurs résultats. Ceci montre que la caractérisation du contexte des mots par ceux qui leur sont très proches semble être la manière la plus adéquate, si l'on se base sur une caractérisation par fenêtre contextuelle. Le fait de choisir des fenêtres de taille plus grande n'améliore pas significativement les résultats de nos expériences.

La deuxième remarque concerne la méthode par représentation syntaxique. Cette méthode mise en œuvre par Gamallo (2008a) donne dans ses expériences de meilleurs résultats que la méthode par représentation graphique. Dans nos expériences, la méthode *syntaxique* reste globalement en deçà de $graphique_k$. Ceci s'explique par trois facteurs. Le premier concerne la taille des corpus. Gamallo (2008a) avait utilisé des corpus de très grande taille (10 millions de mots environs) contrairement à nos corpus spécialisés qui sont de petite taille (600 000, 800 000 et 1 million de mots). Le deuxième facteur, qui est directement lié au premier, concerne la manière de considérer les entrées des vecteurs de contexte de la méthode *syn*. Si dans le vecteur de contexte d'un mot w_i, il existe un mot w_j avec une relation Lmod de w_i ayant un score $S_{w_i^{Lmod}}$ et une autre relation Robj avec un score $S_{w_j^{Robj}}$, alors dans ce vecteur de contexte w_j^{Lmod} et w_j^{Robj} sont considérés comme étant deux mots différents, bien que ce soit le même mot avec deux relations de dépendance distinctes, cela rend la méthode *syntaxique* plus sensible aux petits corpus que $graphique_k$. Ceci explique les performances de la méthode de combinaison *a priori* des contextes. En effet, la méthode a_priori_k comble le manque de la méthode *syntaxique*, car elle considère les deux informations véhiculées par les deux représentations contextuelles. Ainsi, le fait d'exploiter une fenêtre de taille k va permettre d'avoir une information sur le nombre de fois qu'un mot apparaît dans le contexte d'un autre et, comme deuxième information plus fine, la nature des relations qui existent entre les deux mots. Le troisième facteur qui peut en partie expliquer les résultats de l'approche *syntaxique* est le risque d'erreurs engendrées par l'analyseur syntaxique utilisé.

Par ailleurs, nous avons pu constater que la méthode *a priori* était plus sensible aux modifications des mesures d'association et de similarité que la méthode *a posteriori*. Cela s'explique par le fait que la combinaison *a posteriori* agit sur les scores alors que méthode *a priori* agit directement sur le contenu des vecteurs de contexte. L'échec de la combinaison IM-COS peut s'expliquer par la tendance de la mesure de l'information mutuelle à surestimer les cooccurrences de faible fréquence et à sous-estimer les cooccurrences de haute fréquence. L'application de la combinaison *a priori* sur les mots de faible fréquence aura tendance à accentuer l'effet négatif de l'information mutuelle.

5.4 Bilan

Nous nous sommes intéressés dans ce chapitre aux deux principales manières de représenter le contexte des mots, à savoir la représentation graphique et la représentation syntaxique. Nous avons ensuite introduit deux nouvelles techniques de combinaison de ces représentations. Les deux approches de combinaisons contextuelles proposées ont montré des résultats supérieurs à l'utilisation de chaque représentation séparément, pour la plupart des paramètres de configuration. Nous espérons que ce travail ouvrira la voie à une recherche plus approfondie concernant l'enrichissement du contenu des vecteurs de contexte par des informations multiples sur les mots les composants. Si les travaux du présent chapitre se sont limités à deux types d'informations contextuelles, d'autres informations sont envisageables comme l'utilisation de thesaurus ou d'autres informations comme les cognats, les translittérations, les collocations, etc. Nous pouvons à termes, envisager qu'un mot soit étiqueté par plusieurs traits permettant ainsi une caractérisation plus fines des mots.

6

Ré-estimation des cooccurrences

Introduction

Partant de l'hypothèse que les cooccurrences des mots ne sont pas fiables, surtout pour des corpus de petite taille [Zipf, 1949, Evert et Baroni, 2007] et sachant que les méthodes de ré-estimation visent à pallier ce problème, nous proposons une extension de l'*approche directe* en introduisant une étape intermédiaire qui consiste à ré-estimer les cooccurrences des mots observés, soit par des techniques de lissage soit par des techniques de prédiction des cooccurrences. Chaque valeur de cooccurrence (notée $cooc(w_i, w_j)$) est ré-estimée en fonction d'une des méthodes présentées dans les sections 6.1 et 6.2. La nouvelle estimation (notée $cooc^*(w_i, w_j)$) est alors utilisée pour calculer l'association entre w_i et w_j. Il est à noter que nous nous sommes limités à l'estimation des mots observés dans le corpus. Concernant les mots inconnus, Pekar *et al.* (2006) ont montré l'efficacité des méthodes de lissage pour l'alignement des termes peu fréquents.

Dans ce travail, la ré-estimation correspond à deux aspects différents. Un premier aspect porte sur des techniques de lissage qui permettent une redistribution des poids des mots selon certains critères. Ainsi plusieurs techniques de lissage de l'état de l'art sont abordées. Le deuxième aspect est motivé par le manque de corpus de grande taille en domaine de spécialité. L'idée consiste à se projeter dans un corpus de grande taille sans en avoir un à disposition, et ceci partant de l'hypothèse que si deux mots cooccurrent n fois dans un petit corpus, et que ces mots sont fortement liés alors ils apparaîtront ensemble $n \times k$ fois dans un corpus de plus grande taille. Ceci permet d'asseoir plus précisément leur association. Dans ce cas, nous cherchons une fonction qui permet de prédire les cooccurrences des mots dans un grand corpus partant des observations faites sur un corpus de taille plus modeste.

6.1 Ré-estimation par méthodes de lissage

Les méthodes de lissage ont montré leur efficacité dans plusieurs domaines et notamment dans la traduction automatique. Nous présentons les différentes méthodes

exploitées, à savoir la méthode de Laplace (Add-One), la méthode Good-Turing, l'estimation par interpolation linéaire de Jelinek-Mercer, la méthode de Katz Back-off et celle de Kneser-Ney.

6.1.1 Laplace (ou estimation Add-One)

La technique de Laplace [Lidstone, 1920, Johnson, 1932, Jeffreys, 1948] estime la probabilité P en supposant que chaque type de mot absent ou non vu apparaît une fois. Dans ce cas, si nous avons N événements et V mots possibles alors, la probabilité du mot w est :

$$P(w) = \frac{occ(w)}{N} \tag{6.1}$$

où $occ(w)$ est le nombre d'occurrences de w.

L'estimation de P devient alors :

$$P_{addone}(w) = \frac{occ(w) + 1}{N + V} \tag{6.2}$$

L'utilisation de l'estimation de Laplace pour les cooccurrences des mots suppose que si deux mots cooccurrent n fois, alors ils peuvent cooccurrer $n + 1$ fois. Selon l'estimation du maximum de vraisemblance :

$$P(w_{i+1}|w_i) = \frac{cooc(w_i, w_{i+1})}{occ(w_i)} \tag{6.3}$$

En utilisant l'estimation de Laplace nous obtenons :

$$cooc^*(w_i, w_j) = \frac{cooc(w_i, w_j) + 1}{occ(w_i) + V} \tag{6.4}$$

Nous pouvons constater que l'estimation de Laplace, qui est une technique très simple, coïncide avec notre objectif d'augmenter la valeur de cooccurrence des mots observés, et ceci, dans le but de renforcer leurs relations. Ceci étant dit, l'estimateur de Laplace comporte plusieurs désavantages :

- la probabilité des n-grammes fréquents est sous-estimée ;
- la probabilité des n-grammes rares ou absents est sur-estimée ;
- tous les n-grammes absents sont lissés de la même manière ;
- une trop grande masse de probabilité est réservée aux n-grammes absents.

Une amélioration éventuelle serait d'ajouter une valeur inférieure à 1, selon l'équation suivante :

$$cooc^*(w_i, w_j) = \frac{cooc(w_i, w_j) + \delta}{occ(w_i) + \delta \times V} \tag{6.5}$$

avec $\delta \in [0; 1]$.

6.1.2 Estimateur de Good-Turing

L'estimateur de Good-Turing [Good, 1953] fournit une autre manière de lisser les probabilités. Il stipule que pour chaque n-gramme apparaissant r fois, nous pouvons prétendre qu'il apparaît r^* fois. Cette hypothèse converge avec notre idée de prédire les cooccurrences des mots dans un corpus de grande taille en partant des observations faites à partir d'un corpus de petite taille. L'estimateur de Good-Turing utilise les comptes de ce que l'on observe une fois pour estimer les comptes de ce que l'on n'a jamais observé. Pour estimer les fréquences de cooccurrence des mots, nous aurons besoin de calculer N_c qui est le nombre d'événements observés c fois (ceci suppose que tous les événements suivent une loi binomiale). Soit N_r le nombre de n-grammes qui occurrent r fois. N_r peut être utilisé pour fournir une meilleure estimation de r. Étant donnée une distribution binomiale, la fréquence estimée devient alors :

$$r^* = (r+1)\frac{N_{r+1}}{N_r} \tag{6.6}$$

Notre adaptation de l'estimateur de Good-Turing s'écrit comme suit :

$$cooc(w_i, w_j)^*_{GT} = (cooc(w_i, w_j) + 1)\frac{N_{cooc(w_i,w_j)+1}}{N_{cooc(w_i,w_j)}} \tag{6.7}$$

6.1.3 Estimation par interpolation linéaire

Comme alternative aux n-grammes absents, Mercer (1980) proposent d'utiliser l'interpolation linéaire avec le modèle de Good-Turing. Notre adaptation aux cooccurrences des mots est définie selon la formule qui suit :

$$cooc^*_{int}(w_i, w_j) = \lambda \times cooc^*(w_i, w_j) + (1-\lambda) \times occ^*(w_i) \tag{6.8}$$

avec λ qui correspond au facteur de confiance du n-gramme. Il est souvent utile d'interpoler les n-grammes de haut ordre avec les n-grammes d'ordre plus petit, car quand il y a un manque de n-grammes de haut ordre, les n-grammes d'ordre inférieur peuvent compléter et apporter une information non négligeable.

6.1.4 Katz Back-off

Katz (1987) a étendu l'intuition de l'estimateur de Good-Turing en combinant les modèles de plus grand ordre avec ceux de plus petit ordre. Notre adaptation aux cooccurrences des mots est donnée par l'équation suivante :

$$\mathrm{cooc}^*_{\mathrm{katz}}(\mathrm{w_i}, \mathrm{w_j}) = \begin{cases} \mathrm{r}^* & \text{si r} > 0 \\ \alpha(\mathrm{w_i})\mathrm{occ}^*(\mathrm{w_i}) & \text{si r} = 0 \end{cases} \tag{6.9}$$

et :

$$\alpha(\mathrm{w_i}) = \frac{1 - \sum_{\mathrm{w_i:cooc(w_i,w_j)>0}} \mathrm{cooc_{katz}(w_i, w_j)}}{1 - \sum_{\mathrm{w_i:cooc(w_i,w_j)>0}} \mathrm{occ(w_i)}} \tag{6.10}$$

Selon Katz (1987), l'estimation r^* de Good-Turing n'est pas utilisée pour toutes les valeurs de r. En effet, les très grandes valeurs de r sont supposées fiables à partir d'un certain seuil k. Katz (1987) suggère un $k = 5$. Ainsi, $r^* = r$ pour $r > k$ et :

$$\mathrm{r^*} = \frac{(\mathrm{r}+1)\frac{\mathrm{N_{r+1}}}{\mathrm{N_r}} - \mathrm{r}\frac{(\mathrm{k}+1)\mathrm{N_{k+1}}}{\mathrm{N_1}}}{1 - \frac{(\mathrm{k}+1)\mathrm{N_{k+1}}}{\mathrm{N_1}}} \tag{6.11}$$

pour $r \leq k$.

6.1.5 Kneser-Ney

Kneser et Ney (1995) ont proposé une extension de l'estimation par *absolute discounting*. La distribution de grand ordre est estimée en soustrayant une valeur fixe D de chaque valeur de la distribution. La différence avec la méthode *absolute discounting* standard réside dans la manière d'estimer la distribution de plus petit ordre, comme le montre l'équation suivante :

$$\mathrm{cooc^*_{kney}(w_i, w_j)} = \begin{cases} \frac{\mathrm{Max(cooc(w_i,w_j)-D,0)}}{\sum_{w_i} cooc(w_i,w_j)} & \text{si } \mathrm{cooc(w_i, w_j)} > 0 \\ \\ \alpha(\mathrm{w_i, w_j})\mathrm{occ^*(w_i)} & \text{si } \mathrm{cooc(w_i, w_j)} = 0 \end{cases} \tag{6.12}$$

où $\alpha(w_i, w_j)$ est choisi de sorte à ce que la somme de la distribution soit égale à 1 [Chen et Goodman, 1999].

6.2 Ré-estimation par prédiction

Nous partons de l'hypothèse que chaque couple de mots qui cooccurrent plus souvent que par chance, dans un petit corpus, devrait avoir le même comportement dans un corpus de plus grande taille avec une plus grande valeur de cooccurrence. Notre but est d'estimer cette nouvelle valeur. Gardons en tête la question qui motive cette démarche, à savoir étant donné l'ensemble des valeurs de cooccurrences observées, $E_p = \{o_1^p, o_2^p, ..., o_N^p\}$, pour les couples de mots d'un corpus de petite taille, quelles seront les valeurs attendues $E_g = \{o_1^g, o_2^g, ..., o_N^g\}$ de cet ensemble dans un corpus de grande taille ?

Nous proposons, dans ce qui suit, plusieurs méthodes pour estimer l'ensemble E_g :

- une première technique basée sur l'augmentation moyenne des cooccurrences des mots pour chaque valeur de cooccurrence ;

- une deuxième technique basée sur un modèle de régression linéaire simple ;
- deux techniques basées sur le maximum et la moyenne des valeurs observées sur un corpus d'apprentissage de grande taille.

6.2.1 Prédiction par la moyenne

Nous considérons l'utilisation de la moyenne comme une solution intuitive ayant pour principal objectif de fournir une information sur la tendance moyenne de l'augmentation des cooccurrences. Pour estimer les valeurs de l'ensemble E_g, nous utilisons un corpus d'apprentissage divisé en deux échantillons. Le premier échantillon correspond au corpus journalistique de petite taille (500 000 mots), et le second échantillon correspond au corpus journalistique de grande taille (10 millions de mots). Ainsi, nous calculons l'augmentation moyenne observée dans le corpus de grande taille pour tous les couples de mots ayant une valeur de cooccurrence o_1. Nous répétons cette procédure pour chaque valeur de cooccurrence, jusqu'à obtenir un vecteur d'augmentation pour toutes les valeurs de cooccurrences.

Soit $E_p^1 = \{cooc_1^p(w_i, w_j) = 1, i \in [1; N], j \in [1; N]\}$ l'ensemble des couples de mots qui cooccurrent une fois dans le corpus de petite taille.

Soit $E_g^1 = \{cooc_1^g(w_i, w_j) = o_{ij}, i \in [1; N], j \in [1; N]\}$ l'ensemble des couples de mots qui cooccurrent avec leurs valeurs observées o_{ij} dans le corpus de grande taille. La valeur moyenne d'augmentation μ_k est calculée comme suit :

$$\mu_k = \frac{1}{|E_p^k|} \sum_{i=1}^{N} \sum_{j=1}^{M} (cooc_k^g(w_i, w_j) - cooc_k^p(w_i, w_j)) \tag{6.13}$$

Une fois les valeurs $\mu_1, \mu_2, ...\mu_l$ calculées, et pour un corpus de test donné, nous augmentons chaque valeur de cooccurrence observée par la valeur moyenne correspondante comme suit :

$$cooc(w_i, w_j)_{MAE}^* = cooc(w_i, w_j) + \mu_k \tag{6.14}$$

avec μ_k la valeur moyenne correspondante si $cooc(w_i, w_j) = k$.

6.2.2 Prédiction par régression linéaire

La régression linéaire est souvent utilisée pour étudier l'influence d'une variable quantitative X sur une autre variable quantitative Y. La première est souvent appelée variable explicative et la seconde est appelée variable expliquée. Dans notre problématique de prédiction des cooccurrences de mots dans un corpus de grande taille, en partant des observations faites sur un corpus de petite taille, nous sommes confrontés à un type de problème qui peut être résolu en utilisant une régression linéaire. Avec comme variable X, l'ensemble des observations des couples de cooccurrences dans un corpus de petite taille et comme variable Y, l'ensemble des observations des mêmes couples de cooccurrences dans un corpus de grande taille. L'objectif est alors de rechercher une liaison linéaire entre les variables X et Y. Dans notre cas, effectuer une régression linéaire signifie que l'on émet l'hypothèse que la fréquence

de cooccurrence des mots doit croître proportionnellement à la taille du corpus. La droite de régression linéaire $y = ax + b$ a pour but de décrire au mieux la tendance du nuage observé et constitue donc un modèle de prédiction. Ainsi, la ré-estimation de la cooccurrence des mots w_i et w_j par exemple, est représentée par la formule :

$$cooc^*(w_i, w_j) = a \times cooc(w_i, w_j) + b \tag{6.15}$$

avec a et b qui représentent les paramètres de régression.

6.2.3 Les modèles Mean et Max

Les valeurs de cooccurrence croissent en fonction de la taille du corpus. Une autre manière d'estimer cette augmentation est de considérer tout simplement les valeurs observées sur le corpus d'apprentissage de grande taille. Ainsi, l'estimation par la moyenne (Mean) est représentée par l'équation suivante :

$$\mathrm{Mean_k} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{count(k, i)} \tag{6.16}$$

où k représente la valeur observée sur le petit corpus d'apprentissage et i représente la valeur observée pour un couple de mots sur le corpus de grande taille. $count(k, i)$ quant à lui, représente le nombre de couples de cooccurrences qui apparaissent k fois dans un petit corpus et i fois dans un grand corpus. La nouvelle estimation de la cooccurrence des mots w_i et w_j par exemple, est représentée par la formule :

$$cooc(w_i, w_j)^*_{Mean} = cooc(w_i, w_j) + Mean_k \tag{6.17}$$

De la même manière, en utilisant un processus d'estimation par le maximum (Max) chaque paire de cooccurrence est estimée selon l'équation suivante :

$$\mathrm{Max_k} = \frac{1}{\mathrm{N}} \max_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{count(k, i)} \tag{6.18}$$

La nouvelle estimation de la cooccurrence des mots w_i et w_j est alors :

$$cooc(w_i, w_j)^*_{Max} = cooc(w_i, w_j) + Max_k \tag{6.19}$$

6.3 Évaluation

Nous présentons dans ce qui suit les résultats des expériences menées sur les corpus du cancer du sein, des énergies renouvelables et de vulcanologie. L'évaluation est toujours réalisée sur les 3 couples de mesures : TV-JAC, DOR-COS et IM-COS.

6.3.1 Méthodes de lissage

Nous avons comparé l'*approche directe* notée *AD* avec les différentes techniques de lissage appliquées aux vecteurs de contexte de l'*approche directe*, à savoir la technique de Laplace (*Add1*), l'estimateur de Good-Turing (*GT*), la technique par interpolation linéaire de Jelinek-Mercer (*JM*), le Katz-Backoff (*Katz*) et la technique de Kneser-Ney (Kney).

	AD	**Add1**	**GT**	**JM**	**Katz**	**Kney**	
P1	15.5	17.1	18.7	**21.5**	18.7	05.3	IM-COS
P5	31.1	32.7	32.0	**38.3**	33.9	13.4	
P10	34.5	37.0	37.0	**44.8**	38.0	15.2	
MAP	22.6	24.8	25.6	**29.5**	25.9	09.1	
P1	15.8	16.1	16.8	14.6	**17.1**	09.0	DOR-COS
P5	**34.8**	33.6	34.2	33.0	33.9	19.6	
P10	40.4	**41.7**	39.8	38.3	40.1	25.2	
MAP	24.8	24.4	25.2	23.3	**25.3**	14.1	
P1	20.2	**22.4**	14.6	14.6	14.6	16.2	TV-JAC
P5	35.8	**40.5**	27.7	26.7	26.7	29.9	
P10	42.6	**44.2**	34.2	33.3	33.0	33.9	
MAP	27.9	**30.6**	21.4	21.2	21.2	22.9	

TABLE 6.1 – Résultats des expériences sur le corpus du cancer du sein (Mis à part la configuration DOR-COS, les améliorations indiquent un indice de significativité de 0.05 selon le test de Student)

La table 6.1 montre les résultats des expériences menées sur le corpus du cancer du sein. La première observation concerne l'*approche directe* (*AD*). Les meilleurs résultats sont obtenus pour la configuration TV-JAC avec une MAP = 27.9%. Nous pouvons aussi remarquer que seule l'approche *Add*1 améliore significativement les résultats avec une MAP = 30.6%. En revanche, les autres techniques de lissage dégradent les résultats. La deuxième observation concerne la configuration DOR-COS où aucune des techniques de lissage n'améliore significativement les résultats. Bien que les techniques *GT* et *Katz* montrent une légère amélioration avec une MAP = 25.2 % et 25.3 % respectivement, ces résultats ne sont pas significatifs. Le résultat le plus intéressant à noter concerne la configuration IM-COS. Nous pouvons constater que quatre des cinq techniques de lissage améliorent les performances. La meilleure technique étant Jelinek-Mercer (*JM*) avec une MAP = 29.5% et où le top1 est amélioré de 6% et le top10 de 10.3%.

La Table 6.2 montre les résultats des expériences menées sur le corpus des énergies renouvelables. D'une manière générale les résultats suivent le même comportement que celui observé dans la précédente expérience. Les meilleurs résultats de l'*approche directe* sont obtenus en utilisant la configuration TV-JAC avec une MAP = 25.7%. Là encore, seule la technique du *Add*1 améliore les résultats avec

	AD	Add1	GT	JM	Katz	Kney	
P1	07.0	14.0	14.0	**21.0**	16.0	09.0	IM-COS
P5	27.0	32.0	31.0	**37.0**	30.0	17.0	
P10	37.0	42.0	43.0	**51.0**	44.0	28.0	
MAP	17.8	23.6	22.9	**30.1**	24.2	14.1	
P1	12.0	**17.0**	12.0	12.0	12.0	06.0	DOR-COS
P5	31.0	**35.0**	31.0	32.0	28.0	16.0	
P10	38.0	**44.0**	36.0	39.0	35.0	21.0	
MAP	21.8	**26.5**	19.8	20.8	19.7	11.1	
P1	17.0	**22.0**	13.0	13.0	13.0	14.0	TV-JAC
P5	36.0	**38.0**	27.0	27.0	27.0	29.0	
P10	42.0	**50.0**	37.0	38.0	38.0	39.0	
MAP	25.7	**29.7**	20.5	21.3	21.3	22.9	

TABLE 6.2 – Résultats des expériences sur le corpus des énergies renouvelables (Mis à part la configuration DOR-COS, les améliorations indiquent un indice de significativité de 0.05 selon le test de Student)

une MAP = 29.7%. Concernant la configuration DOR-COS, excepté la technique *Add*1 aucune des méthodes de lissage n'améliore significativement les résultats. Finalement, le résultat le plus probant concerne la configuration IM-COS. Là aussi, quatre des cinq techniques de lissage améliorent les résultats, la meilleure étant JM avec une MAP = 30.1% et une amélioration de 14.0% pour le top1 et le top10.

	AD	Add1	GT	JM	Katz	Kney	
P1	13.9	21.6	17.1	**22.9**	21.0	11.2	IM-COS
P5	29.1	38.2	33.1	**43.9**	36.9	23.4	
P10	39.8	45.2	41.4	**53.5**	47.7	30.1	
MAP	21.7	29.7	25.2	**32.3**	29.7	15.7	
P1	20.2	19.1	21.0	20.3	**21.6**	10.1	DOR-COS
P5	**41.7**	39.4	41.4	37.5	40.1	18.4	
P10	51.2	49.0	51.5	51.1	**52.2**	26.1	
MAP	30.3	28.7	**30.4**	29.3	**30.4**	15.6	
P1	37.9	**40.1**	33.7	29.2	29.2	31.2	TV-JAC
P5	**56.3**	56.0	54.7	50.9	50.9	50.3	
P10	**62.0**	60.5	61.1	57.3	57.3	58.5	
MAP	46.8	**47.3**	44.2	39.4	39.4	40.0	

TABLE 6.3 – Résultats des expériences sur le corpus de vulcanologie (Mis à part la configuration DOR-COS, les améliorations indiquent un indice de significativité de 0.05 selon le test de Student)

La Table 6.3 montre les résultats des expériences menées sur le corpus de vulcanologie. Là aussi les résultats confirment le comportement de l'*approche directe* vis-à-vis des techniques de lissage. La technique du *Add*1 améliore les résultats avec une MAP = 47.3% (configuration TV-JAC). Cependant, l'augmentation est moins importante en comparaison avec les autres corpus. Nous remarquons aussi une légère baisse aux tops 5 et 10. Concernant la configuration DOR-COS, aucune des méthodes de lissage n'améliore significativement les résultats. Pour finir, le résultat le

plus important concerne la configuration IM-COS. Là encore, quatre des cinq techniques de lissage améliorent les résultats, la meilleure étant JM avec une MAP = 32.3% et une amélioration de 09.0% pour le top1 et de 13.7% pour le top10.

6.3.2 Méthodes de prédiction

Nous comparons dans cette expérience l'*approche directe* avec les différentes techniques de prédiction appliquées aux vecteurs de contexte de l'*approche directe*, à savoir la technique par sélection du maximum et de la moyenne (Max, $Mean$), le modèle de régression linéaire ($LReg$) et l'augmentation moyenne des cooccurrences (MAE). Nous rajoutons aussi les résultats du modèle de Good-Turing (GT) à titre comparatif, sachant que celui-ci peut être considéré comme un modèle de prédiction.

	AD	**Max**	**Mean**	**LReg**	**MAE**	**GT**	
P1	15.5	**20.2**	13.7	18.0	18.6	18.6	IM-COS
P5	31.1	**35.8**	28.3	**35.8**	34.2	32.0	
P10	34.5	41.1	32.7	**42.0**	38.3	37.0	
MAP	22.6	**27.2**	20.3	26.7	26.4	25.6	
P1	15.8	15.5	11.8	**19.9**	13.7	16.8	DOR-COS
P5	**34.8**	30.2	28.6	34.2	27.7	34.2	
P10	40.4	36.7	35.5	**41.7**	33.0	39.8	
MAP	24.8	22.9	19.8	**27.6**	20.9	25.2	
P1	**20.2**	06.5	16.5	15.5	09.9	14.6	TV-JAC
P5	**35.8**	15.5	33.9	28.6	21.4	27.7	
P10	**42.6**	20.5	38.3	37.3	26.7	34.2	
MAP	**27.9**	11.6	24.6	22.6	15.6	21.4	

TABLE 6.4 – Résultats des expériences sur le corpus du cancer du sein (les améliorations indiquent un indice de significativité de 0.05 selon le test de Student)

La table 6.4 illustre les résultats des expériences menées sur le corpus du cancer du sein. Nous pouvons remarquer que pour la configuration TV-JAC aucun des modèles de prédiction n'améliore les résultats. Bien au contraire, les résultats sont même dégradés. Concernant la configuration DOR-COS, seuls les modèles naïfs $Mean$ et Max sont en deçà de l'approche de référence AD. Le meilleur score est obtenu en utilisant le Modèle $LReg$, avec une MAP = 27.6%. Le résultat le plus notable reste celui concernant la configuration IM-COS. Ainsi, quatre des cinq modèles de prédiction améliorent les résultats. Le meilleur étant le modèle du Max qui atteint une MAP = 27.2% et améliore la précision de 4.8% au top1 et de 6.6% au top10.

La table 6.5 montre les résultats obtenus en exploitant le corpus des énergies renouvelables. Globalement les résultats suivent le même comportement que l'expérience menée sur le corpus du cancer du sein. Là encore, aucun modèle de prédiction n'est efficace quand il est associé à la configuration TV-JAC. Les modèles $Mean$ et Max sont les seuls à montrer des résultats moins bons que ceux de l'*approche directe*

	AD	Max	Mean	LReg	MAE	GT	
P1	07.0	13.0	10.0	**18.0**	15.3	14.0	IM-COS
P5	27.0	34.0	30.0	**37.0**	33.0	31.0	
P10	37.0	**46.0**	36.0	**46.0**	43.0	43.0	
MAP	17.8	23.1	19.2	**28.0**	25.0	22.9	
P1	12.0	09.0	06.0	**14.0**	10.0	12.0	DOR-COS
P5	31.0	20.0	27.0	**32.0**	25.0	31.0	
P10	38.0	26.0	39.0	**40.0**	33.0	36.0	
MAP	21.8	15.7	17.0	**23.3**	18.0	19.8	
P1	17.0	09.0	**18.0**	15.0	18.0	13.0	TV-JAC
P5	**36.0**	16.0	30.0	31.0	29.0	27.0	
P10	42.0	22.0	**45.0**	36.0	36.0	37.0	
MAP	**25.7**	14.0	25.1	22.9	23.7	20.5	

TABLE 6.5 – Résultats des expériences sur le corpus des énergies renouvelables (Mis à part la configuration DOR-COS, les améliorations indiquent un indice de significativité de 0.05 selon le test de Student)

en utilisant la configuration DOR-COS. Les meilleures performances sont données par l'approche *LReg* avec une MAP = 23.3%. La configuration IM-COS est la plus à même d'être améliorée par quatre des cinq modèles de prédiction. Contrairement à l'expérience précédente, c'est le modèle *LReg* qui donne les meilleurs résultats avec une MAP = 28.0% et une amélioration de 11.0% au top1 et de 10.2% au top10.

	AD	Max	Mean	LReg	MAE	GT	
P1	13.9	**22.1**	08.2	20.2	18.9	17.1	IM-COS
P5	29.1	32.9	22.1	**37.3**	34.1	33.1	
P10	39.8	39.2	27.8	**41.7**	40.5	41.4	
MAP	21.7	27.4	14.7	**27.5**	25.8	25.2	
P1	20.2	20.8	13.2	**21.5**	20.8	21.0	DOR-COS
P5	**41.7**	29.7	25.9	36.0	36.7	41.4	
P10	51.2	37.9	31.6	43.0	41.7	**51.5**	
MAP	30.3	26.1	19.5	29.2	28.4	**30.4**	
P1	**37.9**	16.45	16.45	27.2	28.4	33.7	TV-JAC
P5	**56.3**	32.9	32.2	41.7	43.6	54.7	
P10	**62.0**	39.2	38.6	48.7	51.2	61.1	
MAP	**46.8**	23.8	24.5	34.2	35.4	44.2	

TABLE 6.6 – Résultats des expériences sur le corpus de vulcanologie (Mis à part la configuration DOR-COS, les améliorations indiquent un indice de significativité de 0.05 selon le test de Student)

La table 6.6 montre les résultats obtenus sur le corpus de vulcanologie. Nous remarquons là encore qu'aucun modèle de prédiction n'est efficace quand il est associé à la configuration TV-JAC. Contrairement aux autres corpus, les résultats de la prédiction sur la configuration DOR-COS sont moins bons que ceux de l'*approche directe*. La configuration IM-COS obtient les plus grandes améliorations sur quatre des cinq modèles de prédiction. Là aussi, c'est le modèle *LReg* qui donne les meilleurs résultats avec une MAP = 27.5% et une amélioration de 6.3% au top1 et de 1.9%

au top10. Nous notons cependant que ces améliorations sont moins importantes que celles obtenues pour les autres corpus. Cependant la configuration TV-JAC reste la plus performante, et les améliorations apportées aux autres configurations restent bien en deçà.

6.4 Discussion

Les techniques de lissage sont souvent évaluées d'après leur capacité à prédire les n-grammes non observés dans la phase d'apprentissage. Nous nous sommes exclusivement focalisés dans nos expériences sur les cooccurrences observées. Ainsi, le comportement des différentes techniques de lissage n'est pas forcément en adéquation avec les résultats de l'état de l'art. C'est le cas par exemple de la méthode de Laplace (*Add*1), considérée comme une technique pauvre en traitement automatique du langage naturel : elle a néanmoins montré des résultats intéressants avec la configuration TV-JAC. Les bons résultats de l'approche *Add*1 sont sans doute liés à la non considération des cooccurrences non observées, principale faiblesse de *Add*1. Concernant la configuration DOR-COS et malgré une légère amélioration en utilisant *Add*1, les techniques de lissage ont montré des résultats décevants. Si nous ne pouvons expliquer cela de manière formelle, nous supposons néanmoins que cela peut venir de la mesure du *discounted odds-Ratio* qui se base sur la table de contingence. Ainsi, lisser l'information conjointe (les cooccurrences observées) sans lisser les informations disjointes pourrait expliquer ces résultats. Cette remarque est valable aussi pour la mesure du *taux de vraisemblance*. Là encore, aucune méthode de lissage (excepté *Add*1) n'a montré une amélioration des résultats pour la configuration TV-JAC. Ceci reste le principal questionnement auquel nous n'avons pas trouvé de réponse claire. Ainsi, les mesures d'association du *discounted odds-Ratio* et du *taux de vraisemblance* ne semblent pas être compatibles avec la plupart des techniques de lissage. Plus d'investigations seront sans doute nécessaires pour comprendre les raisons de ces échecs.

Le résultat le plus remarquable concerne la configuration IM-COS. Excepté le lissage de Kneser-Ney, toutes les autres techniques ont montré de meilleurs résultats que l'*approche directe* sans lissage. La technique par interpolation linéaire de Jelinek-Mercer a été la plus performante. Nous expliquons ces résultats par le fait que la mesure de l'information mutuelle a tendance à surestimer les cooccurrences de faible fréquence et utiliser des techniques de lissage permet sans doute de corriger cela.

En TALN, le lissage des modèles n-grammes a été traité dans une multitude de travaux [Chen et Goodman, 1999]. L'estimateur de Good-turing qui est rarement utilisée seule, forme la base d'autres techniques comme le Katz-Backoff ou l'interpolation linéaire de Jelinek-Mercer, deux techniques considérées comme performantes de manière générale. La surprise vient sans doute de la technique proposée par Kneser-Ney : si celle-ci est très connue pour être l'une des meilleures techniques de lissage, les résultats obtenus dans nos expériences furent très décevants. L'explication peut venir d'une part, du fait de ne pas considérer les cooccurrences non observées et d'autre part, soustraire une valeur fixe pour toutes les cooccurrences observées pourrait altérer le modèle de cooccurrences de base et renforcer la surestimation des cooccurrences de faible fréquence dans le cas de l'information mutuelle.

En outre, nous avons présenté une autre manière de ré-estimer les cooccurrences des mots en considérant cette tâche comme un problème de prédiction. Nous avons pu constater encore une fois les bonnes performances des modèles de prédiction (hormis MEAN) pour la configuration IM-COS. Là encore, les bons résultats sont sans doute liés au biais engendré par l'information mutuelle sur les cooccurrences de faible fréquence, biais qui semble être corrigé par les modèles de prédiction. Globalement ce sont les méthodes *MAX* et *LReg* qui montrent les meilleures performances. Comme pour les techniques de lissage, les méthodes de prédiction n'ont pas apporté d'amélioration en utilisant la configuration TV-JAC. Alors que les modèles de prédiction tentent d'augmenter les cooccurrences des mots, dans le cas de la mesure du taux de vraisemblance ceci provoque le contraire de l'effet escompté. Nous remarquons le même comportement concernant la mesure du *discounted odds-Ratio* qui hormis une légère amélioration en utilisant le modèle *LReg*, offre de mauvais résultats quand elle est associée aux autres méthodes de prédiction. Si les mesures du taux de vraisemblance et du *discounted odds-Ratio* sont basées sur la table de contingence et s'il nous parait difficile de construire des modèles de prédiction ou de lissage sur les informations disjointe, une autre solution serait de construire ces modèles de prédiction et/ou de lissage une fois les mesures d'association calculées. Des travaux supplémentaires dans ce sens sont nécessaires.

6.5 Bilan

Nous nous sommes intéressés dans ce chapitre à deux manières de ré-estimer les cooccurrences des mots en utilisant différentes méthodes de prédiction et de lissage. Les résultats expérimentaux ont montré que ré-estimer les cooccurrences des mots permettait d'améliorer significativement les performances de l'*approche directe* en utilisant l'information mutuelle comme mesure d'association. Nous notons aussi la performance de la technique *Add*1 associée au taux de vraisemblance. Des travaux futurs seront sans doute nécessaires pour mieux utiliser les modèles de ré-estimation au *taux de vraisemblance* et au *discounted odds-ratio* ainsi qu'à d'autres mesures d'association non abordées dans ce chapitre.

7

Metarecherche

Introduction

Dans ce chapitre, nous reprenons à notre compte l'idée de [Fung, 1998] qui indique que l'extraction de lexiques bilingues à partir de corpus comparables peut être approchée comme un problème de recherche d'information. Dans cette représentation, la requête serait alors le mot à traduire et les documents retournés par le moteur de recherche les candidats à la traduction de ce mot. De la même manière que les documents retournés sont ordonnés suivant leur adéquation avec la requête, les traductions candidates seront classées en fonction de leur pertinence par rapport au mot à traduire. Nous souhaitons donc poursuivre plus en avant cette analogie et proposer une amélioration significative à l'*approche par similarité interlangue* en considérant l'extraction de lexiques bilingues comme un problème de fusion de résultats analogue à celui rencontré par les métamoteurs de recherche d'information. Nous faisons ainsi l'hypothèse que le fait de combiner différentes sources d'information permet de renforcer globalement l'*approche par similarité interlangue*.

7.1 Méthode *Metarecherche*

Cette section décrit notre extension de l'*approche par similarité interlangue*. Les différents modes de fusion définis ici se basent sur les éléments décrits dans la table 7.1. L'approche proposée par [Déjean et Gaussier, 2002] introduit implicitement le problème du choix de la valeur adéquate k des plus proches voisins ($kppv$). D'une manière générale, la valeur optimale de k dépend des données mises en jeu. Cette valeur est souvent définie de façon empirique, bien qu'il soit possible de la déterminer par validation croisée. L'*approche par similarité interlangue* (SIL) appliquée à nos données s'est révélée très sensible vis-à-vis du paramètre k. Pour des valeurs de k supérieures à 20, la précision chute de façon significative. De plus, il n'est pas possible de déterminer des intervalles de stabilité relatives pour k. Le choix du paramètre k devient alors crucial. En partant du principe que chaque mot contribue à la caractérisation du mot à traduire, notre proposition vise non seulement à améliorer la

précision, mais aussi à être plus robuste vis-à-vis du nombre de *ppv*. En poussant l'analogie des approches inspirées de la RI [Fung et Yee, 1998] plus loin, nous proposons une nouvelle façon d'aborder le problème de l'extraction lexicale bilingue à partir de corpus comparables, en le considérant comme un problème de fusion de résultats analogue à celui rencontré par les métamoteurs de recherche.

L'objectif de la *métarecherche* est de fusionner les classements renvoyés par plusieurs systèmes de RI en une liste unique, afin d'obtenir un système combiné qui soit plus performant que les systèmes individuels [Aslam et Montague, 2001]. Puisque chacun des *ppv* produit un classement différent, la *métarecherche* fournit un cadre adéquat pour exploiter l'information véhiculée par chacun des k classements. Un intérêt particulier est donné aux mots candidats à la traduction d'un mot donné. Partant du principe que les corpus contiennent beaucoup de bruit, il n'est pas rare de rencontrer des mots qui soient proches d'un nombre important de mots du dictionnaire, et viennent parasiter le modèle et fausser les résultats. Pour traduire un mot, le système choisit un nombre k de plus proches voisins en langue source, puis il cherche en langue cible les candidats les plus proches des traductions de ces *kppv* sans tenir compte de la relation de ces candidats avec le reste des mots du dictionnaire. Pour pallier cela, nous construisons un modèle qui prend en compte cette information en accordant plus de confiance aux candidats qui sont plus proches des *kppv* que du reste des entrées du dictionnaire.

Symbole	Définition
i	Le mot à traduire
t	Le mot candidat à la traduction de i
s	L'ensemble des plus proches voisins de i
$\overline{s}$	L'ensemble des traductions des plus proches voisins de i
k	Le nombre de plus proches voisins sélectionnés
n	L'ensemble de tous les voisins de t
u	Le nombre total de mots du dictionnaire
$occ_{\overline{s}(t)}$	L'effectif de t ie : avec combien de voisins t est-il en relation ?
$\mathrm{sim}(\overline{s_k}, t)$	Le score de similarité entre le k ième proche voisin de $\overline{s}$ et t
$\max_{\overline{s_k}}$	Le score maximum du candidat le plus proche de $\overline{s_k}$
$\max_{\overline{s}}$	Le score maximum du candidat le plus proche de l'ensemble $\overline{s}$
$\mathrm{sim}_k(s,t)$	Le score de similarité entre s et t par rapport au k ième plus proche voisin
$\mathrm{sim}(s,t)$	Le score de similarité entre s et t pour l'ensemble des plus proches voisins
θ_t	Le paramètre de régulation ou facteur de confiance de t

TABLE 7.1 – Éléments de notation

La première étape de notre méthode consiste à fixer le nombre de plus proches voisins d'un mot à traduire. La valeur de k est déterminée empiriquement. Cependant, intuitivement mais aussi à travers nos expériences, nous pouvons dire que la sélection d'un nombre faible de plus proches voisins est insuffisante dans la plupart des cas pour trouver la bonne traduction, et que la sélection d'un grand nombre de voisins, d'une part contredit la notion de plus proches voisins et d'autre part, induit la prise en compte de voisins éloignés qui peuvent fausser le modèle.

Une fois k fixé, nous considérons chaque liste de candidats renvoyée par un proche voisin indépendamment des autres. Ces candidats sont les mots dont les vecteurs de contexte sont les plus similaires au vecteur de contexte d'un voisin donné. Dans nos expériences, la taille des listes a été fixée à 200. Partant du même principe

que le choix du paramètre k. La taille de la liste joue un rôle important. En effet, une liste trop petite de candidats ne serait pas suffisante pour aider à trouver la bonne traduction, de la même façon, une liste trop importante de mots risque de rajouter du bruit car il faut garder en tête que les mots appartenant à une liste sont les traductions potentielles classées par ordre de score de similarité. Notre modèle privilégie les candidats qui apparaissent dans plusieurs listes, ainsi, plus l'effectif du mot candidat est important plus il a de chances d'être une bonne traduction. Ceci reste valable si le candidat est bien classé, en revanche, s'il apparaît souvent mais en étant toujours mal classé par rapport aux différentes listes, ce mot a de fortes chances d'être une mauvaise traduction.

Le raisonnement qui conduit à ce calcul est le suivant : Les scores des différents classements sont sur la même échelle car donnés par la même mesure de similarité. Si $\max(l) \gg \max(m)$, cela veut dire que selon le système, le classement l est plus sûr que le classement m (indépendamment de la réponse réelle). Nous considérons ici que les classements donnés par les k plus proches voisins du mot à traduire sont le résultat de k moteurs de RI différents. À l'instar des métamoteurs de recherche, nous allons tenter de nous servir des scores des mots pour améliorer l'extraction bilingue. Une des approches majeures en *métarecherche* est le modèle de fusion linéaire [Bartell *et al.*, 1994] où le score final d'un terme i est la somme pondérée de chacun des scores obtenus :

$$\mathrm{sim}(i,t) = \theta_t \times \frac{\sum_{j=1}^{k} \mathrm{sim}_j(i,t)}{\sum_{j=1}^{n} \mathrm{sim}(\overline{s_j},t)} \tag{7.1}$$

Pour réduire l'influence des candidats à la traduction qui apparaissent dans différents contextes lexicaux et qui peuvent par leur forte fréquence d'apparition induire en erreur les systèmes d'extraction lexicale basés sur le contexte, on se propose de prendre en compte ce phénomène en considérant en plus du score calculé à partir des k plus proches voisins, un score défini à partir de toutes les entrées du dictionnaire pour lequel le terme candidat est lié. L'équation 7.1 permet de calculer le score de similarité entre i et t en prenant en considération le score de similarité par rapport aux k plus proches voisins choisis, normalisé par la somme des scores de t par rapport à tous ses voisins et pondéré par le paramètre de confiance θ. Le poids θ est donné par :

$$\theta_t = occ_{\overline{s}}(t) \times \frac{(u - (k - occ_{\overline{s}}(t)))}{(u + occ_n(t))} \tag{7.2}$$

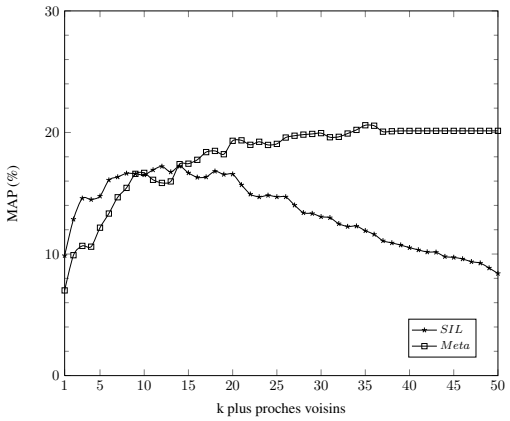
L'équation 7.2 prend en compte l'effectif du candidat par rapport aux k plus proches voisins, c'est-à-dire le nombre de voisins avec lesquels le mot t est en relation. Ceci est représenté par $occ_{\overline{s}}(t)$. Nous privilégions ainsi les mots avec un effectif élevé. Le numérateur $(u - (k - occ_{\overline{s}}(t)))$ permet de considérer l'effectif de t dans l'ensemble $\overline{s}$ par rapport à tous les mots du dictionnaire. On normalisera ensuite par la distribution de t par rapport à tous ses voisins à l'aide de $u + occ_n(t)$. Le paramètre θ permet donc d'accorder plus de confiance à un mot candidat à la traduction qui a un effectif élevé par rapport aux k plus proches voisins mais qui a aussi un effectif faible par rapport au reste de ses voisins.

7.2 Évaluation

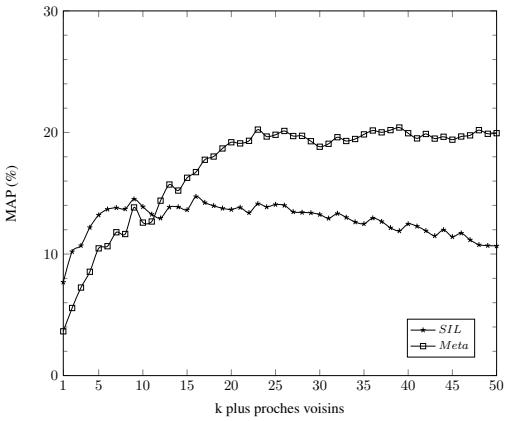
Pour évaluer les performances de notre approche ($Meta$), nous utilisons comme référence l'*approche par similarité interlangue* (SIL) proposée par [Déjean et Gaussier, 2002] dans sa version *plate*. Les k plus proches voisins sont calculés en utilisant la mesure d'association du taux de vraisemblance (TV) et la mesure de similarité du Jaccard (JAC).

7.2.1 Variation des k plus proches voisins

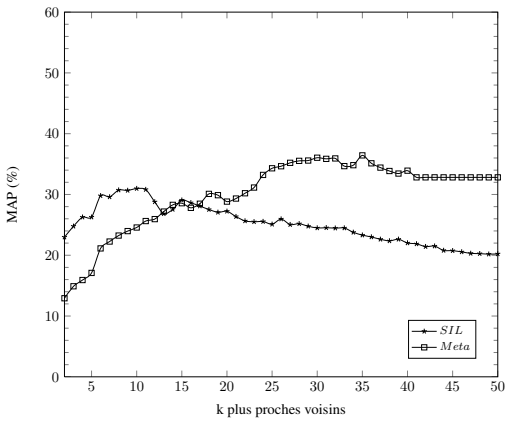
Nous faisons varier dans cette expérience le paramètre k correspondant au nombre de plus proches voisins d'un candidat à la traduction. La comparaison entre la méthode par similarité interlangue (SIL) et la méthode meta-recherche ($Meta$) est illustrée dans la figure qui suit pour les trois corpus de spécialité.



(a) Cancer du sein



(b) Énergies renouvelables



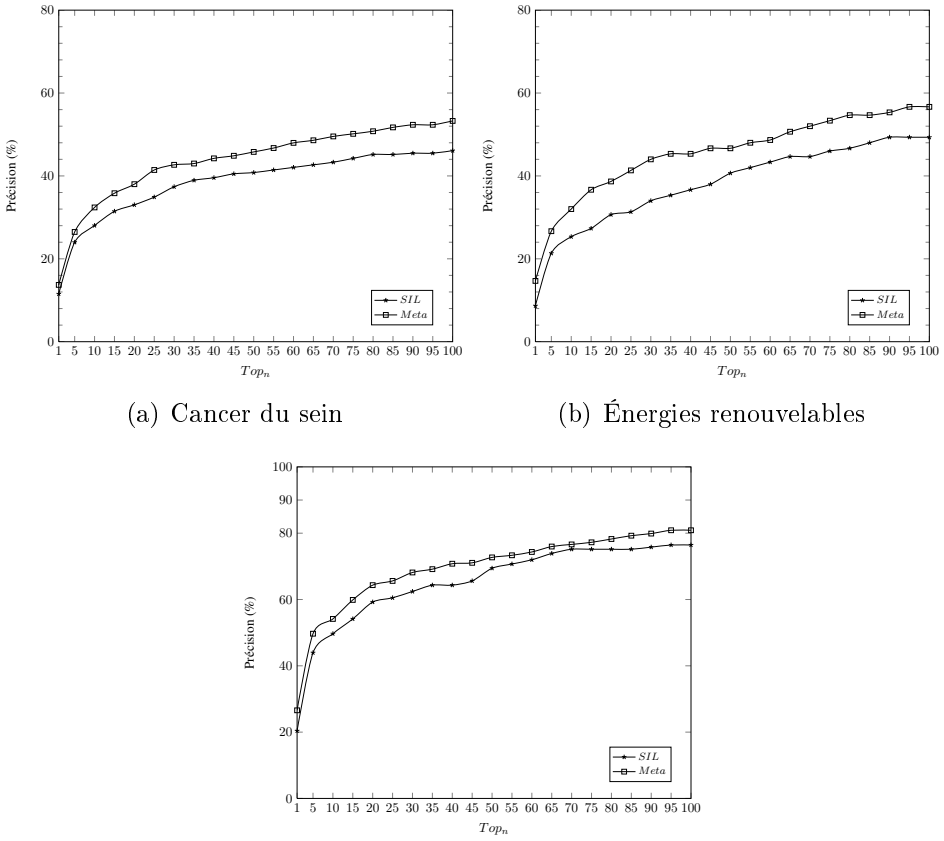
(c) Vulcanologie

FIGURE 7.1 – Comparaison en MAP % de SIL et Meta en fonction des k plus proches voisins

Nous pouvons remarquer que l'approche *Meta* est moins sensible à la variation des *ppv* en comparaison avec l'approche *SIL* à partir d'un $k > n$ ($n \in [15, 25]$). Les résultats sont plus nets sur les 3 corpus à partir de $k = 25$. Les moins bons résultats pour un $k < 15$ s'expliquent par le fait que l'approche *Meta* a besoin d'un minimum de proches voisins pour être efficace.

7.2.2 Représentation des meilleures configurations

Nous comparons dans cette expérience les meilleurs résultats des méthodes *SIL* et *Meta*.



(a) Cancer du sein

(b) Énergies renouvelables

(c) Vulcanologie

FIGURE 7.2 – Comparaison de SIL et Meta en fonction de leur meilleure configuration

La figure 7.2 montre que l'*approche par métarecherche* améliore les résultats de l'*approche par similarité interlangue* pour les trois corpus comparables. L'amélioration est plus nette sur le corpus du cancer du sein et des énergies renouvelables.

7.3 Discussion

Les *approches par similarité interlangue* et par *métarecherche* se basent sur les k plus proches voisins pour identifier les meilleurs candidats à la traduction. L'approche *SIL* effectue une fusion en amont, privilégiant ainsi une vue globale des k

plus proches voisins. Ceci peut se révéler problématique, car une bonne traduction pourrait être noyée dans la masse et ainsi écartée de la liste des candidats, si des mots plus fréquents viennent à apparaître dans le contexte du mot à traduire. Plus précisément, si des mots obtiennent des scores de similarité très élevés par rapport à un seul plus proche voisin et que d'autres obtiennent des scores moins élevés mais proches de plusieurs plus proches voisins du mot à traduire, ces mots seront moins bien classés voire mal classés.

Pour pallier ce problème, l'approche *Meta* considère dans un premier temps, chaque plus proche voisin comme étant une source d'information indépendante des autres, privilégiant ainsi sa liste de candidats en fixant une taille arbitraire (généralement autour de 200 dans nos expériences), pour ensuite effectuer une fusion en aval des plus proches voisins après avoir normalisé les scores de similarité. En outre, l'approche *Meta* introduit une mesure de fiabilité en considérant les plus proches voisins des mots candidats à la traduction comme étant proches des k plus proches voisins du mot à traduire, ainsi que tous les voisins de ces candidats, pour éloigner des mots qui apparaîtraient trop fréquemment et dans trop de contextes. Car ne l'oublions pas, ces approches se basent uniquement sur une représentation graphique des données qui induit un certain volume de bruit, lequel serait sans doute mieux traité par une analyse plus fine du contexte.

Il est évident que plus les mots sont fréquents dans le corpus plus on a une représentation riche de leur contexte. Cette remarque nous amène à nous interroger sur les fréquences des k plus proches voisins du mot candidat à la traduction. Si un plus proche voisin apparaît fréquemment en langue source et que sa traduction en langue cible est faible ou inversement, quel serait l'impact de ce déséquilibre sur les résultats ? Aucune étude à notre connaissance n'a approfondi ce sujet. Quoique rien ne nous permette d'affirmer une quelconque relation entre ce déséquilibre et une éventuelle traduction erronée, nous pouvons néanmoins supposer que cela est nuisible à une représentation riche du contexte du mot, car un mot peu fréquent apporte moins d'information qu'un mot très fréquent et ceci toujours en se basant sur l'idée de la représentation graphique qui caractérise le contexte. Nous nous attellerons dans nos travaux futurs à étudier cette problématique.

Enfin, les plus proches voisins ont été fixés d'une manière empirique dans les deux approches *SIL* et *Meta*, et dans toutes les évaluations. Nous avons fixé un même k pour tous les mots de la liste d'évaluation. L'état de l'art ne spécifie aucune manière efficace de choisir ce paramètre k. Néanmoins, nous sommes en droit de nous interroger sur l'existence d'un nombre k idéal de plus proches voisins qui puisse garantir une bonne traduction de tous les mots de la liste d'évaluation ? Nous serions plutôt tentés de dire qu'il existe un k pour chaque mot à traduire mais que celui-ci varie selon les mots. Là encore, nos travaux futurs devront répondre à cette question clé.

7.4 Bilan

Nous avons présenté dans ce chapitre une nouvelle manière d'aborder le problème de l'extraction lexicale bilingue à partir de corpus comparables en nous appuyant sur le principe des métamoteurs de recherche. Nous avons présenté une nouvelle

approche simple et robuste qui revisite l'approche par similarité interlangue pour présenter un modèle inspiré par les métamoteurs de recherche d'information. Ce modèle qui prend en compte la distribution des candidats à la traduction non seulement par rapport aux k plus proches voisins du mot à traduire mais aussi par rapport à tous leurs voisins, a permis un gain significatif en terme de précision. Les résultats empiriques que nous obtenons montrent que les performances de ce nouveau modèle sont toujours supérieures à celles obtenues avec l'*approche par similarité interlangue* pour $k > 15$.

8

Q-Align : Un nouveau système pour l'extraction de lexiques bilingues à partir de corpus comparables

Introduction

Les méthodes d'extraction de lexiques bilingues à partir de corpus comparables sont pour la plupart basées sur l'hypothèse distributionnelle et considèrent le contexte d'une manière globale. C'est-à-dire que tous les mots contextuels d'un mot donné sont collectés et représentés dans un vecteur, sans tenir compte de leurs positions dans le corpus vis-à-vis du terme à traduire. Nous nous intéressons ici à une représentation locale du contexte, où l'idée est de localiser des passages dans le corpus qui ont de fortes chances de contenir la traduction recherchée. Pour cela, nous tentons d'aligner des passages sources contenant le mot à traduire avec des passages cibles susceptibles de contenir la bonne traduction. Inspiré du domaine de la recherche d'information d'une manière générale et des systèmes de questions réponses (SQR) d'une manière plus spécifique, nous considérons donc l'alignement de termes comme étant analogue à celui de la recherche de la réponse à une question dans un document (par exemple à l'aide d'un moteur de recherche). Ainsi, pour un mot à traduire nous récupérons la fenêtre où ce mot apparaît, ceci constituera notre requête. Une fois la requête traduite, nous mesurons une distance entre cette requête et les segments présents en langue cible. Ces segments sont simplement des fenêtres qui peuvent correspondre à une phrase, un paragraphe ou un document. Nous supposons que la traduction candidate apparaît dans le segment au même titre qu'une réponse à une question dans le cadre des systèmes de question/réponse. Nous avons nommé ce système *Q-Align*.

8.1 La méthode *Q-Align*

La méthode *Q-Align* est basée sur le principe des systèmes de question/réponse et plus précisément, sur une étape intermédiaire qui est l'extraction de passages. En effet, la majorité des SQR passent par l'étape d'extraction de passages susceptibles de contenir la bonne réponse, pour ensuite analyser les passages les plus pertinents et obtenir la réponse recherchée. Par analogie, notre intuition est de dire que si un passage source contenant le mot à traduire w et un passage cible partagent des mots en commun, alors il y a des chances pour que la traduction de w soit présente dans le passage cible. La méthode *Q-Align* peut être décomposée en trois étapes qui sont :

1. Extraction des requêtes source contenant le terme à traduire ;
2. Traduction des requêtes source contenant le terme à traduire ;
3. Alignement des requêtes traduites avec les segments de la langue cible.

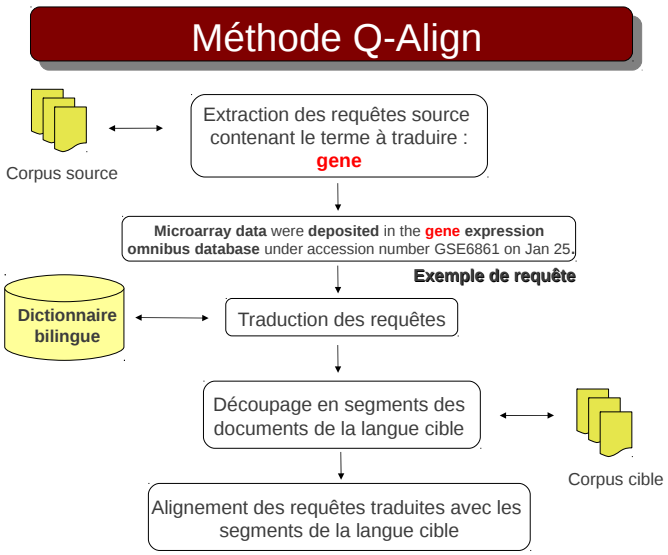


FIGURE 8.1 – Illustration des étapes de la méthode Q-Align

Nous détaillons dans ce qui suit ces différentes étapes.

8.1.1 Extraction des requêtes

La première étape consiste à collecter les requêtes qui contiennent le mot à traduire. Chaque requête est une fenêtre contextuelle. Deux paramètres sont à fixer : (i) La taille de la requête, donc le nombre de mots entourant le mot à traduire (ceci correspond au sac de mots), notons ce paramètre par w_q. Si nous prenons par exemple $w_q = 5$, ceci veut dire que la requête contient deux mots avant et après le mot à traduire. Ceci est illustré dans l'exemple qui suit concernant le mot **replica** (Il est à noter que les mots outils ont été retirés) :

$detail_V$	$painting_N$	**replica**$_N$	$line_V$	$separate_J$

TABLE 8.1 – Exemple d'une requête en anglais du terme **replica**

avec les catégories grammaticales N pour désigner un nom, V pour désigner un verbe et J pour désigner un adjectif.

(ii) Le deuxième paramètre est le nombre de requêtes à sélectionner. Nous partons du principe que toutes les requêtes ne sont pas forcement utiles pour trouver la bonne traduction. Ainsi, nous cherchons à identifier les n meilleures requêtes. Cela revient à se poser la question suivante : comment sélectionner ces requêtes ?

La manière la plus simple est d'affecter à chaque requête un score pour ensuite ordonner ces requêtes et en sélectionner les n meilleures. Ce calcul est fait selon l'équation suivante :

$$Score(query_n) = \sum_{i=1}^{w_q - 1} freq(word_i) \tag{8.1}$$

Ceci dit, rien ne garantit que cette méthode soit la plus efficace. Nous discuterons ce point à la fin de ce chapitre.

8.1.2 Traduction des requêtes

Chaque requête sélectionnée est traduite en langue cible. En reprenant l'exemple précédent concernant le mot **replica** et en prenant le français comme langue cible, nous obtenons :

Mot	Traduction
$detail_V$	$désigner_V$
$painting_N$	$peinture_N$
replica$_N$	**Unknown**$_N$
$line_V$	$marquer_V$
$separate_J$	$indépendant_J$

TABLE 8.2 – Représentation d'une requête du mot **replica** et sa traduction en français

La requête traduite sera utilisée dans le corpus cible comme illustré dans la table suivante :

$désigner_V$	$peinture_N$	$marquer_V$	$indépendant_J$

TABLE 8.3 – Requête traduite du mot **replica**

Il est à noter que les mots de la requête sont traduits à l'aide d'un dictionnaire bilingue et ceci en tenant compte de leurs catégories grammaticales.

8.1.3 Extraction des traductions candidates

Pour sélectionner un mot candidat à la traduction, nous utilisons la mesure de compacité telle que définie dans [Gillard *et al.*, 2007, Voorhees, 2002]. Le principe de la compacité dans un système de question/réponse est de mesurer la similarité entre une question donnée et un segment dans un texte donné. Le segment peut être une phrase, un paragraphe ou un document. Dans notre cas et par analogie aux SQR, nous mesurons la compacité entre la requête traduite et un segment donné en langue cible. Le score de compacité final $Compact_{All}(\bar{w_x})$ du mot $\bar{w_x}$ est simplement la somme des compacités de toutes les requêtes préalablement sélectionnées. Ceci est représenté dans l'équation suivante :

$$Compact_{All}(\bar{w_x}) = \sum_{i \in nbQuery} Compact(\bar{w_x})_i \tag{8.2}$$

Une question qui se pose est : comment choisir les segments de la langue cible ? Tous les documents de la langue cible sont découpés en segments de taille fixe. Chaque segment sera exploré pour voir s'il contient la bonne traduction. Notons par w_{seg} la taille du segment, ce qui correspond au nombre de mots dans un segment. Étant donné une requête traduite et un segment, la compacité de $\bar{w_x}$ pour le segment s est définie par l'équation suivante :

$$Compact_s(\bar{w_x}) = \frac{1}{|WQ|} \sum_{i=\in WQ} Contrib(w_i)_{\bar{w_x}} \tag{8.3}$$

où $Contrib(w_i)_{\bar{w_x}}$ est la contribution de chaque mot de la requête. Prenons un exemple pour illustrer la manière de calculer la contribution et la compacité. Notons QR comme étant l'ensemble des mots de la requête traduite, comme le montre la table 8.4, avec $w_q = 5$ et w_i un mot de la requête traduite. Dans l'exemple $QR = \{w_1, w_2, w_3, w_4\}$ et $\mathbf{Cand}_N$ est le mot dont nous cherchons la traduction.

w_1	w_2	$\mathbf{Cand}_N$	w_3	w_4

TABLE 8.4 – Requête du mot à traduire

Prenons un segment $w_{seg} = 8$. Chaque mot du segment qui n'apparaît pas dans la requête est considéré comme une traduction potentielle. Considérons $\bar{w_x}$ comme l'un des mots candidats :

w_1			w_2	$\bar{w_x}$		w_3	$\bar{w_4}$	w_4
-4	-3	-2	-1	0	1	2	3	4

TABLE 8.5 – Illustration d'un segment donné

Le calcul de la contribution de chaque mot $w_i \in QR$ entourant $\bar{w_x}$ est fait selon l'équation suivante :

$$Contrib(w_i)_{\bar{w_x}} = \frac{|Z|}{D+1} \tag{8.4}$$

où :

$$D = distance(w_i, \bar{w_x}) = |pos(w_i) - pos(\bar{w_x})| \tag{8.5}$$

$pos(w_i)$ est la position de w_i dans un segment donné, par exemple dans la table 8.5 $pos(w_1) = -4$.

$$\mathrm{Z} = \{\mathrm{Y} - distance(Y, \bar{w_x}) < D \quad et \quad Y \in QR\} \cup \{\bar{w_x}\} \tag{8.6}$$

Prenons comme mot : w_1, sa contribution est calculée de la manière suivante :

$$Contrib(w_1)_{\bar{w_x}} = \frac{2+1}{4+1} = \frac{3}{5} \tag{8.7}$$

Nous venons de donner l'équation de la compacité d'un mot pour un segment donné. Puisqu'il y a plusieurs segments dans un corpus, deux possibilités sont à prendre en compte, à savoir considérer le meilleur segment comme le montre l'équation 8.8, ou alors considérer la totalité des segments selon l'équation 8.9.

$$Compact(\bar{w_x}) = \max_s Compact_s(\bar{w_x}) \tag{8.8}$$

$$Compact(\bar{w_x}) = \sum_s Compact_s(\bar{w_x}) \tag{8.9}$$

Nous pouvons aussi considérer les n meilleurs segments. Il est a priori difficile de juger de la pertinence d'un segment d'une manière automatique, nous avons donc fait le choix de considérer le degré de pertinence d'un segment par rapport au nombre de mots de la requête qu'il contient. Ainsi, plus un segment va contenir de mots de la requête plus il sera pertinent. Ceci nous ramène à appréhender les segments de deux façons différentes, c'est-à-dire soit fixer le nombre minimum de mots de la requête min_{req} qui doivent appartenir à un segment, et dans ce cas éliminer les segments dont le nombre de mots de la requête est inférieur à min_{req}. Soit, pondérer le score de compacité d'un terme par rapport à un segment par la similarité entre la requête et le segment. Nous détaillons ce point dans la section 8.1.4.

8.1.4 Paramètres de *Q-Align*

Plusieurs paramètres sont à considérer afin d'exécuter la méthode *Q-Align* :

1. Taille de la requête source w_{req} ;
2. Nombre de requêtes sources nb_{req} ;
3. Taille des segments cibles w_{seg} ;
4. Meilleur segment Max_{seg} versus somme des segments Som_{seg} ;

5. Meilleure requête Max_{req} versus somme des requêtes Som_{req}.

En plus de ces paramètres, nous introduisons trois autres paramètres de pondération des scores de compacité qui sont :

1. La similarité entre une requête et un segment ;
2. L'IDF des mots de la requête ;
3. L'IDF versus la spécificité des candidats à la traduction.

Pondération par similarité requête/segment

Nous partons de l'idée que si une requête est très similaire à un segment, ceci renforce les chances que la bonne traduction se trouve dans ce segment. Partant de là, nous choisissons de pondérer le score d'un candidat par cette similarité qui peut être définie de différentes manières. Trois mesures de similarité entre une requête et un segment ont été choisies.

Le nombre de mots en commun

$$CDM(req, seg) = |\{w : w \in req \wedge w \in seg\}| \tag{8.10}$$

Le nombre de mots en commun pondérés

$$WCM(req, seg) = \sum_{w \in req \wedge w \in seg} \log(\frac{N}{f_w}) \tag{8.11}$$

avec N qui correspond au nombre de segments et f_w le nombre de segments contenant w.

La mesure du cosinus

$$Cos(req, seg) = \frac{\sum_{w \in req \wedge w \in seg}(TF_{seg,w} \times IDF_{seg,w})}{\sqrt{\sum_{w \in req}(TF_{seg,w})^2 \times \sum_{w \in seg}(IDF_{seg,w})^2}} \tag{8.12}$$

L'IDF doit être considéré en remplaçant la notion de document par celle du segment. L'utilisation du WCM par exemple nous donne le résultat de compacité suivant :

$$Compact_{All}(\bar{w_x}) = \sum_{i \in nbQuery} WCM(req, seg) \times Compact(\bar{w_x})_i \tag{8.13}$$

ISF des mots de la requête

Nous voulons considérer différemment les mots d'une requête traduite. Une manière de faire est de pondérer chaque mot par sa fréquence inverse dans un segment (ISF) par analogie à l'IDF en recherche d'information, en supposant que les mots avec un ISF élevé devraient être plus importants. Ceci peut être représenté par l'équation suivante :

$$Compact_s(\bar{w_x}) = \frac{1}{|WQ|} \sum_{i=\in WQ} ISF(w_i) \times Contrib(w_i)_{\bar{w_x}} \tag{8.14}$$

ISF versus la spécificité des candidats à la traduction

Partant de l'intuition que si un mot est peu fréquent en langue source alors sa traduction devrait être peu fréquente en langue cible, nous faisons le choix de pondérer la compacité des mots peu fréquents par l'ISF. Ceci nous donne :

$$Compact_{All}(\bar{w_x}) = \sum_{i \in nbQuery} ISF(\bar{w_x}) \times Compact(\bar{w_x})_i \tag{8.15}$$

Selon le principe de comparabilité des corpus comparables, nous pouvons émettre l'hypothèse que des termes qui sont en relation de traduction suivent la même distribution. Ainsi, nous pouvons pondérer les candidats à la traduction par le rapport de fréquence entre le terme que l'on cherche à traduire et le terme candidat à la traduction. Ceci n'est autre que la spécificité.

$$Compact_{All}(\bar{w_x}) = \sum_{i \in nbQuery} SPEC(\bar{w_x}) \times Compact(\bar{w_x})_i \tag{8.16}$$

8.2 Étude des différents paramètres de *Q-Align*

Nous présentons dans ce qui suit les résultats des expériences menées sur trois corpus de langue de spécialité, à savoir le corpus du cancer du sein, celui des énergies renouvelables et celui de vulcanologie.

8.2.1 Taille de la requête

Le premier paramètre auquel nous nous intéressons est celui de la taille de la requête source w_{req} comme le montre la figure 8.2.

Nous constatons que la taille des requêtes sources influe sur les performances de *Q-Align*. Notons que des requêtes de petite taille donnent globalement les meilleurs résultats pour les trois corpus. Même si la différence des résultats n'est pas très significative pour des fenêtres allant de 3 à 7.

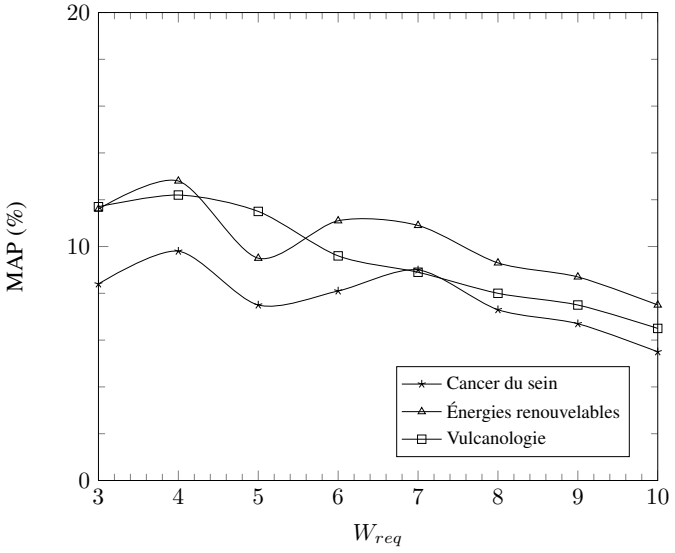


FIGURE 8.2 – Variation de la taille de la requête pour les trois corpus comparables

8.2.2 Nombre de requêtes

Comme deuxième paramètre, nous faisons varier le nombre de requêtes sources du terme à traduire. Ceci est illustré dans la figure 8.3.

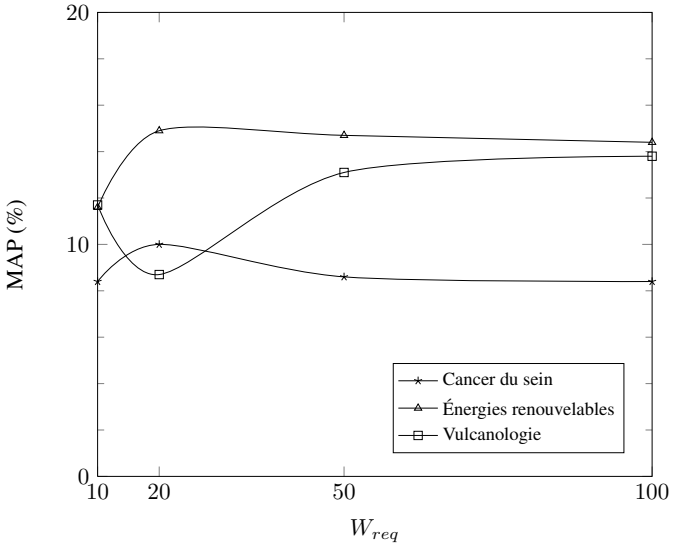


FIGURE 8.3 – Variation du nombre de requêtes source pour les trois corpus comparables

Nous constatons que pour le corpus du cancer du sein, la variation du nombre de requêtes n'a pas une grande influence sur les résultats contrairement aux deux autres corpus. Ceci dit, ces résultats confirment l'importance de choisir des requêtes pertinentes vis-à-vis du terme à traduire. En effet, nous avons pu constater lors des expériences qu'une seule requête pouvait changer fortement le rang d'une traduction

candidate. Ainsi, *Q-Align* est très sensible au choix des requêtes ce qui peut être un inconvénient majeur de cette méthode.

8.2.3 Taille des segments cibles

Le troisième paramètre que nous étudions est la taille des segments cibles.

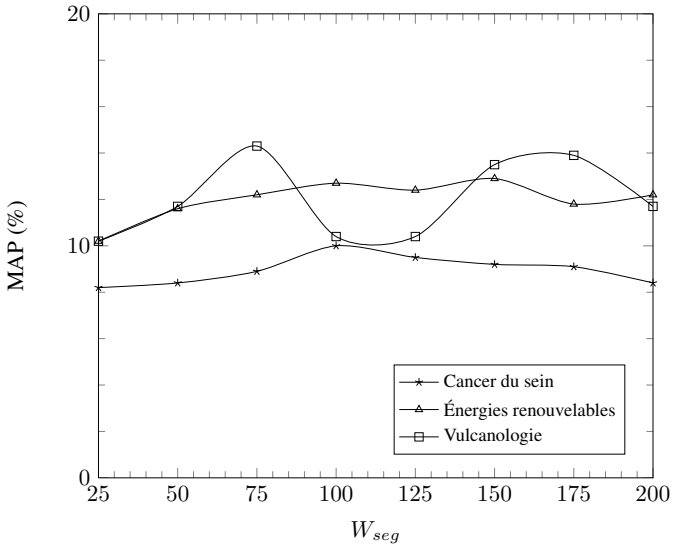


FIGURE 8.4 – Variation de la taille des segments pour les trois corpus comparables

Nous remarquons à travers la figure 8.4 que le corpus de vulcanologie est le plus sensible à la variation de la taille des segments. Les deux autres corpus montrent globalement des variations non significatives. Il est à noter qu'un segment de taille égale à 75 semble être un bon compromis pour les trois corpus. Cette taille coïncide plus ou moins avec la taille d'un paragraphe.

8.2.4 Calcul et pondération de la compacité

Nous nous intéressons ici aux différentes manières de calculer et de pondérer les actants (mots de la requête, candidats à la traduction, segments, etc.) de la méthode *Q-Align*. La table 8.6 résume les différents paramètres.

La première partie de la table 8.6 montre quatre manières de calculer le score final de compacité. Som^{qs} veut dire que le score final d'un candidat à la traduction correspond à la somme de ses scores sur tous les segments et toutes les requêtes. Som^{q} correspond à la somme des segments en ne gardant que la meilleure requête. Som^{s} quant à lui, correspond à choisir le meilleur segment pour chaque requête. Et enfin, Max^{qs} correspond au choix du meilleur segment et de la meilleure requête. Les résultats montrent que la meilleure configuration est celle de Som^{s} pour les trois corpus. Ceci coïncide avec notre idée de choisir le meilleur passage qui contient la bonne traduction. Ainsi, à chaque requête source va être associé le passage (segment) qui lui est le plus proche. Celui-ci, d'après notre hypothèse devrait contenir la bonne traduction.

	Cancer du sein			Énergies renouvelables			Vulcanologie		
	P1	P10	MAP	P1	P10	MAP	P1	P10	MAP
Som^{qs}	02,49	12,15	05,80	02,66	19,33	07,70	06.32	15,82	09,30
Som^{q}	01,24	08,10	03,40	00,66	12,00	04,30	02.53	08,86	04,60
Som^{s}	**04,98**	**15,57**	**08,40**	**06,00**	**24,00**	**11,60**	**08,22**	**22,15**	**11,70**
Max^{qs}	00,93	08,72	03,60	01,33	10,66	04,20	01,89	07,59	04,30
None	04,98	15,57	08,40	06,00	24,00	11,60	**08,22**	**22,15**	11,70
CDM	04,98	15,57	08,50	06,66	24,66	12,10	07,59	20,25	11,80
WCM	**05,60**	**17,13**	**09,20**	**07,33**	**26,00**	**12,80**	07,59	**22,15**	**12,70**
None+	**05,91**	**17,75**	10,30	06,00	26.00	12,50	08,22	22,15	12,70
CDM+	05,60	**17,75**	09,90	**08,00**	26,00	**13,80**	**08,96**	**22,78**	**13,50**
WCM+	05,60	**17,75**	**10,40**	**08,00**	**26,66**	13,70	07,22	20,88	12,90
SPEC	**10,60**	**17,75**	**12,90**	06,00	22,66	11,60	**08,22**	**30,38**	**15,20**
ISF	08,60	**17,75**	11,40	**07,33**	**26,66**	**13,40**	06,96	21,51	12,80

TABLE 8.6 – Comparaison de différents paramètres de la méthode *Q-Align*

La deuxième partie de la table 8.6 montre les différentes mesures de similarité entre une requête et un segment. *None* correspond à l'utilisation de *Q-Align* sans pondération par une mesure de similarité. Nous constatons que la pondération WCM est celle qui donne les meilleurs résultats. Ceci montre l'intérêt de l'utilisation de la similarité requête/segment comme moyen de pondération.

La troisième partie de la table 8.6 montre la combinaison des mesures de similarité avec l'ISF des mots de la requête. Nous pouvons constater une augmentation des résultats pour toutes les configurations et ceci pour les trois corpus. Ainsi, considérer différemment les mots de la requête est un paramètre à prendre en compte dans la configuration finale de Q-Align.

La quatrième et dernière partie de la table 8.6 présente deux manières de pondérer les candidats à la traduction, soit par l'ISF soit par l'indice de spécificité. Ce paramètre est indépendant et peut être appliqué à n'importe quelle méthode d'alignement de termes. Nous constatons que les résultats diffèrent d'un corpus à l'autre. En fait, les résultats ne dépendent pas des corpus mais des listes d'évaluation. Si les couples de traductions ont une distribution similaire, c'est donc la spécificité qui devrait être privilégiée, ce qui semble être le cas pour le corpus de vulcanologie avec une MAP de 15,20 %.

8.3 Synthèse

Vu la multitude de paramètres nécessaires à la méthode *Q-Align*, il est difficile de déterminer la meilleure configuration, d'autant plus qu'il y a des paramètres in-

terdépendants comme par exemple la taille du segment et la taille de la requête. En effet, on aurait tendance à choisir un segment de petite taille pour une petite requête et inversement, même si c'est difficilement vérifiable pour chaque terme à traduire. Nous avons pu constater aussi l'importance du choix des requêtes sources. Le fait d'en prendre beaucoup ne garantit pas forcément un résultat positif et peut au contraire dégrader les performances de *Q-Align*, d'où la nécessité d'investir plus d'efforts dans le traitement des requêtes sources. Une analyse sémantique serait sûrement un plus. Enfin, d'après les différents résultats obtenus, la meilleure configuration semble être $Som^s + WCM + (SPEC$ ou $ISF)$.

8.4 Bilan

Nous avons proposé dans ce chapitre une autre manière d'aborder le tâche d'extraction de lexiques bilingues à partir de corpus comparables, inspirée des systèmes de question/réponse. La méthode *Q-Align* se base sur une représentation locale du contexte facilitant l'extraction de passages susceptibles de contenir la bonne traduction et peut ainsi servir d'outil d'aide à la traduction en proposant au traducteur le passage illustrant le candidat à la traduction. Un inconvénient serait la multitude de paramètres à définir, mais ceci devrait être moins problématique si des informations linguistiques venaient à être ajoutées à *Q-Align*. Comme par exemple, utiliser des paragraphes au lieu de fixer empiriquement la taille des segments ou encore la taille des requêtes.

9

Espace vectoriel, extraction lexicale bilingue et corpus comparables

Introduction

Le but de ce chapitre est d'explorer différents espaces vectoriels. Notre première motivation derrière l'idée du changement d'espace de représentation des mots du corpus vient de l'observation concernant l'orthogonalité des vecteurs bases qui n'est pas vérifiée. Rappelons que l'indépendance (orthogonalité) des vecteurs bases d'un espace vectoriel garantit la non redondance de l'information contenue dans chaque vecteur base, et permet ainsi une représentation plus discriminante des données. Cette propriété essentielle n'étant pas respectée dans l'*approche directe*, nous voulions dans un premier temps vérifier si ceci avait un impact sur l'extraction terminologique bilingue, et dans un second temps sous réserve de résultats positifs, étudier les différentes transformations mathématiques existantes, qui ont déjà prouvé leur efficacité dans plusieurs autres domaines telles que : l'analyse en composantes principales (PCA) qui est largement utilisée en reconnaissance faciale ainsi que dans la compression de données, la génétique, etc. ; l'analyse en composantes indépendantes (ICA) qui est utilisée dans plusieurs applications dont la plus connue reste la résolution du problème de la séparation de sources, illustré par le problème de la soirée cocktail ; et l'analyse sémantique latente (LSA) principalement utilisée en recherche d'information, etc.

Construire des espaces vectoriels à partir de données textuelles n'est pas nouveau. Une attention particulière leur a été portée dès les années 1990 (LSA [Deerwester *et al.*, 1990,Landauer et Dumais, 1997]) dans divers domaines d'application en traitement automatique des langues tels que : la recherche d'information [Dumais *et al.*, 1988,Salton *et al.*, 1975], la désambiguïsation de sens [Schütze, 1993,Schütze, 1998], le classement des mots [McCarthy *et al.*, 2004], la segmentation et la catégorisation de textes [Choi *et al.*, 2001,Sahlgren et Cöster, 2004], la correction d'orthographe en contexte [Jones et Martin, 1997], l'extraction de thésaurus [Grefenstette, 1994b,Lin, 1998a], etc. Ainsi qu'en sciences cognitives dans différentes études de simulation du comportement humain comme par exemple : la similarité de jugements [Mc-

Donald, 2000], l'amorçage et différents tests de connaissances sémantiques [Lund et Burgess, 1996, Landauer et Dumais, 1997, Lowe et Mcdonald, 2000, McDonald et Brew, 2004, Lund *et al.*, 1995, Karlgren et Sahlgren, 2001], mais aussi la compréhension de textes [Landauer et Dumais, 1997, Foltz *et al.*, 1998], etc.

La popularité de la représentation en espace vectoriel des mots vient de la capacité à construire un modèle sémantique simplement en utilisant le principe de la statistique distributionnelle, qui ne nécessite aucune connaissance linguistique ou sémantique au préalable. Si la manière standard de construire un espace vectoriel à partir de données textuelles consiste en une représentation matricielle de type : mots-mots ou termes-documents, des transformations mathématiques plus sophistiquées, présentant des propriétés plus intéressantes, ont fait évoluer cette représentation vers des modèles plus pratiques et beaucoup plus efficaces pour certaines tâches comme la LSA ou la PCA par exemple. Et bien que notre principale motivation soit l'amélioration de la représentation des données, il y a d'autres raisons justifiant l'intérêt des chercheurs pour ces techniques. La première motivation, et non des moindres, est la réduction des dimensions de l'espace vectoriel. En effet, dans la représentation standard en espace vectoriel des mots, la taille des vecteurs correspond au nombre de mots de leurs contextes, et le nombre de dimensions au nombre de mots du vocabulaire. Imaginons un corpus de plusieurs milliers de mots distincts, ceci reviendrait à traiter des milliers de vecteurs ce qui est très coûteux en termes de temps de traitement, et cela même avec les capacités actuelles de calcul. Une deuxième motivation découle du phénomène de la loi de Zipf [Zipf, 1949]. En effet, la majorité des mots occurrent dans un nombre limité de contextes. Cela engendre une représentation en espace vectoriel avec des vecteurs bases de faible densité, c'est-à-dire contenant beaucoup de zéros.

En ce qui concerne l'alignement bilingue, les premiers travaux de [Gaussier *et al.*, 2004] ont montré que l'utilisation de la LSA, de la CCA et de la PLSA n'amélioraient pas les performances de l'alignement de mots. Ceci était dû pour la PLSA à l'utilisation des cooccurrences des mots comme mesure d'association, ce qui n'est peut être pas suffisant comparé à d'autres mesures plus efficaces telles que l'information mutuelle ou le taux de vraisemblance. Concernant la CCA, les auteurs expliquent que ses résultats insuffisants étaient dus au bruit que pouvait modéliser la CCA dont la représentation des directions canoniques en combinaison linéaire avait tendance à emmagasiner différents mots du vocabulaire considérés comme du bruit. La CCA a été utilisée en 2008 par [Haghighi *et al.*, 2008] qui ont montré que sous certaines conditions, celle-ci était efficace pour l'alignement bilingue. D'autres travaux plus récents de [Rubino et Linarès, 2011] ont montré des résultats encourageants en utilisant l'analyse de Dirichlet latente (LDA). Ils ont montré que la modélisation de thématiques par la LDA en plus de l'exploitation du contexte classique améliorait les performances de leur système d'alignement.

9.1 Représentation géométrique

L'*approche directe* est basée sur une représentation vectorielle standard des mots. Cette représentation est faite suivant les bases de l'espace euclidien. Nous nous basons dans cette étude sur le travail fait par [Gaussier *et al.*, 2004] concernant

cette question.

Soit s_i, $1 \leq i \leq p$ et t_j, $1 \leq j \leq q$ les mots sources et cibles appartenant au dictionnaire bilingue D, avec D qui représente l'ensemble des n paires de traductions (s_i, t_j). L'ensemble D peut être représenté par une matrice M de taille $p \times q$ où $M_{ij} = 1$ si $(s_i, t_j) \in D$ et $M_{ij} = 0$ autrement. Étant donné m mots distincts en langue source $e_1, e_2, ...e_m$ et r mots distincts en langue cible $f_1, f_2, ...f_r$, la mesure d'association $a(v, e)$ peut être vue comme les coordonnées du vecteur de contexte $\overrightarrow{v}$ à m dimensions dans l'espace vectoriel formé par les vecteurs bases orthogonaux $(e_1, e_2, ...e_m)$. De la même manière, la mesure d'association $a(w, f)$ peut être vue comme les coordonnées du vecteur de contexte $\overrightarrow{w}$ à r dimensions dans l'espace vectoriel formé par les vecteurs bases orthogonaux $(f_1, f_2, ...f_m)$. Le seul lien qui relie le corpus source au corpus cible est le dictionnaire bilingue D, ceci conduit à une projection des mots de la langue source (respectivement la langue cible) dans un sous espace vectoriel formé uniquement par les mots de langue source appartenant au dictionnaire. Soit une matrice de projection P_s de taille $p \times m$. Ce nouvel espace vectoriel est représenté par les vecteurs bases orthogonaux $(s_1, s_2, ...s_p)$ pour la langue source et $(t_1, t_2, ...t_q)$ pour la langue cible avec une matrice de projection P_t de dimensions $q \times r$.

Sachant que la similarité entre un mot source v et un mot cible w se fait grâce à des mesures comme le cosinus, les coefficients de Dice ou du Jaccard, et que le produit scalaire joue un rôle essentiel dans toutes ces mesures, nous considérons la similarité donnée par le produit scalaire entre $\overrightarrow{v}$ et le vecteur traduit de $\overrightarrow{w}$ comme suit :

$$\langle \overrightarrow{v}, \overrightarrow{tr(w)} \rangle = \sum_e a(v, e) \sum_{f,(e,f)\in D} a(w, f) = \sum_{(e,f)\in D} a(v, e)a(w, f) \tag{9.1}$$

Puisque M encode les relations entre les mots sources et cibles appartenant au dictionnaire bilingue D, l'équation précédente peut être réécrite comme suit :

$$S(v, w) = \langle \overrightarrow{v}, \overrightarrow{tr(w)} \rangle = (P_s \overrightarrow{v})^T M (P_T \overrightarrow{w}) \tag{9.2}$$

où T dénote la transposée. Nous pouvons aussi noter que la matrice M peut être représentée comme $S^T T$, avec S une matrice de dimensions $n \times p$ et T une matrice de dimensions $n \times q$ encodant les relations entre les couples de mots du dictionnaire bilingue D. Ainsi :

$$S(v, w) = \overrightarrow{v}^T P_s^T S^T T P_T \overrightarrow{w} = \langle S P_s \overrightarrow{v}, T P_T \overrightarrow{w} \rangle \tag{9.3}$$

Ce qui montre que l'application de l'*approche directe* revient à effectuer un produit scalaire dans un espace vectoriel formé par les n couples de traductions $((s_1, t_l), ..., (s_p, t_k))$ qui sont supposés être orthogonaux. Cette supposition n'est jamais vérifiée en réalité. De plus, elle n'est pas possible en pratique. C'est pourquoi, nous nous sommes tournés vers des transformations mathématiques qui permettent entre autres d'assurer l'orthogonalité des vecteurs bases.

9.2 Analyse en composantes indépendantes (ICA)

L'analyse en composantes indépendantes (ICA en anglais) est une transformation mathématique utilisée pour découvrir des informations dites cachées à partir d'observations faites sur un ensemble de données. L'ICA s'assure que ces données sont statistiquement indépendantes. Dans la version classique de l'ICA [Jutten et Hérault, 1991, Comon, 1994, Hyvarinen *et al.*, 2001], chaque observation $x = (x_1, x_2, ..., x_n)^T$ est représentée comme une somme pondérée de variables aléatoires indépendantes $s = (s_1, ..., s_k, ..., x_n)^T$, telle que :

$$x = AS \tag{9.4}$$

où A représente la matrice qui contient les poids qui sont supposés être différents pour chaque variable observée, et S le vecteur des composantes indépendantes. Notons les colonnes de A par a_i, alors le modèle peut être défini comme suit :

$$x = \sum_{i=1}^{D} a_i S_i \tag{9.5}$$

Le modèle statistique défini par l'équation 9.5 est appelé le modèle ICA qui décrit comment sont générées les données observées, et ceci par un processus de mélange des composantes S_i. Les deux matrices de mélange A et de composantes indépendantes S sont apprises d'une manière non supervisée à partir des données observées X. L'ICA émet l'hypothèse que les composantes S_i sont statistiquement indépendantes. L'ICA peut être vue comme une extension de l'analyse en composantes principales (PCA) et l'analyse factorielle (factor analysis). La principale différence entre l'ICA et la PCA, est qu'alors que la PCA cherche les projections avec un maximum de variance, l'ICA cherche les projections qui sont statistiquement indépendantes (non-gaussiennes). La PCA est utile comme étape préalable pour réduire les dimensions des données avec un minimum d'erreur. Le but de l'ICA par contre, n'est pas de réduire les dimensions de l'espace vectoriel. Pour notre analyse, nous utilisons la bibliothèque *FastICA* fournie par [Hyvarinen, 1999] où la matrice de données x est considérée comme une combinaison linéaire de composantes indépendantes. Les colonnes de S contiennent les composantes indépendantes et A représente la matrice de mélange linéaire. Les dimensions des données sont d'abord réduites par la PCA. Après la phase de normalisation de la variance (the whitened data), n composantes indépendantes sont extraites grâce à l'ICA.

9.3 Méthode

Nous proposons d'évaluer les transformations mathématiques (LSA, PCA et ICA) d'abord individuellement, puis de manière conjointe, c'est-à-dire en combinant les

sorties (scores) de chaque approche. Pour chaque transformation, notre méthode consiste en l'utilisation d'un nouvel espace de représentation des données qui soit le plus discriminant possible. Notre choix c'est d'abord porté sur l'ICA [1] pour ses propriétés citées plus haut. À titre comparatif, nous évaluons aussi la PCA et la LSA.

9.3.1 Représentation des données

Dans notre cas, les données observées x sont une matrice de mots $M \times N$ où les colonnes représentent les contextes et les lignes représentent les mots. Les N mots de la langue cible qui apparaissent dans le dictionnaire bilingue sont sélectionnés pour construire la matrice X. Chaque colonne de la matrice X représente un vecteur de contexte d'un mot i avec $i \in N$. Étant donné un élément X_{cr} de la matrice X, X_{cr} représente la mesure d'association entre le mot de la r :ième ligne et le contexte de la c :ième colonne. Les mesures d'association utilisées peuvent être l'information mutuelle, le taux de vraisemblance, etc.

		Variables				
		mot_1	mot_2	mot_3	...	mot_n
Échantillons	$(word_1, mot_1)$	0	0	65		1
	$(word_2, mot_2)$	2	0	0		3
	$(word_3, mot_3)$	6	0	0		45
	$(word_4, mot_4)$	0	0	0		7
	$(word_5, mot_5)$	15	2	32		280
	⋮	⋮	⋮	⋮	⋮	⋮
	$(word_m, mot_n)$	2	1	1		0

TABLE 9.1 – Représentation de la matrice des données

La représentation des données par l'ICA (LSA ou PCA) a pour but de construire un espace s de données où tous les mots du corpus présents dans le dictionnaire sont utilisés. Ainsi chaque composante indépendante s_k va contenir ou encoder une certaine quantité d'information extraite des N mots de la langue cible. Pour chaque méthode (ICA, LSA et PCA), nous utilisons les mêmes étapes de caractérisation du contexte et de transfert de vecteur que l'*approche directe*. La principale différence réside dans la construction d'un nouvel espace vectoriel en utilisant l'ICA (LSA ou PCA) qui transforme la matrice X en un nouvel espace de composantes indépendantes $s = (s_1, ..., s_k, ..., x_n)^T$. La matrice X peut être vue comme la concaténation des N vecteurs de contextes des mots de la langue cible présents dans le dictionnaire bilingue.

1. Cette représentation offre un double intérêt. Les propriétés mathématiques de l'ICA assurent une meilleure représentation des données et l'utilisation de la PCA comme étape préalable fournit une réduction de dimension très utile lorsqu'il s'agit de traiter d'énormes quantités de données.

9.3.2 Projection des mots

Une fois l'espace vectoriel S construit, les vecteurs de contexte traduits des mots candidats sont projetés dans le nouvel espace. Soit $\mathbf{i}$ le vecteur de contexte du mot i. La projection de $\mathbf{i}$ notée $\mathbf{i}_p$ est définie comme suit :

$$\mathbf{i}_p = \mathbf{i}^T \times S \tag{9.6}$$

9.3.3 Mesure de distance

Comme pour l'*approche directe*, les candidats à la traduction d'un mot donné sont les mots ordonnés selon un score de similarité ou de dissimilarité. Ici nous n'utiliserons que la mesure de dissimilarité ou de distance. La mesure choisie n'est autre que la distance euclidienne normalisée appelée aussi "Chord distance [Korenius *et al.*, 2006]". La distance euclidienne normalisée est définie par l'équation suivante :

$$d(\mathbf{i},\mathbf{j}) = \sqrt{\sum_{k=1}^{n} (\frac{\mathbf{i_k}}{||\mathbf{i}||} - \frac{\mathbf{j_k}}{||\mathbf{j}||})^2} \tag{9.7}$$

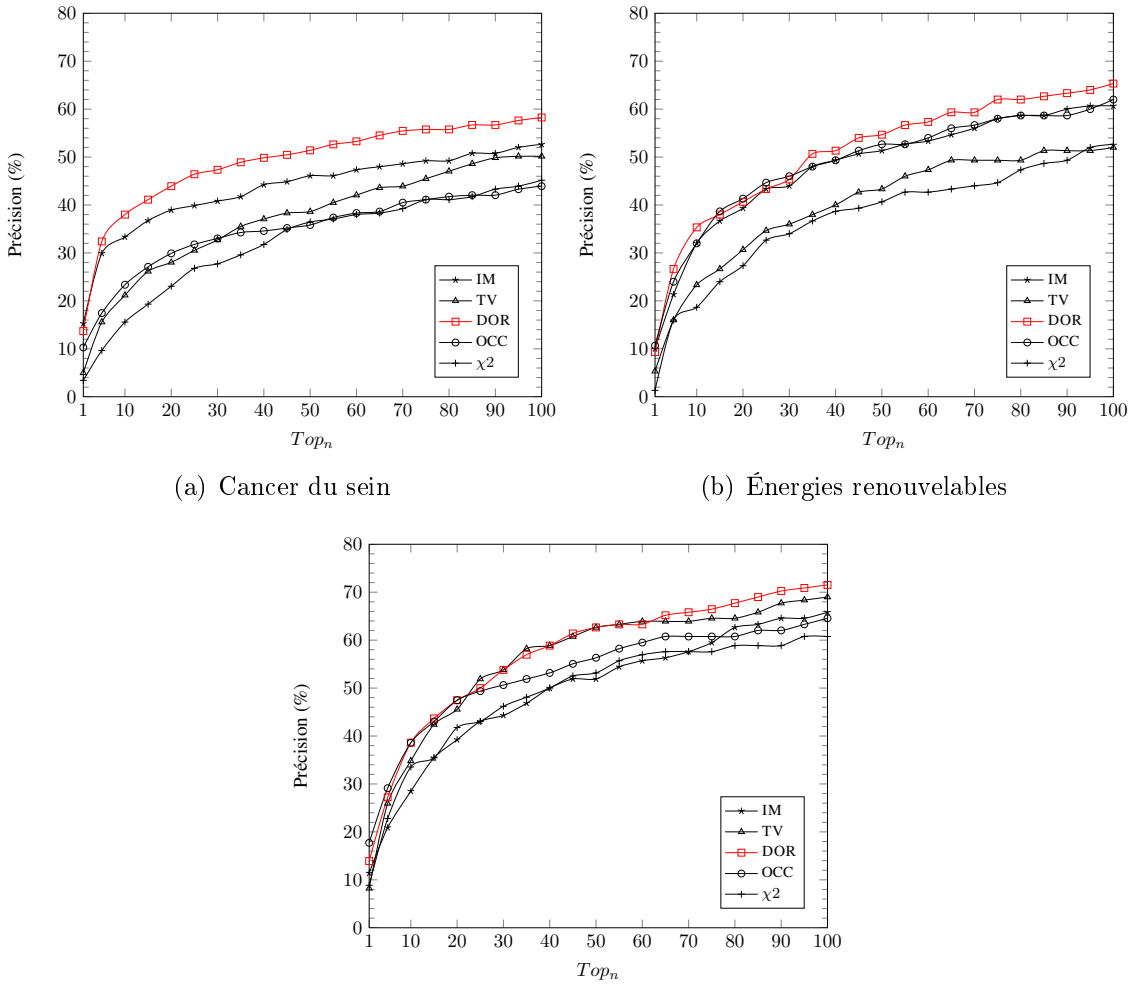
9.4 Évaluation

Nous présentons dans cette section les résultats des différentes transformations mathématiques traitées tout au long de ce chapitre. Dans la première expérience, chaque méthode sera évaluée individuellement afin de définir ses meilleurs paramètres pour ensuite faire une comparaison de ces dernières entre elles et définir la transformation la plus adaptée à notre tâche d'extraction terminologique bilingue à partir de corpus comparables.

9.4.1 Comparaison des mesures d'association

Dans cette première expérience nous faisant varier différentes mesures d'associations pour chaque transformation mathématique. Le but étant de déterminer la mesure d'association la plus adéquate.

La figure 9.1 montre les résultats obtenus sur les trois corpus en faisant varier différentes mesures d'association sur la LSA. Nous constatons que la meilleure mesure est le *discounted odds-ratio* sur les trois corpus. Nous pouvons aussi remarquer que pour le corpus de vulcanologie le *taux de vraisemblance* est très proche du *discounted odds-ratio*. Nous retiendrons ceci dit le *discounted odds-ratio* qui semble être la mesure la plus stable.

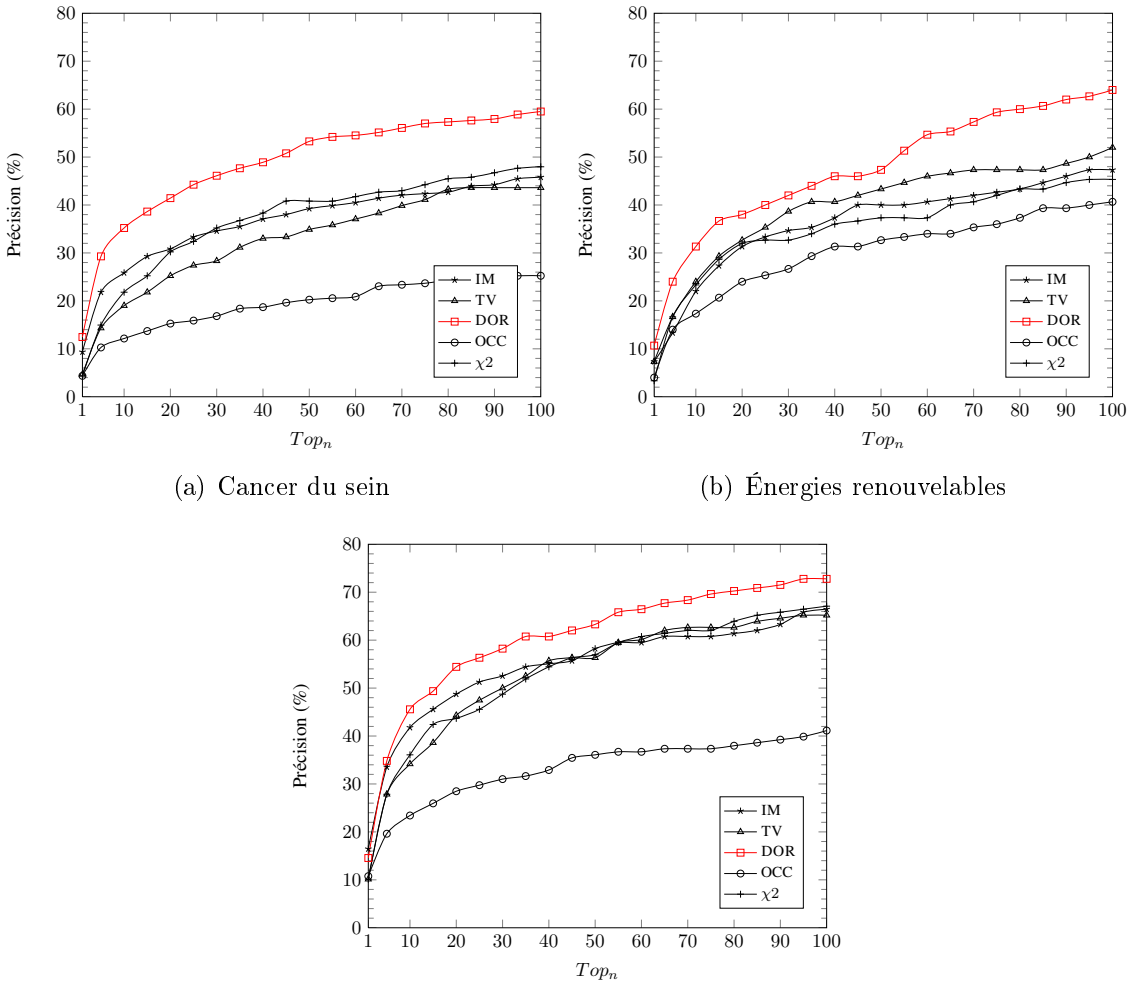


(a) Cancer du sein

(b) Énergies renouvelables

(c) Vulcanologie

FIGURE 9.1 – Comparaison des mesures d'association pour la LSA

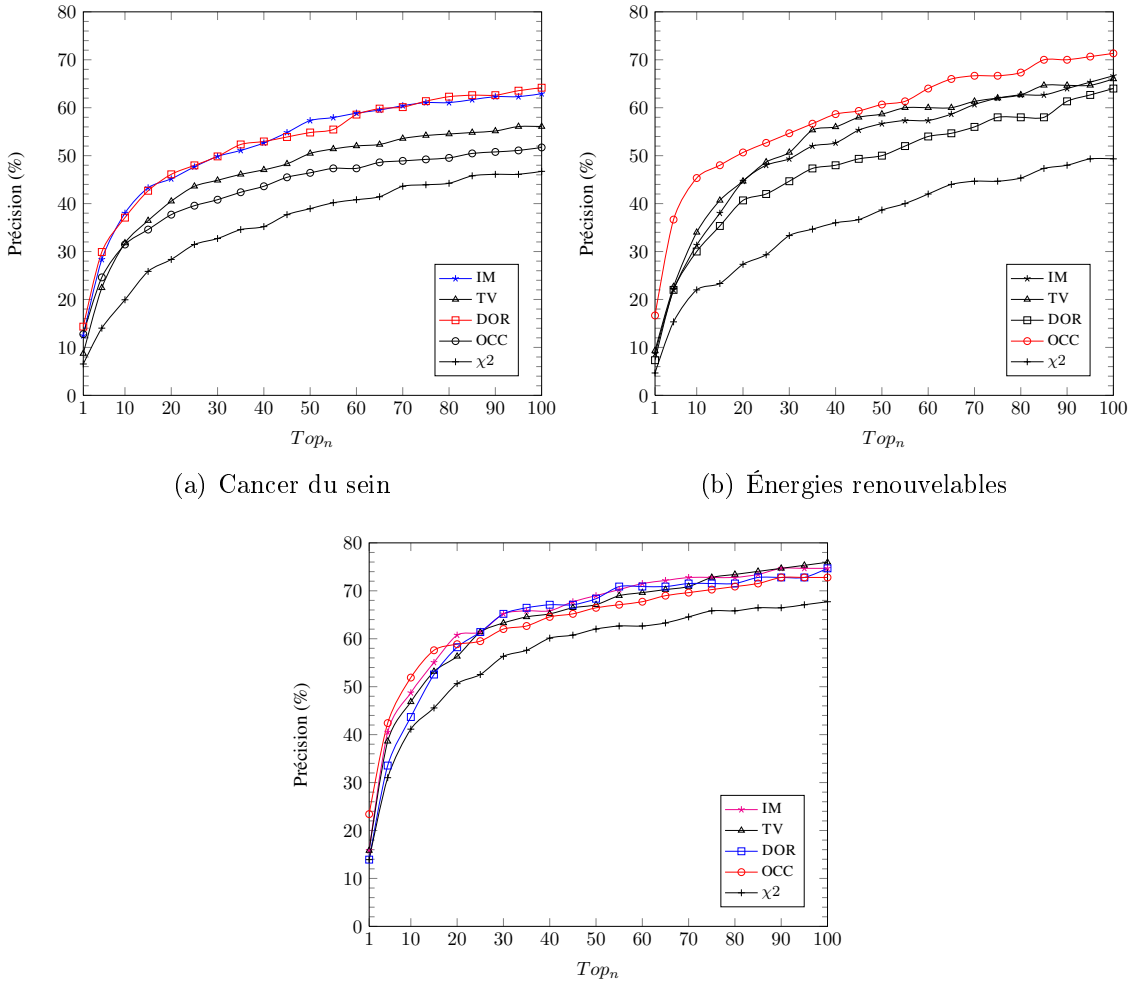


(a) Cancer du sein

(b) Énergies renouvelables

(c) Vulcanologie

FIGURE 9.2 – Comparaison des mesures d'association pour la PCA



(a) Cancer du sein

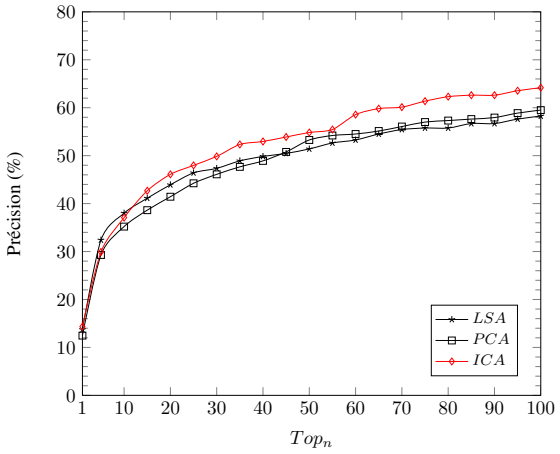
(b) Énergies renouvelables

(c) Vulcanologie

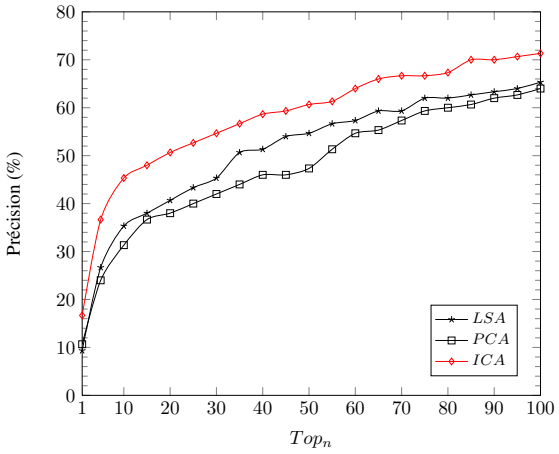
FIGURE 9.3 – Comparaison des mesures d'association pour l'ICA

La figure 9.2 montre les résultats des différentes mesures d'association en utilisant la PCA. Nous pouvons aussi constater que la meilleure mesure est le *discounted odds-ratio*. La différence est plus nette sur les trois corpus. Ainsi, nous retiendrons aussi la mesure du *discounted odds-ratio* comme la plus adéquate pour la PCA.

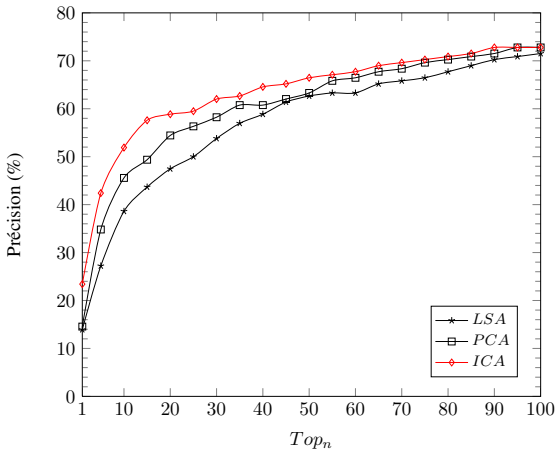
La figure 9.3 montre les résultats des différentes mesures d'association en utilisant l'ICA. Contrairement aux deux figures précédentes, les résultats de l'ICA par rapport aux mesures d'association varient selon les corpus. En effet, les meilleures configurations pour le corpus du cancer du sein sont le DOR et l'IM, alors que pour le corpus des énergies renouvelables, c'est la mesure Occ qui montre les meilleurs résultats. Enfin, concernant le corpus de vulcanologie, les résultats sont plus variables avec un léger avantage pour Occ dans les premiers Tops. Par la suite, c'est le DOR et l'IM qui semblent donner les meilleurs résultats.



(a) Cancer du sein



(b) Énergies renouvelables



(c) Vulcanologie

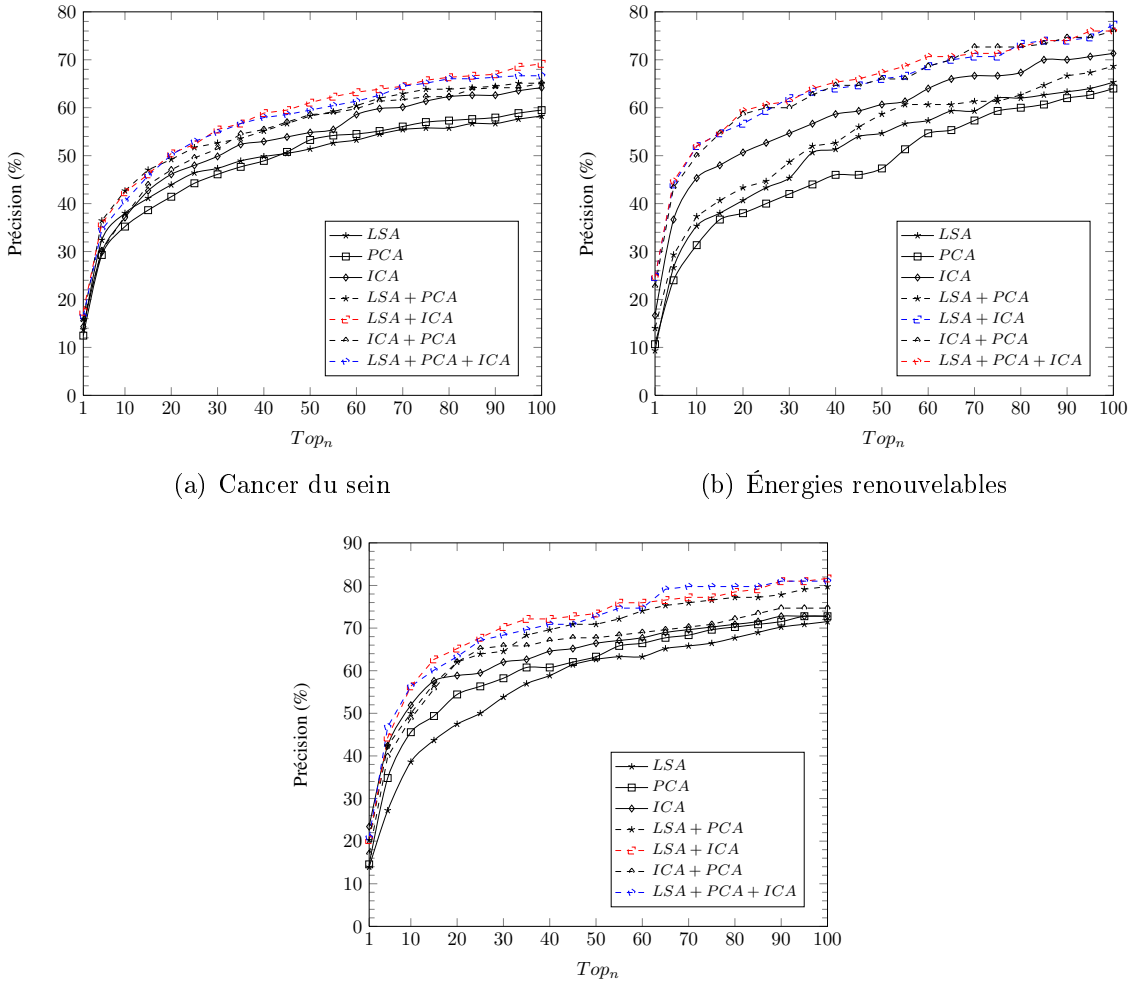
FIGURE 9.4 – Comparaison des transformations mathématiques

9.4.2 Comparaison des transformations mathématiques

Cette expérience a pour but de comparer les différentes transformations mathématiques, à savoir la LSA, la PCA et l'ICA.

La figure 9.4 montre que les meilleurs résultats sont obtenus en utilisant l'ICA. Si la différence des résultats est plus nette pour le corpus des énergies renouvelables, celle-ci reste plus irrégulière pour les deux autres corpus et plus particulièrement pour le corpus du cancer du sein. D'une manière générale, l'ICA semble être la plus adaptée pour l'extraction de lexique bilingue compte tenu de sa meilleure performance sur les trois corpus de spécialité.

Une question naturelle vient à l'esprit : Que donnerait la combinaison des différentes transformations ? Nous tentons de répondre à cette question dans la prochaine expérience.



(a) Cancer du sein

(b) Énergies renouvelables

(c) Vulcanologie

FIGURE 9.5 – Combinaison de transformations mathématiques

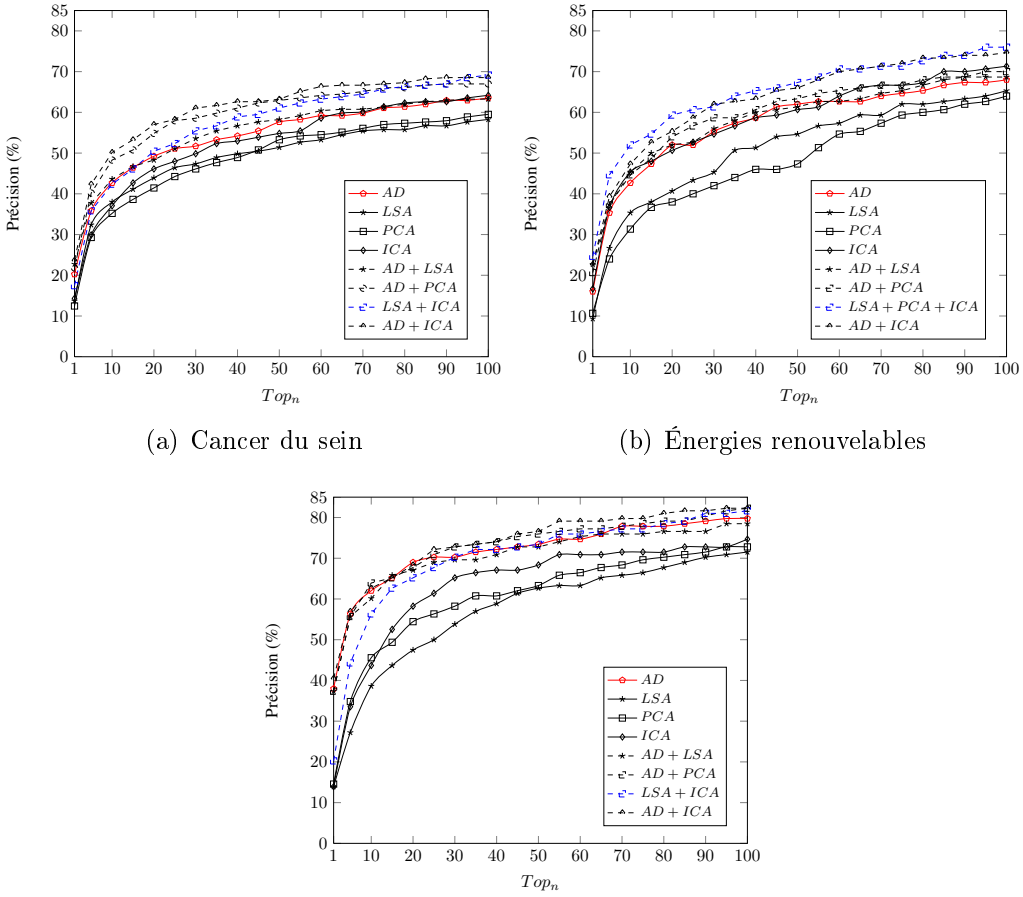
9.4.3 Combinaison des transformations mathématiques

Dans cette expérience nous combinons les différentes transformations mathématiques dans le but de vérifier leur complémentarité. Cette combinaison (a posteriori des résultats) est une moyenne arithmétique des rangs.

La figure 9.5 montre les résultats de la combinaison des transformations mathématiques. D'une manière générale combiner les différentes transformations d'espace de représentation vectorielle améliore les résultats. La meilleure configuration est $LSA + ICA$ pour les corpus du cancer du sein et de vulcanologie. L'apport de la PCA semble être négligeable contrairement au corpus des énergies renouvelables où les meilleurs résultats sont obtenus en combinant les trois modèles.

9.4.4 Apport de l'*approche directe*

Dans la continuité de l'expérience précédente, nous menons une dernière expérience afin de tester la combinaison cette fois de l'*approche directe* avec les autres modèles mathématiques. Sachant que l'*approche directe* se base sur une mesure de



(a) Cancer du sein

(b) Énergies renouvelables

(c) Vulcanologie

FIGURE 9.6 – Combinaison de l'*approche directe* avec les autres transformations mathématiques

similarité et que les autres modèles se basent sur une mesure de distance, nous utilisons la combinaison arithmétique des rangs. Nous comparons les résultats obtenus avec les meilleures configurations de l'expérience précédente.

La figure 9.6 montre les résultats de la combinaison de l'*approche directe* avec les trois transformations mathématiques. Cette figure reprend aussi les modèles séparément ainsi que leur meilleure combinaison. Nous remarquons que l'*approche directe* (SA) est meilleure que les trois modèles pris séparément pour le corpus du cancer du sein bien que suivie de près par l'ICA. Nous pouvons remarquer par ailleurs que la combinaison $LSA+ICA$ donne de meilleurs résultats que la SA. Nous constatons que les meilleurs scores sont obtenus par la combinaison $SA + ICA$ suivie de près par la $SA + PCA$. Concernant le corpus des énergies renouvelables, l'ICA et la SA se superposent avec un avantage global pour l'ICA. Contrairement au corpus du cancer du sein, c'est la combinaison des trois modèles $LSA + PCA + ICA$ qui donne les meilleurs résultats. Nous remarquons aussi que la combinaison $SA+ICA$ se rapproche de $LSA+PCA+ICA$ à partir du Top 30. Enfin, concernant le corpus de vulcanologie, les résultats de la combinaison ne sont pas significatifs. La méthode SA reste globalement meilleure dans les premiers tops. Néanmoins nous constatons de meilleurs scores pour la $SA + ICA$ à partir du top 30.

9.5 Discussion

Dans un cadre théorique, un espace vectoriel dont les vecteurs bases sont orthogonaux constitue un modèle adéquat pour représenter les données. Néanmoins, dans un cadre pratique la plupart des transformations mathématiques dépendent des données initiales. Nous avons pu constater par exemple que la PCA et la LSA obtenaient des résultats en deçà de l'ICA. Ceci laisse à penser qu'elles sont plus sensibles que l'ICA aux données de départ. Ceci étant, le point commun des trois modèles est le choix du nombre de vecteurs bases représentant l'espace vectoriel (vecteurs propres dans le cas de la PCA). Ce paramètre influe grandement sur la performance de ces modèles. On parle de variables et d'échantillons. Les variables représentent les vecteurs bases et les échantillons représentent les entrées de chaque vecteur base. L'une des difficultés de ces modèles est donc le choix du nombre de variables et du nombre d'échantillons. Dans notre cas, les variables et les échantillons sont les mots du dictionnaire présents dans la langue cible. Ceci implique une plus forte dépendance de ces modèles à la couverture du dictionnaire que l'*approche directe*. La principale question non résolue dans ce chapitre reste le choix des variables dites discriminantes. Étant dans un domaine de spécialité nous pourrions nous orienter vers la terminologie du domaine pour choisir les variables. Cette piste a été explorée sans succès. Il semble que sélectionner tous les mots du vocabulaire reste le choix le plus approprié. Nous restons convaincus qu'il y a des mots à éliminer car non porteurs de sens.

Nous avons pu constater que la combinaison des modèles apportait un gain dans la plupart des cas. Ceci étant la meilleure performance reste celle obtenue en combinant l'*approche directe* avec l'ICA pour les corpus du cancer du sein et de vulcanologie alors que c'est la combinaison des trois modèles $LSA + PCA + ICA$ qui donne les meilleurs résultats pour le corpus énergies renouvelables. Ici aussi nous remarquons des différences de performance selon les corpus ce qui une fois encore confirme la dépendance des modèles aux données d'entrée.

9.6 Bilan

Nous avons présenté et évalué différentes transformations mathématiques sur trois corpus de spécialité. Les résultats obtenus ont montré un avantage certain pour l'analyse en composantes indépendantes (ICA) en comparaison avec la PCA et la LSA. Les meilleures performances obtenues lors des expériences grâce à la combinaison de modèles nous amènent à définir un modèle d'extraction terminologique bilingue multi-sources constitué de l'*approche directe* ainsi que de la LSA, la PCA et la ICA.

10

Vers un système multi-sources

Introduction

Les différents résultats expérimentaux ont montré des avantages et des inconvénients pour toutes les méthodes étudiées ou proposées durant ces travaux. Ce constat nous amène à la conclusion qu'il est difficile de concevoir une approche qui couvre tous les cas de figure de notre problématique d'extraction bilingue. Nous avons toutefois pu remarquer que la combinaison d'approches pouvait améliorer significativement les performances. Nous proposons dans ce chapitre une architecture englobant un ensemble de modules nécessaires à notre tâche. Cette architecture représente notre premier prototype d'un système multi-sources pour l'alignement bilingue à partir de corpus comparables.

10.1 Architecture (premier prototype)

La figure 10.1 illustre l'architecture de notre système d'alignement multi-sources. Une fois les textes bruts récupérés à partir des corpus source et cible, ils sont d'abord pré-traités pour ensuite servir à la construction des vecteurs de contexte. Ces vecteurs de cooccurrence sont passés en entrée aux différentes méthodes d'alignement. Enfin, les sorties de chaque méthode sont récupérées et passées en entrée au module de fusion qui renvoie la liste des traductions, celle-ci après post-traitements constituera le résultat final du système multi-sources.

Le module de pré-traitements s'occupe de la partie nettoyage et mise en forme des textes bruts. Après filtrage des mots outils, les mots passent par les étapes de tokenisation, de lemmatisation et d'étiquetage morphosyntaxique. Ensuite, les vecteurs de contexte de tous les mots des corpus source et cible sont construis selon trois manières : soit par représentation graphique, soit par représentation syntaxique ou alors par la combinaison des deux représentations contextuelles. Dans notre cas, et au vu des résultats expérimentaux, la combinaison *a priori* des contextes sera

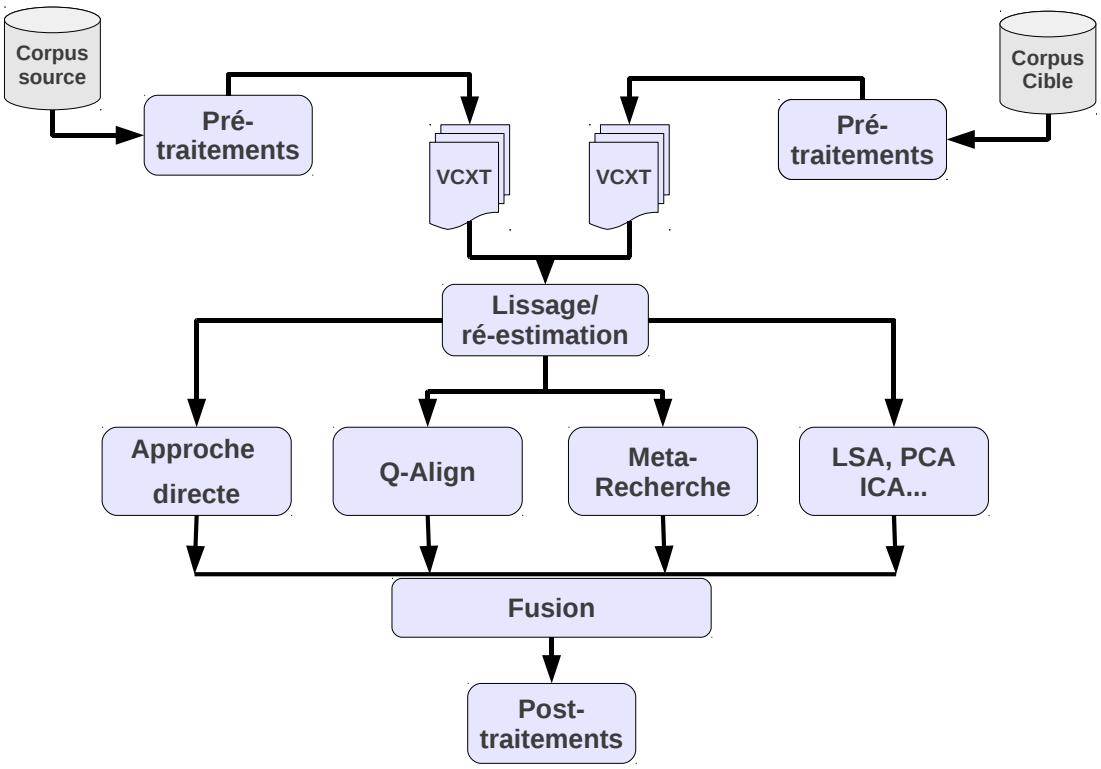


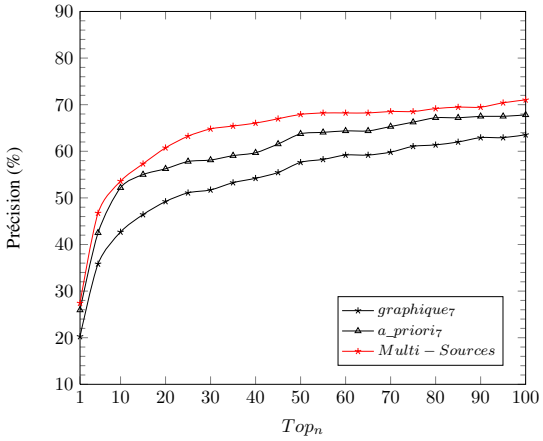
FIGURE 10.1 – Architecture multi-sources pour l'extraction de lexiques bilingues à partir de corpus comparables

privilégiée. Les vecteurs de contexte passent ensuite par le module de ré-estimation des cooccurrences des mots. Les cooccurrences des vecteurs de contexte sont lissées et/ou prédites en fonction des configurations choisies. Une fois les listes des candidats à la traduction renvoyées par les différentes approches, le module de fusion s'occupera de les combiner pour produire une nouvelle liste de candidats. Celle-ci pourra aussi être post-traitée par filtrage des catégories grammaticales ou reclassement selon certains critères comme la spécificité ou la fréquence.

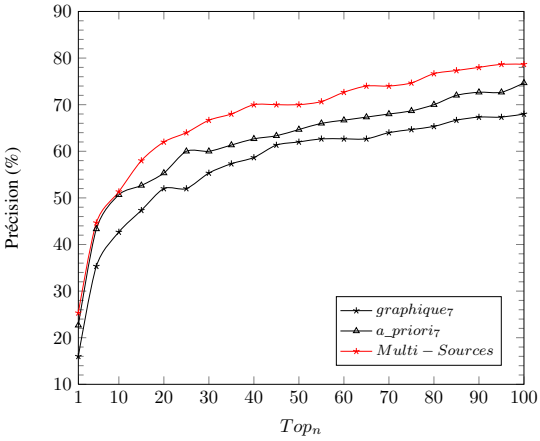
10.2 Évaluation

Nous présentons dans cette section les résultats du système multi-sources. L'évaluation est effectuée en utilisant les mêmes ressources linguistiques que celles des précédentes sections. Nous comparons le système de combinaison multi-sources à l'*approche directe* et à l'approche par combinaison *a priori* des contextes.

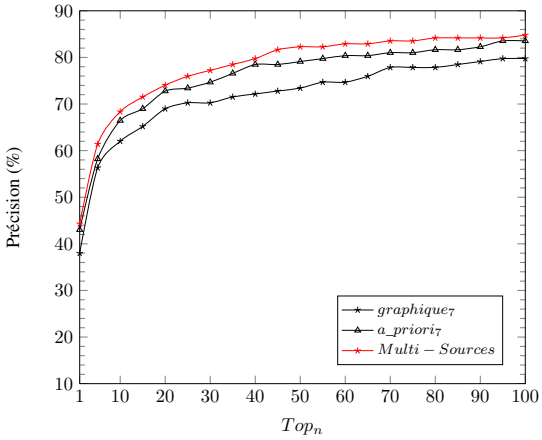
La figure 10.2 illustre les premiers résultats d'une expérience de fusion d'approches sans optimisation particulière. La fusion se base sur une simple combinaison des rangs des mots de chaque liste de candidats renvoyée par une approche donnée. La figure 10.2 montre que le système multi-sources améliore les résultats de l'alignement bilingue sur les trois corpus comparables que ce soit par rapport à



(a) Cancer du sein



(b) Énergies renouvelables



(c) Vulcanologie

FIGURE 10.2 – Premiers résultats du système multi-sources

l'*approche directe* ou par rapport à l'approche par combinaison *a priori* des contextes. L'amélioration est beaucoup plus nette si nous comparons le système multi-sources à l'*approche directe*. Concernant l'approche par combinaison *a priori*, celle-ci fait partie du système multi-sources, elle est représentée ici à titre comparatif.

10.3 Bilan

Nous avons présenté dans ce chapitre le premier prototype d'un système multi-sources basé sur une simple combinaison arithmétique des rangs. Nous avons pu constater que cette approche permettait d'augmenter les performances de l'extraction terminologique bilingue à partir de corpus comparables. Ces premiers résultats nous confortent dans l'idée qu'il est intéressant de combiner les approches proposées et de poursuivre les travaux d'amélioration des différents modules. L'utilisation de techniques de fusion plus évoluées constitue la principale perspective du système multi-sources.

Conclusion et perspectives

Nous nous sommes intéressés dans le présent manuscrit à la tâche de l'extraction terminologique bilingue à partir de corpus comparables en nous concentrant sur le cas des termes simples, considérés comme amorce à d'autres tâches comme par exemple la traduction de termes complexes ou la traduction automatique d'une manière générale. L'étude des travaux antérieurs a soulevé plusieurs questionnements et nous a conduits à nous intéresser principalement aux méthodes distributionnelles qui nous ont paru les plus adaptées à notre problématique. Nous nous sommes volontairement limités à l'exploitation de corpus comparables sans nous appuyer sur d'autres ressources comme par exemple les thesaurus ou les corpus parallèles, par souci d'indépendance vis-à-vis de ressources supplémentaires plus difficiles à obtenir.

Le premier chapitre présente un état de l'art autour de l'extraction lexicale bilingue à partir de corpus comparables. Après un bref retour sur la notion de corpus et la dualité existant entre les corpus parallèles et comparables, nous avons présenté les principales approches traitant le cas des termes simples, à savoir les méthodes distributionnelles et celles dédiées aux termes complexes que sont les méthodes compositionnelles. L'étude des méthodes distributionnelles a souligné l'importance de la caractérisation du contexte des mots et nous a ainsi conduit à lui porter un intérêt particulier. La caractérisation du contexte des mots devient alors le fil conducteur de ces travaux de thèse.

Le deuxième chapitre tourne autour de cette notion de contexte et de ce que nous considérons comme sa plus petite unité, à savoir le *mot* ou le *terme*. À travers l'étude de la littérature nous donnons des définitions traditionnelles et pragmatiques de ces notions pour en garder celle énoncée par [L'Homme, 2004] qui considère que le sens d'un terme est à envisager par rapport à un domaine de spécialité. La question principale qui en est ressortie fait suite à la définition des relations syntagmatiques et paradigmatiques entre les unités lexicales et qui est : comment exploiter au mieux les relations du premier et second ordre pour améliorer la qualité du contexte d'une unité lexicale et ainsi augmenter les performances de l'alignement bilingue ? Nous revenons ensuite sur deux principales représentations contextuelles. La représentation graphique basée sur une fenêtre contextuelle et la représentation syntaxique basée sur les relations de dépendance syntaxique. Notre questionnement revient sur le choix d'une représentation du contexte, en d'autres termes, doit-on choisir entre les deux représentations ou plutôt les combiner ? Nous concluons ce chapitre par

l'introduction de quelques unes des mesures d'association les plus usitées dans le domaine. Là encore se pose la question du choix des mesures les plus appropriées.

Le chapitre trois présente une étude approfondie de l'approche directe. Nous avons discuté plusieurs paramètres et pu constater leur impact sur la qualité de l'alignement. Et notamment le dictionnaire bilingue qui joue le rôle de pont entre les langues concernées. Bien qu'il existe des travaux de recherche visant à s'en passer, le dictionnaire reste une ressource indispensable dans la tâche de l'extraction lexicale bilingue à partir de corpus comparables. Nous avons pu constater que la taille, la qualité du dictionnaire et du corpus comparable influaient d'une façon significative sur les performances des méthodes distributionnelles.

Dans le chapitre quatre et jusqu'à la fin de ce présent manuscrit, nous avons présenté nos contributions que nous pouvons diviser en deux parties bien distinctes. La première concerne les améliorations directement liées à l'approche directe, la deuxième partie quant à elle concerne la proposition de nouvelles méthodes d'alignement. Ces contributions sont à concevoir comme des briques qui feront partie d'un système global plus complexe que nous avons défini comme étant un système multi-sources. Concernant les améliorations de l'approche directe, nous avons commencé par présenter une première méthode de combinaison de contextes qui tire parti des deux principales représentations contextuelles, à savoir la représentation par fenêtre ou sac de mots et celle par relations de dépendance syntaxique. Nous avons pu constater que l'exploitation conjointe des deux représentations améliorait significativement les résultats de l'approche directe. Nous retiendrons plus particulièrement la combinaison *a priori* car moins coûteuse en termes de temps d'exécution. Si l'exploitation de l'information graphique et syntaxique a montré son efficacité, nous pouvons à terme réfléchir à introduire d'autres informations contextuelles afin d'enrichir le contexte.

Nous avons présenté ensuite des techniques de ré-estimation et de lissage appliquées aux cooccurrences des mots dans le but de rendre plus fiables ces valeurs de cooccurrence. Nous avons constaté que les techniques de lissage et plus particulièrement l'interpolation linéaire de Jelinek-Mercer donnaient les meilleurs résultats sur deux des trois configurations retenues (IM-COS et DOR-COS). Concernant la configuration TV-JAC, seule la technique de Laplace (Add-One) a montré une amélioration des résultats. Si l'utilisation des techniques de lissage a un intérêt non négligeable, il reste néanmoins une question non résolue et qui nécessite encore des efforts, à savoir l'échec de la plupart des techniques de ré-estimation quand elles sont associées à la mesure du taux de vraisemblance.

Dans la deuxième partie des contributions, nous avons commencé par présenter une méthode inspirée de la méthode par similarité inter-langue où nous avons défini une autre manière d'exploiter les k plus proches voisins d'un mot à traduire. Si la méthode Metarecherche a montré que l'on pouvait améliorer les résultats en combinant différemment les k plus proches voisins, celle-ci reste intrinsèquement liée à l'approche par similarité interlangue. Ensuite, une méthode nommée *Q-Align* directement inspirée des systèmes de question/réponse a été présentée. Si *Q-Align* constitue une autre manière d'appréhender l'alignement bilingue, les différents paramètres qui lui incombent nécessitent une attention particulière dans la mesure où *Q-Align* est très sensible au choix des contextes locaux (requêtes). Enfin, nous avons abordé un autre point essentiel qui est la représentation des mots du corpus dans différents

espaces vectoriels. Après l'introduction des principes de base, nous avons présenté d'autres transformations mathématiques (LSA, PCA et ICA) et pu constater leurs performances ainsi que l'intérêt de les combiner notamment avec l'approche directe. L'un des problèmes non résolu reste le choix du nombre de dimensions des transformations mathématiques qui est fait empiriquement jusqu'à présent.

Ces travaux nous ont amené à définir le premier prototype d'un système multi-sources. Si là encore, nous avons pu relever une amélioration des performances en choisissant de combiner les approches les plus efficaces, un travail de fond reste à faire quant à la manière d'optimiser la phase de combinaison. Mais avant cela une étape préalable est nécessaire. Rappelons que notre principal but était d'améliorer la caractérisation du contexte. Nous avons observé que la combinaison *a priori* avait un grand intérêt. Une première perspective est donc d'intégrer cette représentation dans l'approche par similarité inter-langue, l'approche Metarecherche ainsi que les multiples représentations mathématiques. Concernant *Q-Align*, intégrer la représentation par relations de dépendance syntaxique devrait être plus en adéquation avec l'idée d'une comparaison plus fine des contextes locaux. Pour toutes les approches, les techniques de ré-estimation sont envisageables, sauf peut-être les techniques de lissage pour les transformations mathématiques plus évoluées (LSA, PCA, ICA) qui incluent déjà une étape de lissage dans leur processus de construction d'espace vectoriel. Pour conclure, l'amélioration de la qualité de l'extraction terminologique bilingue à partir de corpus comparables doit d'abord passer par une description contextuelle discriminante, ainsi que par un corpus comparable de qualité et par un dictionnaire bilingue adapté.

V

Annexe

A

Espace vectoriel

A.1 Transformations mathématiques

Tout en gardant le principe de la représentation vectorielle des mots, les chercheurs se sont tournés vers des transformations qui offraient d'intéressantes propriétés mathématiques telles que l'analyse en composantes principales (PCA), l'analyse en composantes indépendantes (ICA), etc. Dans ce qui suit, nous introduirons les principales méthodes utilisées pour le changement d'espace vectoriel. Nous commencerons par la LSA qui reste un classique dans le domaine de la recherche d'information, pour ensuite introduire la PCA qui est une technique incontournable en ce qui concerne la réduction de dimensionnalité tout en maximisant la variance des données. L'analyse en composantes canoniques (CCA) a aussi montré des résultats prometteurs notamment grâce aux travaux de [Haghighi *et al.*, 2008], c'est pourquoi nous nous intéresserons aussi à cette technique. La PCA choisit des traits caractéristiques qui représentent au mieux les données mais pas nécessairement les plus discriminants. L'analyse linéaire discriminante (LDA) tente de résoudre ce problème en utilisant des critères supervisés pour choisir un ensemble de traits caractéristiques à partir des données initiales. Étant intéressé par des méthodes non supervisées, nous nous tournons vers une autre technique connue pour sa capacité à discriminer les données notamment pour la séparation de sources qui est l'ICA.

A.1.1 Analyse sémantique latente (LSA)

Introduite par [Deerwester *et al.*, 1990], la LSA a suscité un grand engouement parmi les chercheurs depuis le début des années 1990. Principalement dédiée à la résolution des problèmes de polysémie et de synonymie dans le domaine de la recherche d'information, elle est aussi considérée comme l'une des techniques classiques de la réduction de dimensions en traitement de textes. L'idée principale de la LSA est de voir la production d'un texte comme un processus de génération de fréquences de mots qui peuvent être caractérisées par un plus petit nombre de facteurs sous-jacents. La LSA exploite la structure sémantique implicite des associations entre

les mots et les documents dans le but d'améliorer l'extraction de documents pertinents en fonction des termes d'une requête donnée. Elle suppose que les termes des documents n'apparaissent pas indépendamment les uns des autres mais plutôt selon certaines spécificités liées aux domaines, aux styles et aux sujets traités dans un texte donné. Une grande collection de documents peut être utilisée pour estimer les dépendances des mots observés et extraire leurs facteurs sous-jacents, aussi appelés variables cachées ou variables latentes.

Partant d'une matrice X représentant les relations entre les termes et les documents, la LSA utilise la décomposition en valeurs singulières (SVD) pour dériver le modèle de la structure sémantique latente à partir de la matrice X. La SVD est étroitement liée à un bon nombre de techniques mathématiques et statistiques dans une large variété de domaines incluant la décomposition en vecteurs propres, l'analyse spectrale et factorielle. Le principe de la SVD est que chaque matrice rectangulaire (X par exemple) peut être décomposée en un produit de trois autres matrices :

$$X = T_0 S_0 D_0^t \tag{A.1}$$

tel que T_0 et D_0 ont des colonnes orthonormales et S_0 est diagonale. Cette représentation est appelée la décomposition en valeurs singulières de X où T_0 et D_0 représentent les matrices de vecteurs singuliers gauche et droit, et S_0 la matrice des valeurs singulières. La SVD est unique jusqu'à un certain nombre de permutations des lignes, colonnes et signes des matrices. Par convention, les valeurs de la diagonale de S_0 sont positives et ordonnées de manière décroissante. Les k premières plus grandes valeurs de la diagonale peuvent être retenues et le reste des valeurs mis à zéro. Le produit des matrices résultantes est alors la matrice $\widehat{X}$ de rang k et qui est approximativement égale à X. En remplaçant les valeurs diagonales non retenues dans S_0 par des zéros, la représentation peut être simplifiée en supprimant les lignes et colonnes de S_0 contenant donc les zéros pour obtenir une nouvelle matrice diagonale S, puis en supprimant les colonnes correspondantes des matrices T_0 et D_0 pour obtenir les matrices T et D correspondantes. Le résultat est le modèle réduit suivant :

$$\widehat{X} = TSD^t \tag{A.2}$$

La valeur k correspondant au nombre de dimensions retenues est critique, idéalement k doit couvrir la structure réelle des données mais aussi être assez petite pour ne pas couvrir les erreurs et informations inutiles.

La décomposition SVD permet trois types de comparaisons :

- Une comparaison de termes (quel est le degré de similarité des deux termes i et j par exemple ?)
- Une comparaison de documents (quel est le degré de similarité des deux documents i et j par exemple ?)

– Une comparaison de termes et de documents (quel est le degré d'association du terme i au document j par exemple ?)

Nous utiliserons donc pour les trois types de comparaisons la matrice $\widehat{X}$ qui est censée représenter un modèle fiable des données de la matrice initiale X. Puisque $\widehat{X} = TSD^t$, le calcul des valeurs pertinentes peut être effectué juste en utilisant les matrices T, D et S.

A.1.2 Analyse en composantes principales (PCA)

L'idée centrale de l'analyse en composantes principales (PCA) est de réduire le nombre de dimensions d'un large ensemble de données composées de variables reliées entre elles, en gardant un maximum de variations présentes dans l'ensemble de départ. Ceci est fait en appliquant une transformation mathématique qui permet de construire un nouvel espace vectoriel composé des n premiers vecteurs propres, en déterminant des valeurs propres décorrélées et ordonnées de façon à ce que les premières valeurs propres retiennent le maximum de variance contenue dans l'ensemble des variables originales. La base formée par ces vecteurs génère alors un espace utilisé pour représenter les données. Dans notre cas, les données correspondent au vocabulaire d'un corpus. Après projection dans ce nouvel espace, chaque mot se voit attribuer un vecteur composé de ses coefficients de projection.

Pour construire le nouvel espace vectoriel, que l'on peut aussi appeler "espace propre", la première étape consiste à construire une matrice qui représente les données de départ. Soit donc une matrice X de type mots-contextes de dimension $n \times m$, où les lignes correspondent aux n mots du corpus appartenant au dictionnaire bilingue $(s_1, s_2..., s_n)$, et les colonnes aux m mots du corpus $(e_1, e_2, ...e_m)$. Le fait de se limiter dans notre représentation aux n mots appartenant au dictionnaire s'explique par le fait que l'on se place dans le cadre bilingue. Partant d'un corpus comparable bilingue composé de textes source et cible, si nous voulons trouver la traduction d'un mot source en langue cible, nous passerons par les mêmes étapes que l'*approche directe* sauf que l'on rajoute une étape qui consiste à réduire le nombre de dimensions de notre espace, soit les étapes suivantes :

1. Construire le vecteur de contexte du mot à traduire w ;
2. Construire la matrice X en langue cible ;
3. Appliquer la PCA sur la matrice X et obtenir le nouvel espace Ξ ;
4. Traduire les entrées du vecteur de contexte du mot w ;
5. Projeter le vecteur traduit dans l'espace propre Ξ ainsi que tous les mots de la langue cible ;
6. Déterminer la bonne traduction en comparant tous les mots cibles projetés avec la projection du mot w et ordonner les candidats selon une mesure de distance.

La PCA utilise la transformation de Karhunen Loeve (KLT) qui est une transformation linéaire qui diagonalise la matrice de covariance ou de corrélation d'une

séquence de variables aléatoires discrètes. Elle est considérée comme la transformation la plus optimale dans le sens de compactage d'énergie, c'est-à-dire qu'elle concentre le maximum d'information possible dans un minimum de coefficients. Les vecteurs bases de la KLT dépendent des statistiques des données d'entrées, ce qui implique une dépendance statistique, qui veut dire que si les statistiques changent la KLT change aussi.

Bien qu'elle n'ignore pas la covariance et la corrélation, la PCA se concentre sur la variance. La première étape est de chercher une fonction linéaire $\alpha_1^t x$ des variables m du vecteur x qui ont un maximum de variance, où α_1 est un vecteur de m constantes $\alpha_{11}, \alpha_{12}, ..., \alpha_{1m}$ représenté comme suit :

$$\alpha_1^t x = \alpha_{11}x_1 + \alpha_{12}x_2 + ... + \alpha_{1n}x_m = \sum_{j=1}^{m} \alpha_{1j}x_j \tag{A.3}$$

De la même manière on continue la recherche d'une deuxième fonction linéaire $\alpha_2^t x$ non corrélée avec $\alpha_1^t x$ et qui contient un maximum de variance et ainsi de suite jusqu'à la k ième étape où l'on retrouve une fonction linéaire $\alpha_k^t x$ non corrélée avec les $\alpha_1^t x$,$\alpha_2^t x$,...,$\alpha_{k-1}^t x$ fonctions linéaires précédentes. $\alpha_k^t x$ est la k ième composante principale parmi les m possibles. En général, le maximum de variance est contenu dans les k premières composantes principales avec $k \ll m$. Ayant défini les composantes principales, il reste maintenant à savoir comment les trouver.

Soit X notre matrice de données initiales et Σ sa matrice de covariance. Chaque élément de la matrice Σ représenté par le couple (i, j) correspond à la covariance entre i et j si $i \neq j$, et à la variance si $i = j$. Pour dériver les composantes principales, il faut premièrement considérer $\alpha_1^t x$; le vecteur α_1 maximise $var[\alpha_1^t x] = \alpha_1^t \Sigma \alpha_1$. Il est clair que le maximum ne sera pas atteint pour des valeurs finies de α_1, c'est pourquoi une contrainte de normalisation est fixée. La contrainte utilisée dans la dérivation est $\alpha_1^t \alpha_1 = 1$. D'autres types de contraintes de normalisation peuvent être utilisés comme par exemple : $\max_j |\alpha_{1j}| = 1$ mais ceux-ci pourraient poser plus de difficultés d'optimisation. Pour maximiser $\alpha_1^t \Sigma \alpha_1$ sous la contrainte $\alpha_1^t \alpha_1 = 1$, l'approche standard est d'utiliser la technique du multiplicateur de Lagrange. Maximiser donc :

$$\alpha_1^t \Sigma \alpha_1 - \lambda(\alpha_1^t \alpha_1 - 1) \tag{A.4}$$

où λ est le multiplicateur de Lagrange. La différentiation en fonction de α_1 donne :

$$\Sigma \alpha_1 - \lambda \alpha_1 = 0 \tag{A.5}$$

ou alors :

$$(\Sigma - \lambda I_m)\alpha_1 = 0 \tag{A.6}$$

avec I_m la matrice d'identité de dimensions $m \times m$. Ainsi, λ est une valeur propre de Σ et α_1 le vecteur propre correspondant. Pour décider lesquels des m vecteurs propres donnent $\alpha_1^t x$ avec un maximum de variance, il faut maximiser la quantité suivante :

$$\alpha_1^t \Sigma \alpha_1 = \alpha_1^t \lambda \alpha_1 = \lambda \alpha_1^t \alpha_1 = \lambda \tag{A.7}$$

La valeur de λ doit être aussi grande que possible, Ainsi α_1 est le vecteur propre correspondant à la plus grande valeur propre de Σ et $var[\alpha_1^t x] = \alpha_1^t \Sigma \alpha_1 = \lambda_1$ la plus grande valeur propre. De façon générale, la k ième composante principale de X est $\alpha_1^t x$, et $var[\alpha_k^t x] = \lambda_k$ où λ_k est la k ième plus grande valeur propre de Σ, avec α_k le vecteur propre correspondant.

Pour finir, et afin d'utiliser la PCA d'une manière optimale, nous rappelons les étapes suivantes :

1. Construire la matrice de données initiales X ;
2. Soustraire la moyenne à toutes les variables de la matrice X ;
3. Calculer la matrice de covariance de X, à savoir Σ ;
4. Calculer les valeurs et vecteurs propres de la matrice Σ.

La soustraction de la moyenne est une étape préalable importante, car elle permet de retirer tout ce qui est commun aux échantillons.

A.1.3 Analyse en composantes canoniques (CCA)

Introduite par [Hotelling, 1936], la CCA est une technique d'extraction de relations de corrélation entre deux variables multi-dimensionnelles. Elle peut être utilisée pour identifier des directions sources et cibles qui sont corrélées de façon maximale. Elle est connue pour pouvoir traiter des ensembles de variables dépendantes et indépendantes. Alors que la régression linéaire multiple permet de prédire une seule variable dépendante à partir d'un ensemble de variables indépendantes, la CCA permet de prédire simultanément plusieurs variables dépendantes à partir d'un ensemble de variables dépendantes et indépendantes. Les données d'un corpus bilingue peuvent être naturellement représentées par une matrice ζ de taille $n \times (m + r)$, où les lignes correspondent aux couples de traductions du dictionnaire bilingue $((s_1, t_1), ..., (s_n, t_n))$, et les colonnes aux mots source $(e_1, e_2, ...e_m)$ et cible $(f_1, f_2, ...f_r)$ du vocabulaire. La matrice ζ montre que chaque couple de traduction soutient deux vues grâce aux vecteurs de contextes des langues source et cible. Chaque vue est connectée à l'autre grâce aux couples de traductions. La CCA peut donc être utilisée pour identifier les directions dans la vue source (les m premières colonnes de la matrice ζ), la vue cible (les r premières colonnes de la matrice ζ) qui sont corrélées de façon maximale. Ainsi, la CCA recherche les directions dans l'espace vectoriel source et cible défini par les vecteurs bases $(e_1, e_2, ...e_m)$ et $(f_1, f_2, ...f_r)$ de telle sorte que les projections des couples de traduction dans ce nouvel espace soient corrélés de façon maximale. Intuitivement, ces directions définissent des axes

sémantiques latents qui capturent les relations entre les couples de traductions. Soit ξ_s et ξ_t les directions dans les corpus source et cible. Ceci peut être représenté comme suit :

$$\rho = \max_{\xi_s \xi_t} \frac{\sum_i \langle \xi_s, \overrightarrow{s_i} \rangle \langle \xi_t, \overrightarrow{t_i} \rangle}{\sqrt{\sum_i \langle \xi_s, \overrightarrow{s_i} \rangle \sum_j \langle \xi_t, \overrightarrow{t_j} \rangle}} \tag{A.8}$$

Comme pour l'analyse en composantes principales, une fois les deux premières directions identifiées (ξ_s^1, ξ_t^1), le processus peut être répété dans le sous-espace orthogonal à celui formé par les directions déjà identifiées. Une solution plus générale basée sur les valeurs propres peut être proposée. En suivant l'exemple de (Back et Jordan 2001), le problème ci-dessus peut être reformulé comme suit :

$$\mathbf{B}\xi = \rho \mathbf{D}\xi \tag{A.9}$$

où R_s et R_t sont respectivement les m premières et r dernières colonnes de la matrice ζ, avec :

$$\mathbf{B} = \begin{pmatrix} 0 & R_t R_t^T R_s R_s^T \\ R_s R_s^T R_t R_t^T & 0 \end{pmatrix}, \mathbf{D} = \begin{pmatrix} (R_s R_s^T)^2 & 0 \\ 0 & (R_t R_t^T)^2 \end{pmatrix}, \xi = \begin{pmatrix} \xi_s \\ \xi_t \end{pmatrix}$$

L'une des manières de résoudre l'équation A.9 est d'utiliser la décomposition incomplète de Cholesky de la matrice D (Bach et Jordan 2001). Ainsi, nous obtenons un nouveau sous-espace vectoriel défini par les couples sources et cibles $(\xi_s^1, \xi_t^1), ..., (\xi_s^l, \xi_t^l)$. Une fois projetés, les mots des corpus source et cible pourront être comparés grâce à un produit scalaire ou en utilisant la mesure du cosinus. Soit le nouvel espace $\Xi_s = [\xi_s^1, ...\xi_s^l]^T$, et $\Xi_t = [\xi_t^1, ...\xi_t^l]^T$, alors la similarité sera représentée comme suit :

$$S(v, w) = \langle \Xi_s \overrightarrow{v}, \Xi_t \overrightarrow{w} \rangle = \overrightarrow{v}^T \Xi_s^T, \Xi_t \overrightarrow{w} \tag{A.10}$$

Les dimensions du nouvel espace construit grâce à la CCA seront définies par les l vecteurs retenus dans chaque langue.

Il est à noter que la CCA a donné de mauvais résultats dans nos expériences.

A.2 Approche GLICA

Nous avons développé une approche nommée GLICA [Hazem et Morin, 2012], qui consiste en une combinaison de deux représentations de l'ICA, une représentation globale (GICA) et une représentation locale (LICA) décrites ci dessous.

A.2.1 Rappel : représentation des données

Les données observées x sont une matrice de mots $N \times N$ où les colonnes représentent les contextes et les lignes représentent les mots. Les N mots de la langue cible qui appartenant au dictionnaire bilingue sont sélectionnés pour construire la matrice X. Chaque colonne de la matrice X représente un vecteur de contexte d'un mot i avec $i \in N$. Étant donné un élément X_{cr} de la matrice X, X_{cr} représente la mesure d'association entre le mot de la r :ième ligne et le contexte de la c :ième colonne. Les mesures d'association utilisées peuvent être l'information mutuelle, le taux de vraisemblance, etc.

Soient X_g et X_l deux représentations initiales des représentations globale et locale. En utilisant l'ICA sous ses deux forment (globale et locale), les matrices initiales X_g et X_l seront transformées en deux nouveaux espaces de composantes indépendantes $s^g = (s_1^g, ..., s_k^g, ..., x_n^g)^T$ et $s^l = (s_1^l, ..., s_k^l, ..., x_n^l)^T$.

A.2.2 Représentation globale : GICA

La représentation globale des données appelée *GICA* a pour but de construire un espace global s^g de données où tous les mots du corpus présents dans le dictionnaire sont utilisés. Ainsi chaque composante indépendante s_k^g va contenir ou encoder une certaine quantité d'information extraite des N mots de la langue cible.

A.2.3 Représentation locale : LICA

La représentation locale des données appelée *LICA* a pour but de construire un espace de représentation partiel qui n'est construit qu'à partir des mots faisant partie du contexte proche du mot à traduire.

Chaque composante indépendante s_k^l va encoder des informations apprises à partir des M mots faisant partie du contexte du mot à traduire. Le but de cette représentation est de capturer une information spécifique au mot à traduire sans prendre en compte l'ensemble des mots du corpus.

A.2.4 Modèle final : GLICA

Notons par $d_{GL}(i,j)$, ($d_G(i,j)$ et $d_L(i,j)$), les distances de GLICA, GICA et LICA. Le modèle GLICA est une somme pondérée ou une interpolation linéaire des modèles GICA et LICA comme suit :

$$d_{GL}(i,j) = \lambda \times d_G(i,j) + (1-\lambda) \times d_L(i,j) \tag{A.11}$$

B

Listes d'évaluation

Cancer du sein

321 couples En/Fr

Termes anglais	Termes français	# occ_{en}	# occ_{fr}
birth	naissance	74	18
relapse	rechute	67	526
carcinogenesis	carcinogenèse	17	7
pregnancy	grossesse	95	181
cyclin	cycline	23	8
diet	alimentation	35	5
survival	survie	442	677
life	vie	214	171
history	antécédent	162	208
light	lumière	40	37
abnormality	anomalie	21	114
technology	technologie	34	29
medicine	médecine	25	15
estradiol	estradiol	21	45
lymphoma	lymphome	10	13
time	temps	577	343
trust	confiance	5	20
node	ganglion	582	401
adjustment	ajustement	59	26
docetaxel	docétaxel	34	36
adjuvant	adjuvant	340	398
oncogene	oncogène	19	26
anatomy	anatomie	11	10
electronic	électronique	17	15
interview	entretien	84	28
ligand	ligand	26	17

skin	peau	184	121
concentration	concentration	310	153
health	santé	229	55
resection	résection	83	66
lymph	lymphe	309	6
metastasis	métastase	246	495
necrosis	nécrose	20	119
resistance	résistance	33	33
causality	causalité	7	10
human	humain	407	62
recovery	guérison	8	14
ph	ph	61	6
review	revue	243	50
incidence	incidence	183	252
breast	sein	3477	3007
future	futur	53	13
injection	injection	59	167
policy	principe	17	49
carcinoma	carcinome	201	384
sample	échantillon	352	124
drug	médicament	161	38
internet	internet	14	5
iridium	iridium	11	7
macrophage	macrophage	18	6
histology	histologie	56	79
osteoclast	ostéoclaste	26	12
infection	infection	36	35
publication	publication	30	75
genomic	génomique	50	13
artery	artère	47	14
risk	risque	1080	1621
women	femme	46	1223
depression	dépression	26	14
surgery	chirurgie	371	499
care	soin	131	30
menopause	ménopause	12	127
fibrosis	fibrose	13	112
randomization	randomisation	17	31
cdna	adnc	43	9
death	décès	121	71
chromatography	chromatographie	22	7
heart	coeur	215	29
back	dos	6	7
consensus	consensus	32	34
confusion	confusion	6	22
nature	nature	37	39
heat	chaleur	15	55
software	logiciel	66	48

recommendation	recommandation	48	99
acetate	acétate	6	13
vincristine	vincristine	7	6
lumpectomy	zonectomie	42	7
lymphocyte	lymphocyte	33	43
ultrasound	échographie	32	124
employment	emploi	8	24
collaboration	collaboration	9	12
chest	poitrine	121	7
proliferation	prolifération	196	97
examination	examen	86	404
mortality	mortalité	123	68
igg	igg	16	16
radiologist	radiologue	41	39
adult	adulte	15	11
arm	bras	103	223
pathology	pathologie	31	127
affect	affect	153	2
lesion	lésion	151	350
disease	maladie	653	377
chemotherapy	chimiothérapie	596	680
recurrence	récidive	464	736
receptor	récepteur	402	437
sensitivity	sensibilité	108	137
probe	sonde	44	53
family	famille	211	61
cisplatin	cisplatine	7	8
production	production	63	53
hypertension	hypertension	9	10
oestrogen	oestrogène	39	54
separation	séparation	8	8
intention	intention	7	21
prolactin	prolactine	6	9
education	éducation	60	6
inhibition	inhibition	136	31
work	travail	98	178
observation	observation	80	56
population	population	243	306
melanoma	mélanome	5	11
segmentectomy	quadrantectomie	15	12
projection	projection	16	26
evaluation	évaluation	74	260
algorithm	algorithme	112	6
osteoblast	ostéoblaste	98	21
futility	limitation	6	10
character	caractère	6	70
femur	fémur	6	5
fluorescence	fluorescence	51	15

absorption	absorption	8	13
prevention	prévention	23	105
bias	biais	80	92
hormone	hormone	163	82
radiology	radiologie	6	6
calcium	calcium	23	14
satisfaction	satisfaction	30	46
aromatase	aromatase	51	48
tobacco	tabac	8	10
price	prix	6	16
thorax	thorax	14	15
dna	adn	241	138
excision	exérèse	68	266
uncertainty	incertitude	15	46
administration	administration	43	104
oxygen	oxygène	21	6
safety	sécurité	24	28
scar	cicatrice	52	132
attitude	attitude	47	52
behavior	comportement	24	7
blood	sang	125	19
environment	environnement	21	16
comment	commentaire	16	6
parity	parité	25	11
hand	main	56	11
phenotype	phénotype	57	23
animal	animal	75	55
communication	communication	24	18
exemestane	exemestane	17	18
dissection	curage	57	295
literature	littérature	69	197
pressure	pression	17	7
cold	froid	10	10
apoptosis	apoptose	129	53
oncology	cancérologie	62	14
brachytherapy	curiethérapie	40	124
vein	veine	14	6
selenium	sélénium	6	11
liver	foie	87	12
lymphedema	lymphoedème	6	44
canada	canada	20	10
temperature	température	48	6
epithelium	épithélium	5	20
vegf	vegf	125	7
edema	oedème	6	37
taxane	taxane	46	33
cancer	cancer	3599	3308
consultation	consultation	29	91

mri	irm	72	195
membrane	membrane	118	28
role	rôle	249	142
homeostasis	homéostasie	7	5
therapeutic	thérapeutique	69	418
therapy	thérapie	378	31
bone	os	249	23
involvement	envahissement	80	259
specificity	spécificité	45	60
reconstruction	reconstruction	93	543
trastuzumab	trastuzumab	44	210
cyclophosphamide	cyclophosphamide	41	48
color	couleur	8	10
anthracycline	anthracycline	45	65
estrone	estrone	5	7
tyrosine	tyrosine	38	15
association	association	309	319
attention	attention	20	25
knowledge	connaissance	133	49
sensation	sensation	6	12
charge	frais	8	10
reading	lecture	25	19
cytology	cytologie	7	19
tamoxifen	tamoxifène	187	414
hip	hanche	17	7
oncologist	oncologue	44	25
boost	surdosage	167	19
scintigraphy	scintigraphie	7	11
morbidity	morbidité	26	77
hyperplasia	hyperplasie	23	35
lymphoscintigraphy	lymphoscintigraphie	20	23
mouse	souris	351	18
head	tête	20	18
micrometastasis	micrométastase	6	49
antigen	antigène	32	24
identification	identification	58	59
immunohistochemistry	immunohistochimie	21	60
glucose	glucose	21	6
ice	glace	10	6
anxiety	anxiété	38	24
abdomen	abdomen	15	10
fear	peur	22	16
ovary	ovaire	6	119
marrow	moelle	99	13
vitamin	vitamine	76	32
alteration	altération	42	68
science	science	32	5
radiotherapy	radiothérapie	163	790

growth	croissance	277	99
inhibitor	inhibiteur	187	71
allele	allèle	64	13
estrogen	estrogène	195	283
angiogenesis	angiogenèse	56	9
probability	probabilité	79	91
diagnosis	diagnostic	298	395
complication	complication	35	289
sex	sexe	7	15
staining	coloration	89	7
biopsy	biopsie	185	102
eosin	éosine	17	9
epidemiology	épidémiologie	8	6
irradiation	irradiation	223	626
insulin	insuline	31	7
prevalence	prévalence	17	31
syndrome	syndrome	8	65
tumor	tumeur	1456	1386
gene	gène	486	325
milk	lait	24	18
electrophoresis	électrophorèse	21	16
perception	perception	20	7
brain	cerveau	103	9
progesterone	progestérone	83	98
fibroblast	fibroblaste	13	6
screening	dépistage	246	225
schedule	planning	22	4
adenocarcinoma	adénocarcinome	10	40
colon	côlon	11	17
muscle	muscle	35	104
hematoxylin	hématoxyline	17	5
wine	vin	56	9
prognosis	pronostic	172	218
efficiency	rendement	19	5
hybridization	hybridation	15	16
detection	détection	113	230
paclitaxel	paclitaxel	48	32
drainage	drainage	23	97
distress	douleur	23	110
genome	génome	6	23
morphology	morphologie	21	22
face	face	25	76
nausea	nausée	8	8
doxorubicin	doxorubicine	31	33
inflammation	inflammation	31	11
transfer	transfert	42	58
motion	déplacement	11	13
adolescent	adolescent	32	5

movement	mouvement	10	44
axilla	aisselle	24	18
serine	sérine	5	5
sarcoma	sarcome	6	21
toxicity	toxicité	88	163
fibroadenoma	fibroadénome	9	9
clinician	clinicien	22	19
glass	verre	10	6
spleen	rate	7	12
pain	douleur	32	110
mammography	mammographie	90	234
smoking	tabagisme	57	15
neck	cou	7	15
community	communauté	11	5
mutation	mutation	110	188
erythema	érythème	10	6
consumption	consommation	154	21
exon	exon	9	8
radiation	radiation	224	31
rna	arn	48	29
epirubicin	épirubicine	48	12
permit	permis	14	48
hospitalization	hospitalisation	6	31
insurance	assurance	21	11
transplantation	greffe	31	44
mastectomy	mastectomie	178	406
centrifugation	centrifugation	19	16
treatment	traitement	1237	1898
pcr	pcr	88	28
diffusion	diffusion	5	22
keratin	kératine	7	5
paraffin	paraffine	16	18
classification	classification	105	99
mitoxantrone	mitoxantrone	32	17
heterogeneity	hétérogénéité	23	31
antibody	anticorps	202	157
anesthesia	anesthésie	58	39
kinase	kinase	102	22
orientation	orientation	8	18
research	recherche	228	179
hospital	hôpital	126	23
marker	marqueur	163	212
chromosome	chromosome	37	23
spine	rachis	14	6
ability	capacité	106	47
nipple	mamelon	56	44
prostate	prostate	41	41
play	jeu	89	18

lung	poumon	208	57
symptom	symptôme	35	48
protein	protéine	384	226
metabolism	métabolisme	16	20
serum	sérum	179	82
anastrozole	anastrozole	32	16
margin	berge	144	83

Énergies renouvelables

150 couples En/Fr

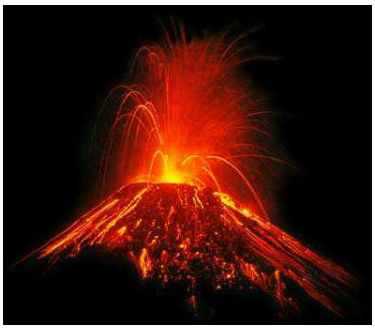
Termes anglais	Termes français	# occ_{en}	# occ_{fr}
tower	tour	503	77
steel	acier	58	12
two-bladed	bipale	12	18
generator	générateur	558	621
viscous	visqueux	12	16
inverter	onduleur	50	184
betz	betz	13	6
stall	décrochage	159	14
power	puissance	1483	2422
load	charge	182	412
horizontal	horizontal	78	100
sound	sonore	673	20
rotation	rotation	67	383
watt	watt	9	10
cable	câble	115	22
obstacle	obstacle	18	9
measurement	dimensionnement	99	72
distribution	distribution	165	119
energy	énergie	1429	1687
variable	variable	162	322
gas	gaz	126	215
cylinder	cylindre	28	17
battery	batterie	72	383
direction	direction	95	60
wind	vent	4714	1272
speed	vitesse	706	1653
aerodynamic	aérodynamique	175	125
quality	qualité	67	62
yaw	orientation	79	74
shaft	arbre	94	91
cyclic	cyclique	9	62
kilowatt	kilowatt	21	16
height	hauteur	132	71
hybrid	hybride	21	90
coefficient	coefficient	206	242
average	moyen	178	443
mwh	mwh	42	27
synchronous	synchrone	64	129

brake	frein	79	13
profile	profil	72	76
law	loi	65	107
cooling	refroidissement	5	19
economics	économie	10	44
trail	fuite	23	22
mast	mât	7	39
visual	visuel	57	14
curve	courbe	100	173
diagram	diagramme	25	24
turbine	turbine	2808	595
fix	fixe	64	121
viscosity	viscosité	8	13
biological	biologique	19	8
total	total	150	161
corrosion	corrosion	5	5
coupling	couplage	14	16
gust	rafale	14	16
vertical	vertical	85	86
blade	pale	1425	477
impact	impact	593	393
support	support	180	12
mechanism	dispositif	48	136
connection	connexion	125	84
maximum	maximal	147	249
project	projet	1233	351
rectifier	redresseur	14	103
factor	facteur	292	134
induction	induction	81	26
hinge	articulation	8	13
offshore	offshore	395	41
global	global	58	81
thyristor	thyristor	14	10
stator	stator	47	106
domestic	domestique	37	19
anemometer	anémomètre	15	80
indirect	indirect	15	41
annoyance	nuisance	84	25
slipstream	sillage	27	62
kwh	kwh	25	85
potential	potentiel	338	73
sensor	capteur	13	115
shear	cisaillement	14	6
bird	oiseau	281	13
park	parc	65	363
tip	bout	108	16
map	carte	92	91
lattice	treillis	8	6

loss	perte	150	564
drag	traînée	84	28
benefit	bénéfice	83	13
base	base	288	95
hub	moyeu	114	23
kw	kw	85	265
mechanical	mécanique	122	224
farm	ferme	585	76
foundation	fondation	101	15
density	densité	79	50
voltage	voltage	214	34
lift	portance	130	37
level	niveau	529	318
transformer	transformateur	45	34
conversion	conversion	69	319
source	source	285	368
efficiency	rendement	107	231
local	local	199	101
velocity	vitesse	183	1653
mw	mw	189	241
darrieus	darrieus	28	24
sustainable	durable	41	15
transmission	transmission	85	64
angle	angle	385	253
nacelle	nacelle	69	157
renewable	renouvelable	200	286
resource	ressource	172	57
site	site	619	331
wake	sillage	50	62
manufacturer	fabricant	34	26
storage	stockage	94	787
converter	convertisseur	102	524
page	page	262	30
slip	glissement	11	32
axis	axe	158	214
regulation	régulation	117	144
pitch	calage	283	144
pump	pompe	79	31
shadow	ombre	46	5
location	implantation	247	84
gearbox	multiplicateur	78	54
rotor	rotor	848	467
parameter	paramètre	79	250
scale	échelle	96	86
vawt	vawt	61	22
direct	direct	76	97
disk	disque	60	45
solar	solaire	60	80

safety	sécurité	154	55
availability	disponibilité	51	25
savonius	savonius	13	25
onshore	terrestre	60	14
current	courant	196	563
edge	bord	50	13
asynchronous	asynchrone	6	215
emission	émission	286	94
three-bladed	tripale	29	36
hawt	hawt	109	32
structure	structure	186	206
breeze	brise	12	9
vortex	tourbillon	132	6
turbulence	turbulence	49	47
air	air	196	146
noise	bruit	484	45

Vulcanologie



158 couples En/Fr

Termes anglais	Termes français	# occ_{en}	# occ_{fr}
pressure	pression	238	222
tower	tour	19	78
flank	flanc	279	188
storm	tempête	40	17
sulfur	soufre	108	180
catastrophe	catastrophe	39	144
thermal	thermique	36	42
lapilli	lapilli	40	29
horizontal	horizontal	34	17
central	central	125	122
temperature	température	221	307
chamber	chambre	93	73
basin	bassin	39	31
hydrovolcanic	hydrovolcanique	18	8
rigid	rigide	20	18
ignimbrite	ignimbrite	10	16
seismograph	sismographe	33	26
stratosphere	stratosphère	54	56
delta	delta	14	8
mantle	manteau	190	236
tephra	tephra	101	15
tube	tunnel	147	38
magnitude	magnitude	89	16
froth	mousse	9	24
gas	gaz	689	599
volcanism	volcanisme	123	268
time	temps	690	301
geologist	géologue	216	104
effusive	effusif	16	42
flood	inondation	143	28
speed	vitesse	89	129
erosion	érosion	64	61
caldera	caldera	367	47
obsidian	obsidienne	14	19
iceberg	iceberg	7	11
geophysicist	géophysicien	37	23
peak	piton	171	161

geothermal	géothermique	25	40
lahar	lahar	136	48
fragmentation	fragmentation	19	13
continental	continental	74	94
crisis	crise	41	63
pumice	ponce	109	70
geology	géologie	59	75
ozone	ozone	79	48
rain	pluie	138	139
sulfuric	sulfurique	22	25
activity	activité	497	503
hydrogen	hydrogène	28	34
red	rouge	113	133
magmatic	magmatique	19	153
active	actif	483	325
lava	lave	2217	1282
vertical	vertical	55	35
vulcanology	vulcanologie	22	5
ice	glace	226	174
maar	maar	15	18
melting	fusion	45	155
grey	gris	28	67
explosive	explosif	281	147
scoria	scorie	40	49
rhyolite	rhyolite	29	17
fire	feu	203	225
extrusion	extrusion	23	13
volatile	volatil	24	7
volcanologist	volcanologue	213	230
global	global	123	34
intrusion	intrusion	60	23
topography	topographie	16	13
oil	huile	24	14
vulcanian	vulcanien	24	38
globe	globe	35	156
block	bloc	147	149
cinder	cendre	126	581
geophysics	géophysique	14	50
chemistry	chimie	30	26
tree	arbre	120	45
gaseous	gazeux	13	45
soil	sol	65	206
geological	géologique	212	157
crust	croûte	261	178
rate	taux	113	13
volcanology	volcanologie	47	127
ball	boule	14	20
fumarole	fumerolle	59	105

vapor	vapeur	32	129
dyke	dyke	28	21
volcanic	volcanique	1413	1153
blast	explosion	224	331
lake	lac	350	562
pyroclastic	pyroclastique	270	77
black	noir	140	181
top	sommet	202	194
basalt	basalte	156	104
plateau	plateau	23	63
plinian	plinienne	46	5
conduit	conduit	83	33
crater	cratère	698	637
zone	zone	356	388
shock	choc	39	26
climate	climat	72	65
rift	rift	189	65
earthquake	séisme	535	126
pool	piscine	31	14
terrace	terrasse	15	16
alert	alerte	32	39
edifice	édifice	24	132
tsunami	tsunami	149	18
vulcano	vulcano	15	41
fountain	fontaine	97	57
volcano	volcan	3714	2602
belt	ceinture	36	38
glass	verre	59	55
surface	surface	628	484
island	ile	1023	9
pipe	pipe	35	18
vulcan	vulcan	17	6
neck	neck	16	15
fissure	fissure	192	147
pahoehoe	pahoehoe	103	8
sill	sill	24	10
mud	boue	95	191
summit	sommet	389	194
earth	terre	672	730
cloud	nuage	309	212
strombolian	strombolien	43	57
mass	masse	99	50
rock	roche	1183	517
column	colonne	101	70
sand	sable	54	54
field	domaine	162	56
tuff	tuf	55	14
glacier	glacier	73	163

eruption	éruption	2849	1561
avalanche	avalanche	153	56
vegetation	végétation	34	70
olivine	olivine	13	35
magma	magma	1064	515
basaltic	basaltique	138	73
fragment	fragment	172	58
shield	bouclier	157	27
bomb	bombe	78	124
acid	acide	65	114
center	centre	117	236
geophysical	géophysique	38	50
vulcania	vulcania	8	105
geyser	geyser	36	70
tremor	tremblement	112	92
collapse	effondrement	215	71
mountain	montagne	565	340
ash	cendre	923	581
sulphur	soufre	48	180
vulcanologist	vulcanologue	82	34
hydrothermal	hydrothermal	27	31
point	point	143	341
volcanological	volcanologique	7	49
air	air	229	149
sleep	sommeil	21	44

Bibliographie

[Abdul-Rauf et Schwenk, 2009] Abdul-Rauf, S. et Schwenk, H. (2009). On the use of comparable corpora to improve smt performance. *In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL'09)*, pages 16–23, Athens, Greece. Cited p. 12.

[Afli *et al.*, 2012] Afli, H., Barrault, L. et Schwenk, H. (2012). Parallel texts extraction from multimodal comparable corpora. *In Advances in Natural Language Processing - 8th International Conference on NLP (JapTAL'12)*, pages 40–51, Kanazawa, Japan. Cited p. 12.

[Aker *et al.*, 2013] Aker, A., Paramita, M. et Gaizauskas, R. (2013). Extracting bilingual terminologies from comparable corpora. *In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13)*, pages 402–411, Sofia, Bulgaria. Cited p. 12.

[Andrade *et al.*, 2011] Andrade, D., Matsuzaki, T. et Tsujii, J. (2011). Effective use of dependency structure for bilingual lexicon creation. *In Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'11)*, pages 80–92, Tokyo, Japan. Cited p. 23.

[Aslam et Montague, 2001] Aslam, J. A. et Montague, M. (2001). Models for Metasearch. *In Proceedings of the 24th Annual SIGIR Conference (SIGIR'01,)*, pages 275–284, New Orleans, Louisiana. Cited pp. 82 et 110.

[Babych *et al.*, 2008] Babych, B., Sharoff, S. et Hartley, A. (2008). Generalising lexical translation strategies for MT using comparable corpora. *In Proceedings of the International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. Cited p. 12.

[Baldwin et Tanaka, 2004] Baldwin, T. et Tanaka, T. (2004). Translation by machine of complex nominals : Getting it right. *In Proceedings of the ACL'04 Workshop on Multiword Expressions : Integrating Processing.* Cited pp. 32 et 33.

[Bartell *et al.*, 1994] Bartell, B. T., Cottrell, G. W. et Belew, R. K. (1994). Automatic combination of multiple ranked retrieval systems. *In Proceedings of the 17th Annual SIGIR Conference (SIGIR'94)*, pages 173–181. Cited p. 111.

[Bouamor *et al.*, 2013] Bouamor, D., Semmar, N. et Zweigenbaum, P. (2013). Utilisation de la similarité sémantique pour l'extraction de lexiques bilingues à partir de corpus comparables. *In Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'13)*, pages 327–338, Les Sables d'Olonne, France. Cited p. 76.

[Bowker et Pearson, 2002] Bowker, L. et Pearson, J. (2002). *Working with Specialized Language : A Practical Guide to Using Corpora*. Routeledge, New York, USA. Cited pp. 9, 11 et 35.

[Brent, 1991] Brent, M. R. (1991). Automatic acquisition of subcategorization frames from untagged text. *In Proceedings of the 29th Meeting of the Association for Computational Linguistics (ACL'91)*, pages 209–214. Cited p. 48.

[Brown *et al.*, 1991] Brown, P. F., Pietra, S. D., Pietra, V. J. D. et Mercer, R. L. (1991). Word sense disambiguation using statistical methods. *In Proceedings of the 29th Meeting of the Association for Computational Linguistics (ACL'91)*, pages 264–270. Cited p. 8.

[Brown *et al.*, 1992] Brown, P. F., Pietra, V. J. D., de Souza, P. V., Lai, J. C. et Mercer, R. L. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479. Cited p. 48.

[Cartoni, 2009] Cartoni, B. (2009). Lexical morphology in machine translation : A feasibility study. *In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL'09)*, pages 130–138, Athens, Greece. Cited p. 33.

[Catizone *et al.*, 1989] Catizone, R., Russell, G. et Warwick-Armstrong, S. (1989). Deriving translation data from bilingual texts. *In* Zernik, éditeur : *Lexical Acquisition Workshop*, Detroit. Cited p. 8.

[Chen et Goodman, 1999] Chen, S. F. et Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–393. Cited pp. 100 et 107.

[Chiao, 2004] Chiao, Y.-C. (2004). Extraction lexicale bilingue à partir de textes médicaux comparables : application à la recherche d'information translangue. Cited pp. 8, 22 et 25.

[Chiao et Zweigenbaum, 2003] Chiao, Y.-C. et Zweigenbaum, P. (2003). The Effect of a General Lexicon in Corpus-Based Identification of French-English Medical Word Translations. *In The New Navigators : from Professionals to Patients, Actes Medical Informatics Europe*, volume 95 de *Studies in Health Technology and Informatics*, pages 397–402, Amsterdam. Cited p. 21.

[Choi *et al.*, 2001] Choi, F. Y. Y., Wiemer-Hastings, P. et Moore, J. (2001). Latent semantic analysis for text segmentation. *In Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP'01)*, pages 109–117, Vancouver, Canada. Cited p. 129.

[Church et Hanks, 1990] Church, K. W. et Hanks, P. W. (1990). Word association norms, mutual information and lexicography. *Computational Linguistics*, 16(1): 22–29. Cited p. 48.

[Comon, 1994] Comon, P. (1994). Independent component analysis—a new concept ? *Signal Processing*, 36:287–314. Cited p. 132.

[Dagan et Church, 1994] Dagan, I. et Church, K. (1994). Termight : Identifying and translating technical terminology. Cited p. 8.

[Dagan *et al.*, 1999] Dagan, I., Lee, L. et Pereira, F. C. N. (1999). Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34(1-3):43–69. Cited p. 24.

[Daille, 1994] Daille, B. (1994). *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques.* Thèse de doctorat, Université Paris 7. Cited p. 54.

[Daille *et al.*, 1994] Daille, B., Gaussier, E. et Langé, J.-M. (1994). Towards automatic extraction of monolingual and bilingual terminology. *In Proceedings of the 15th International Conference on Computational Linguistics (COLING'94).* Cited p. 8.

[Daille et Morin, 2005] Daille, B. et Morin, E. (2005). French-english terminology extraction from comparable corpora. *In Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCLNP'05)*, pages 707–718, Jeju Island, Korea. Cited p. 62.

[de Saussure, 1916] de Saussure, F. (1916). *Cours de linguistique générale.* Paris :Payot. Cited p. 47.

[Deerwester *et al.*, 1990] Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W. et Harshman, R. A. (1990). Indexing by latent semantic analysis. *JASIS*, 41(6):391–407. Cited pp. 48, 49, 129 et 153.

[Déjean *et al.*, 2002] Déjean, H., Gaussier, É. et Sadat, F. (2002). An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. *In Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 1–7, Taipei, Taiwan. Cited pp. 13, 62 et 66.

[Delpech *et al.*, 2012] Delpech, E., Daille, B., Morin, E. et Lemaire, C. (2012). Extraction of domain-specific bilingual lexicon from comparable corpora : Compositional translation and ranking. *In Proceedings of the 24th International Conference on Computational Linguistics (COLING'12)*, pages 745–762, Mumbai, India. Cited pp. 33 et 34.

[Dumais *et al.*, 1988] Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S. et Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 281–285, New York, USA. ACM. Cited p. 129.

[Dunning, 1993] Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74. Cited pp. 20, 48, 55, 62 et 75.

[Déjean et Gaussier, 2002] Déjean, H. et Gaussier, E. (2002). Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica, Alignement Lexical dans les Corpus Multilingues*, pages 1–22. Cited pp. 8, 11, 14, 21, 22, 27, 28, 29, 35, 81, 109 et 112.

[Evert, 2005] Evert, S. (2005). *The statistics of word cooccurrences : word pairs and collocations.* Thèse de doctorat, University of Stuttgart, Holzgartenstr. 16, 70174 Stuttgart. Cited pp. 56, 62 et 76.

[Evert et Baroni, 2007] Evert, S. et Baroni, M. (2007). zipfr : Word frequency modeling in r. *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, Prague, Czech Republic. Cited p. 97.

[Fano, 1961] Fano, R. M. (1961). *Transmission of Information : A Statistical Theory of Communications.* MIT Press, Cambridge, MA, USA. Cited pp. 20, 48, 53 et 62.

[Felber, 1984] Felber, H. (1984). *Basic Principles and Methods for the Preparation of Terminology Standards.* in C.G. Interrante and F.J.Heymann (eds). Cited p. 44.

[Firth, 1957] Firth, J. R. (1957). A synopsis of linguistic theory 1930–1955. *In Studies in Linguistic Analysis (special volume of the Philological Society)*, pages 1–32. Blackwell, Oxford. Cited p. 13.

[Foltz *et al.*, 1998] Foltz, P., Kintsch, W. et Landauer, T. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25(2&3):285–307. Cited p. 130.

[Fung, 1995a] Fung, P. (1995a). Compiling Bilingual Lexicon Entries From a non-Parallel English-Chinese Corpus. *In Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA'95)*, pages 1–16, Langhorne, PA, USA. Cited pp. 8, 11, 13, 16 et 17.

[Fung, 1995b] Fung, P. (1995b). A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. *In Proceedings of the 33rd Meeting of the Association for Computational Linguistics (ACL'95)*, pages 236–243, Boston, MA, États-Unis d'Amérique. Cited p. 11.

[Fung, 1998] Fung, P. (1998). A statistical view on bilingual lexicon extraction : From parallel corpora to non-parallel corpora. *In Proceedings of Machine Translation and the Information Soup, 3rd Conference of the Association for Machine Translation in the Americas (AMTA'98)*, pages 1–17, Langhorne, PA, USA. Cited pp. 10, 11, 13, 20, 21, 24, 29, 82 et 109.

[Fung et Cheung, 2004] Fung, P. et Cheung, P. (2004). Mining very-non-parallel corpora : Parallel sentence and lexicon extraction via bootstrapping and EM. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'04)*, pages 57–63, Barcelona, Spain. Cited p. 12.

[Fung et Mckeown, 1997] Fung, P. et Mckeown, K. (1997). Finding terminology translations from non-parallel corpora. *In Proceedings of the 5th Annual Workshop on Very Large Corpora (VLC'97)*, pages 192–202, Hong Kong. Cited pp. 9, 17, 18 et 62.

[Fung et Yee, 1998] Fung, P. et Yee, L. Y. (1998). An ir approach for translating new words from non parallel, comparable texts. *In Proceedings of the 17th international conference on Computational linguistics (COLING'98)*, pages 414–420, Quebec, Canada. Cited pp. 13 et 110.

[Gamallo, 2008a] Gamallo, O. (2008a). Evaluating two different methods for the task of extracting bilingual lexicons from comparable corpora. *In Proceedings of LREC 2008 Workshop on Comparable Corpora (LREC'08)*, pages 19–26, Marrakech, Marroco. Cited pp. 25, 26, 50, 51, 67, 76, 81, 82, 86 et 95.

[Gamallo, 2008b] Gamallo, O. (2008b). The meaning of syntactic dependencies. *Linguistik Online.* Cited p. 51.

[Garera *et al.*, 2009] Garera, N., Callison-burch, C. et Yarowsky, D. (2009). Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences. Cited pp. 25 et 50.

[Gaussier *et al.*, 2004] Gaussier, E., Renders, J.-M., Matveeva, I., Goutte, C. et Déjean, H. (2004). A geometric view on bilingual lexicon extraction from comparable corpora. *In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04)*, pages 526–533, Barcelona, Spain. Cited pp. 13, 30, 31 et 130.

[Gillard *et al.*, 2007] Gillard, L., Bellot, P. et El-Bèze, M. (2007). D'une compacité positionnelle à une compacité probabiliste pour un système de questions / réponses. *In Proceedings of the 4th French Information Retrieval Conference (CORIA'07)*, pages 271–286, Saint-Étienne, France. Cited p. 120.

[Goeuriot, 2009] Goeuriot, L. (2009). *Découverte et caractérisation des corpus comparables spécialisés.* Thèse de doctorat, Departament of Computer Science, Lina, Nantes. Cited pp. 11 et 35.

[Goldman et Payne, 1981] Goldman, A. et Payne, E. (1981). *A Taxonomic approach to the Lexis of Science.* In L.Selinker et al. Cited p. 46.

[Good, 1953] Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40:16–264. Cited p. 99.

[Grefenstette, 1993] Grefenstette, G. (1993). Automatic thesaurus generation from raw text using knowledge-poor techniques. in making sense of words. *In Proceedings of the 9th Annual Conference of the UW Centre for the OED and Text Research.* Cited p. 48.

[Grefenstette, 1994a] Grefenstette, G. (1994a). Corpus-Derived First, Second and Third-Order Word Affinities. *In Proceedings of the 6th Congress of the European Association for Lexicography (EURALEX'94)*, pages 279–290, Amsterdam, The Netherlands. Cited pp. 47, 48 et 62.

[Grefenstette, 1994b] Grefenstette, G. (1994b). *Explorations in Automatic Thesaurus Discovery.* Kluwer Academic Publisher, Boston, MA, USA. Cited pp. 21, 30, 63 et 129.

[Grefenstette, 1999] Grefenstette, G. (1999). The World Wide Web as a resource for example-based machine translation tasks. *In Translating and the Computer 21 : Proceedings of the 21st International Conference on Translating and the Computer.* Cited pp. 32 et 33.

[Groc, 2011] Groc, C. D. (2011). Babouk : Focused Web Crawling for Corpus Compilation and Automatic Terminology Extraction. *In Proceedings of The IEEE WICACM International Conferences on Web Intelligence*, pages 497–498, Lyon, France. Cited p. 64.

[Haghighi *et al.*, 2008] Haghighi, A., Percy, L., Taylor, B.-K. et Dan, K. (2008). Learning bilingual lexicons from monolingual corpora. *In Proceedings of the 46nd Annual Meeting of the Association for Computational Linguistics (ACL'08)*, pages 771–779, Columbus, Ohio. Cited pp. 14, 130 et 153.

[Harastani *et al.*, 2012] Harastani, R., Daille, B. et Morin, E. (2012). Neoclassical compound alignments from comparable corpora. *In Proceedings of the 13th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'12)*, pages 72–82. Cited pp. 34 et 35.

[Harastani *et al.*, 2013] Harastani, R., Daille, B. et Morin, E. (2013). Identification, alignement, et traductions des adjectifs relationnels en corpus comparables. *In Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'13)*, pages 313–326, Les Sables d'Olonne, France. Cited pp. 34 et 35.

[Harris, 1971] Harris, Z. S. (1971). *Structures mathématiques du langage.* Dunod. Traduit de l'Américain par C. Fuchs. Cited p. 62.

[Hazem et Morin, 2012] Hazem, A. et Morin, E. (2012). Ica for bilingual lexicon extraction from comparable corpora. *In Proceedings of the 5th Workshop on Building and Using Comparable Corpora : Comparable Corpora and the Web (BUCC'12)*, Istanbul, Turkey. Cited p. 158.

[Herbert, 1965] Herbert, A. (1965). *The Structure of Technical English.* London : Longman. Cited p. 46.

[Hindle, 1990] Hindle, D. (1990). Noun classification from predicate-argument structures. *In Proceedings of the 28th Meeting of the Association for Computational Linguistics (ACL'90)*, pages 268–275. Cited p. 49.

[Hoffmann, 1985] Hoffmann, L. (1985). *Kommunikationsmittel Fachsprache.* Tubingen : Gunter Narr Verlag. Cited pp. 45 et 46.

[Hofmann, 1999] Hofmann, T. (1999). Probabilistic latent semantic analysis. *In UAI*, pages 289–296. Cited p. 30.

[Hotelling, 1936] Hotelling, H. (1936). Relation between two sets of variates. 28(3):321–377. Cited p. 157.

[Hunsicker *et al.*, 2012] Hunsicker, S., Ion, R. et Stefanescu, D. (2012). Hybrid parallel sentence mining from comparable corpora. *In Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT'12)*, Trento, Italy. Cited p. 12.

[Hyvarinen, 1999] Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component a analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634. Cited p. 132.

[Hyvarinen *et al.*, 2001] Hyvarinen, A., Karhunen, J. et Oja, E. (2001). Independent component analysis. *John Wiley Sons.* Cited p. 132.

[Ismail et Manandhar, 2010] Ismail, A. et Manandhar, S. (2010). Bilingual lexicon extraction from comparable corpora using in-domain terms. *In Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 481–489. Cited pp. 23 et 57.

[ISO 1087-1 :2000, 2000] ISO 1087-1 :2000, . (2000). *Terminology - Vocabulary.* International Standard Norme. Cited p. 45.

[ISO 1087-1990 :5, 1990] ISO 1087-1990 :5, . (1990). *Terminology - Vocabulary.* International Standard Norme. Cited p. 45.

[Jeffreys, 1948] Jeffreys, H. (1948). *Theory of Probability.* Clarendon Press, Oxford. Cited p. 98.

[Johansson et Hofland, 1989] Johansson, S. et Hofland, K. (1989). *Frequency Analysis of English Vocabulary and Grammar : Based on the Lob Corpus Volume 1 : Tag Frequencies and Word Frequencies.* Clarendon Press. Cited p. 36.

[Johnson, 1932] Johnson, W. (1932). Probability : the deductive and inductive problems. *Mind*, 41(164):409–423. Cited p. 98.

[Jones et Martin, 1997] Jones, M. P. et Martin, J. H. (1997). Contextual spelling correction using latent semantic analysis. *In ANLP*, pages 166–173. Cited p. 129.

[Jutten et Hérault, 1991] Jutten, C. et Hérault, J. (1991). Blind separation of sources. part i. an adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10. Cited p. 132.

[Karlgren et Sahlgren, 2001] Karlgren, J. et Sahlgren, M. (2001). From words to understanding. *In Foundations of Real-World Intelligence*, pages 294–308. Cited p. 130.

[Katz, 1987] Katz, S. M. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(3):400–401. Cited pp. 99 et 100.

[Kilgarriff, 2001] Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):1–37. Cited pp. 35 et 36.

[Kneser et Ney, 1995] Kneser, R. et Ney, H. (1995). Improved backing-off for M-gram language modeling. *In Proceedings of the 20th International Conference on Acoustics, Speech, and Signal Processing (ICASSP'95)*, pages 181–184, Michigan, USA. Cited p. 100.

[Koehn, 2004] Koehn, P. (2004). Europarl : A parallel corpus for statistical machine translation. *In Proceedings of the Machine Translation Summit (MT Summit'05)*, pages 1–8. Cited p. 10.

[Koehn et Knight, 2002] Koehn, P. et Knight, K. (2002). Learning a translation lexicon from monolingual corpora. *In Proceedings of the workshop of the ACL Special Interest Group on the Lexicon (SIGLEX)*, pages 9–16. Cited p. 22.

[Korenius *et al.*, 2006] Korenius, T., Laurikkala, J., Juhola, M. et Järvelin, K. (2006). Hierarchical clustering of a finnish newspaper article collection with graded relevance assessments. *Inf. Retr.*, 9(1):33–53. Cited p. 134.

[Koutsoudas et Humecky, 1957] Koutsoudas, A. et Humecky, A. (1957). *Ambiguity of syntactic function resolved by linear context.* Word. Cited p. 9.

[Landauer et Dumais, 1997] Landauer, T. K. et Dumais, S. T. (1997). A solution to plato's problem : The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, pages 211–240. Cited pp. 129 et 130.

[Laroche et Langlais, 2010] Laroche, A. et Langlais, P. (2010). Revisiting context-based projection methods for term-translation spotting in comparable corpora. *In Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 617–625, Beijing, China. Cited pp. 8, 13, 20, 24, 56, 75, 81, 82 et 86.

[Lewis *et al.*, 1967] Lewis, P. A. W., Baxendale, P. B. et Bennett, J. L. (1967). Statistical discrimination of the synonymy/antonymy relationship between words. *J. ACM*, 14(1):20–44. Cited p. 30.

[L'Homme, 2004] L'Homme, M.-C. (2004). *La Terminologie : Principes et Techniques.* La Presses de l'Université de Montréal. Cited pp. 46 et 147.

[Li et Gaussier, 2010] Li, B. et Gaussier, É. (2010). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. *In Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 644–652, Beijing, China. Cited pp. 14, 35, 36 et 77.

[Li *et al.*, 2011] Li, B., Gaussier, E., Morin, E. et Hazem, A. (2011). Degré de comparabilité, extraction lexicale bilingue et recherche d'information interlingue. *In Actes de la 18ème Conférence Traitement Automatique des Langues Naturelles (TALN'11)*, pages 283–293, Montpellier, France. Cited p. 14.

[Lidstone, 1920] Lidstone, G. J. (1920). Note on the general case of the bayes-laplace formula for inductive or a posteriori probabilities. *Transactions of the Faculty of Actuaries*, 8:182–192. Cited p. 98.

[Lin, 1998a] Lin, D. (1998a). Automatic retrieval and clustering of similar words. *In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 768–774, Montréal, Quebec, Canada. Cited p. 129.

[Lin, 1998b] Lin, D. (1998b). Dependency-based evaluation of minipar. *In Proceedings of the Workshop on the Evaluation of Parsing Systems, 1st International Conference on Language Resources and Evaluation (LREC'98)*, Granada, Spain. Cited p. 50.

[Liu et Zhang, 2013] Liu, S. et Zhang, C. (2013). Termhood-based comparability metrics of comparable corpus in special domain. *In Proceedings of the 13th Chinese conference on Chinese Lexical Semantics (CLSW'12)*, pages 134–144, Wuhan, China. Cited p. 37.

[Lowe et Mcdonald, 2000] Lowe, W. et Mcdonald, S. (2000). *The direct route : mediated priming in semantic space*, pages 806–811. Cited p. 130.

[Lund et Burgess, 1996] Lund, K. et Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods*, 28(2):203–208. Cited p. 130.

[Lund *et al.*, 1995] Lund, K., Burgess, C. et Atchley, R. A. (1995). Semantic and associative priming in high-dimensional semantic space. *In Proceedings of the 17th Annual Conference of the Cognitive Science Society*, pages 660–665. Cited p. 130.

[Maia, 2003] Maia, B. (2003). What are comparable corpora. *In Proceedings of the Corpus Linguistics workshop on Multilingual Corpora : Linguistic requirements and technical perspectives*, Lancaster, U.K. Cited p. 35.

[Manning, 1993] Manning, C. D. (1993). Automatic acquisition of a large subcategorization dictionary from corpora. *In Proceedings of the 31st Meeting of the Association for Computational Linguistics (ACL'93)*, pages 235–242. Cited p. 48.

[McCarthy *et al.*, 2004] McCarthy, D., Koeling, R., Weeds, J. et Carroll, J. A. (2004). Finding predominant word senses in untagged text. *In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04)*, pages 279–286, Barcelona, Spain. Cited p. 129.

[McDonald, 2000] McDonald, S. (2000). *Environmental Determinants of Lexical Processing Effort*. Thèse de doctorat, University of Edinburgh. Cited p. 129.

[McDonald et Brew, 2004] McDonald, S. et Brew, C. (2004). A distributional model of semantic context effects in lexical processing. *In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04)*, pages 17–24, Barcelona, Spain. Cited p. 130.

[McEnery et Xiao, 2007a] McEnery, A. et Xiao, Z. (2007a). Parallel and comparable corpora ? *In Incorporating Corpora : Translation and the Linguist. Translating Europe. Multilingual Matters.* Cited p. 35.

[McEnery et Xiao, 2007b] McEnery, A. et Xiao, Z. (2007b). *Parallel and comparable corpora : What is happening.* Cited p. 9.

[Mercer et Jelinek, 1980] Mercer, L. et Jelinek, F. (1980). Interpolated estimation of markov source parameters from sparse data. *In Proceedings of the Workshop on pattern recognition in Practice*, Amsterdam, The Netherlands. Cited p. 99.

[Morin, 2009] Morin, E. (2009). Apport d'un corpus comparable déséquilibré à l'extraction de lexiques bilingues. *In Actes de la 16ème Conférence Traitement Automatique des Langues Naturelles (TALN'09)*, Senlis, France. Cited pp. 11, 31, 81, 82, 83 et 86.

[Morin et Daille, 2004] Morin, E. et Daille, B. (2004). Extraction terminologique bilingue à partir de corpus comparables d'un domaine spécialisé. *Traitement Automatique des Langues. TAL*, 45(3):103–122. Cited pp. 8 et 14.

[Morin et Daille, 2010] Morin, E. et Daille, B. (2010). Compositionality and lexical alignment of multi-word terms. *Language Resources and Evaluation*, 44(1-2):79–95. Cited pp. 32, 33 et 34.

[Morin *et al.*, 2004] Morin, E., Dufour-Kowalski, S. et Daille, B. (2004). Extraction de terminologies bilingues à partir de corpus comparables. *In Actes de la 11ème Conférence Traitement Automatique des Langues Naturelles (TALN'04)*, pages 309–318, Fès, Maroc. Cited pp. 8 et 13.

[Morin et Prochasson, 2011] Morin, E. et Prochasson, E. (2011). Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. *In Proceedings of the 4th Workshop on Building and Using Comparable Corpora : Comparable Corpora and the Web (BUCC'11)*, pages 27–34, Portland, Oregon. Cited p. 12.

[Munteanu et Marcu, 2005] Munteanu, D. S. et Marcu, D. (2005). Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31:477–504. Cited pp. 12 et 35.

[Munteanu et Marcu, 2006] Munteanu, D. S. et Marcu, D. (2006). Extracting parallel sub-sentential fragments from non-parallel corpora. *In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 81–88, Stroudsburg, PA, USA. Cited pp. 12 et 35.

[Nakagawa et Mori, 2003] Nakagawa, H. et Mori, T. (2003). Automatic term recognition based on statistics of compound nouns and their components. *Terminology*, 9(2):201–219. Cited p. 33.

[Oard et Diekema, 1998] Oard, D. et Diekema, A. (1998). Cross-language information retrieval. *Annual Review of Information Science (ARIST)*. Cited p. 8.

[Och et Ney, 2000] Och, F. J. et Ney, H. (2000). Improved statistical alignment models. *In Proceedings of the 38th Annual Meeting on Association for Computational Linguistics (ACL'00)*, pages 440–447, Hong Kong. Cited p. 38.

[Otero, 2007] Otero, P. G. (2007). Learning bilingual lexicons from comparable english and spanish corpora. *In Proceedings of Machine Translation Summit XI*, pages 191–198. Cited p. 62.

[Patry et Langlais, 2010] Patry, A. et Langlais, P. (2010). Paradocs : A language independant go-between for mating parallel documents. *TAL*, 51(2):41–63. Cited p. 10.

[Pearson, 1998] Pearson, J. (1998). *Terms in context.* Dublin City University. Cited p. 44.

[Pekar *et al.*, 2006] Pekar, V., Mitkov, R., Blagoev, D. et Mulloni, A. (2006). Finding translations for low-frequency words in comparable corpora. *Machine Translation*, 20(4):247–266. Cited pp. 21, 24 et 97.

[Prochasson, 2010] Prochasson, E. (2010). *Extraction lexicale bilingue à partir de corpus comparables.* Thèse de doctorat, Departament of Computer Science, Lina, Nantes. Cited p. 76.

[Prochasson et Fung, 2011] Prochasson, E. et Fung, P. (2011). Rare word translation extraction from aligned comparable documents. *In Proceedings of the 49nd Annual Meeting of the Association for Computational Linguistics (ACL'11)*, pages 1327–1335, Portland, Oregon, USA. Cited p. 24.

[Prochasson et Morin, 2009] Prochasson, E. et Morin, E. (2009). Influence des points d'ancrage pour l'extraction lexicale bilingue à partir de corpus comparables spécialisés. *In Actes de la 16ème Conférence Traitement Automatique des Langues Naturelles (TALN'09)*, Senlis, France. Cited pp. 27, 57 et 81.

[Rapp, 1995] Rapp, R. (1995). Identify Word Translations in Non-Parallel Texts. *In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL'95)*, pages 320–322, Boston, MA, USA. Cited pp. 8, 9, 11, 13, 14, 16, 17, 23 et 62.

[Rapp, 1999] 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'99)*, pages 519–526, College Park, MD, USA. Cited pp. 8, 13, 20, 21, 23, 29, 62, 81 et 82.

[Rapp, 2002] Rapp, R. (2002). The computation of word associations : Comparing syntagmatic and paradigmatic approaches. *In Proceedings of the 15th International Conference on Computational Linguistics (COLING'02)*, Taipei, Taiwan. Cited pp. 47 et 48.

[Resnik *et al.*, 1999] Resnik, P., Olsen, M. B. et Diab, M. T. (1999). The bible as a parallel corpus : Annotating the 'book of 2000 tongues'. *Computers and the Humanities*, 33(1-2):129–153. Cited p. 10.

[Robitaille *et al.*, 2006] Robitaille, X., Sasaki, Y., Tonoike, M., Sato, S. et Utsuro, T. (2006). Compiling french-japanese terminologies from the web. *In Proceedings of the 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference (EACL'06)*, Trento, Italy. Cited pp. 32, 33 et 34.

[Rondeau, 1984] Rondeau, G. (1984). *Introduction à la terminologie.* Québec : Gaetan Morin. Cited p. 44.

[Rose *et al.*, 1997] Rose, T., Haddock, N. et Tucker, R. (1997). The effects of corpus size and homogeneity on language model quality. Cited pp. 35 et 36.

[Rubino et Linarès, 2011] Rubino, R. et Linarès, G. (2011). Une approche multi-vue pour l'extraction terminologique bilingue. *In Actes de la 8ème Conférence en Recherche d'Information et Applications (CORIA'11)*, pages 97–111, Avignon, France. Cited p. 130.

[Sadat *et al.*, 2003] Sadat, F., Yoshikawa, M. et Uemura, S. (2003). Learning bilingual translations from comparable corpora to cross-language information retrieval : hybrid statistics-based and linguistics-based approach. *In IRAL*, pages 57–64. Cited pp. 23 et 25.

[Sager, 1990] Sager, J. (1990). *A Practical Course in Terminology Processing.* Amsterdam : John Benjamins Publishing Company. Cited p. 44.

[Sahlgren et Cöster, 2004] Sahlgren, M. et Cöster, R. (2004). Using bag-of-concepts to improve the performance of support vector machines in text categorization. *In Proceedings of the 20th International Conference on Computational Linguistics (COLING'04)*, Geneva, Switzerland. Cited p. 129.

[Salton et Lesk, 1968] Salton, G. et Lesk, M. E. (1968). Computer evaluation of indexing and text processing. *Journal of the Association for Computational Machinery*, 15(1):8–36. Cited pp. 21 et 63.

[Salton *et al.*, 1975] Salton, G., Wong, A. et Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620. Cited p. 129.

[Saralegi *et al.*, 2008] Saralegi, X., Vicente, I. et Gurrutxaga, A. (2008). Automatic extraction of bilingual terms from comparable corpora in a popular science domain. *In Proceedings of the Workshop on Comparable Corpora (LREC'08)*, Marrakech, Morocco. Cited p. 35.

[Savary et Jacquemin, 2000] Savary, A. et Jacquemin, C. (2000). Reducing information variation in text. *In Proceedings of the 8th ELSNET Summer School Text- and Speech-Triggered Information Access*, pages 145–181, Chios Island, Greece. Cited p. 32.

[Schütze, 1993] Schütze, H. (1993). Word space. *In* Hanson, S. J., Cowan, J. D. et Giles, C. L., éditeurs : *Advances in Neural Information Processing Systems 5*, pages 895–902. Morgan Kaufmann. Cited pp. 49 et 129.

[Schütze, 1998] Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123. Cited p. 129.

[Schütze et Pedersen, 1993] Schütze, H. et Pedersen, T. (1993). A vector model for syntagmatic and paradigmatic relatedness. *In Proceedings of the UW Centre for the OED and Text Research.* Cited p. 49.

[Sekine, 1997] Sekine, S. (1997). The domain dependence of parsing. *In Proceedings of the 5th conference on Applied natural language processing (ANLC'97)*, pages 96–102, Washington, DC. Cited p. 36.

[Shao et Ng, 2004] Shao, L. et Ng, H. T. (2004). Mining new word translations from comparable corpora. *In Proceedings of the 21st International Conference on Computational Linguistics (COLING'04).* Cited p. 23.

[Sharoff, 2007] Sharoff, S. (2007). Classifying web corpora into domain and genre using automatic feature identification. *In Proceedings of the 3rd Web as Corpus Workshop*, Louvain-la-Neuve, Belgium. Cited p. 35.

[Sinclair, 1996] Sinclair, J. (1996). Preliminary recommendations on corpus typology. Cited p. 9.

[Smajda, 1993] Smajda, F. (1993). *Retrieving Collocations from Text : Xtract.*, volume 19. Computational Linguistics. Cited p. 48.

[Smith *et al.*, 2010] Smith, J. R., Quirk, C. et Toutanova, K. (2010). Extracting parallel sentences from comparable corpora using document level alignment. *In Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT'10)*, pages 403–411, Los Angeles, California. Cited pp. 12 et 36.

[Su et Babych, 2012] Su, F. et Babych, B. (2012). Measuring comparability of documents in non-parallel corpora for efficient extraction of (semi-)parallel translation equivalents. *In Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, pages 10–19, Avignon, France. Cited pp. 35 et 38.

[Tanaka, 2002] Tanaka, T. (2002). Measuring the similarity between compound nouns in different languages using non-parallel corpora. *In Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, Taipei, Taiwan. Cited p. 33.

[Teubert, 1996] Teubert, W. (1996). Comparable or parallel corpora ? *International Journal of Lexicography*, 9(3):238–276. Cited p. 11.

[Trimble et Trimble, 1978] Trimble, L. et Trimble, R. (1978). *The Development of EFL Materials for Occupational English : The technical manual.* Oregon State University. Cited pp. 45 et 46.

[Voorhees, 2002] Voorhees, E. M. (2002). Overview of the trec 2002 question answering track. *In TREC*. Cited p. 120.

[Véronis et Langlais, 2000] Véronis, J. et Langlais, P. (2000). Evaluation of parallel text alignement systems. *The Arcade Project, In Jean Veronis (ed), Parallel Text Processing - Alignement and Use of Translation Corpora, Kluwer Academic Publishers.* Cited pp. 8, 9 et 10.

[Weller *et al.*, 2011] Weller, M., Gojun, A., Heid, U., Daille, B. et Harastani, R. (2011). Simple methods for dealing with term variation and term alignment. *In Proceedings of the 9th International Conference on Terminology and Artificial Intelligence*, Paris, France. Cited p. 33.

[Wettler et Rapp, 1993] Wettler, M. et Rapp, R. (1993). Computation of word associations based on the co-occurrences of words in large corpora. *In Proceedings of the Workshop on Very Large Corpora (VLC'93)*, volume:84–93. Cited p. 14.

[Wong, 2009] Wong, W. Y. (2009). *Learning lightweight ontologies from text across different domains using the web as background knowledge.* Thèse de doctorat, University of Western Australia. Cited p. 24.

[Wuster, 1968] Wuster, E. (1968). *The machine tool : An Interlingual dictionary of basic concepts.* London : Technical Press. Cited p. 44.

[Yang, 1986] Yang, H. (1986). *A new technique for identifying scientific an technical terms and describing science texts. Literary and Linguistic Computing.* Oxford : Oxford University Press. Cited p. 46.

[Yu et ichi Tsujii, 2009] Yu, K. et ichi Tsujii, J. (2009). Extracting bilingual dictionary from comparable corpora with dependency heterogeneity. *In Proceedings of Human Language Technologies : Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL'09) (Short Papers)*, pages 121–124. Cited p. 26.

[Zipf, 1949] Zipf, G. K. (1949). *Human Behaviour and the Principle of Least Effort : an Introduction to Human Ecology.* Addison-Wesley. Cited pp. 97 et 130.

[Zweigenbaum et Habert, 2006] Zweigenbaum, P. et Habert, B. (2006). Faire se rencontrer les parallèles : regards croisés sur l'acquisition lexicale monolingue et multilingue. *GLOTTOPOL*, 8:22–44. Cited p. 83.

l'université nantes angers le mans
PÔLE DE RECHERCHE ET D'ENSEIGNEMENT SUPÉRIEUR

UNIVERSITÉ DE NANTES

Thèse de Doctorat

Amir HAZEM

Extraction de lexiques bilingues à partir de corpus comparables

Résumé

La plupart des travaux en acquisition de lexiques bilingues à partir de corpus comparables reposent sur l'hypothèse distributionnelle qui a été étendue au scénario bilingue. Deux mots ont de fortes chances d'être en relation de traduction s'ils apparaissent dans les mêmes contextes lexicaux. Ce postulat suppose donc une définition claire et rigoureuse du contexte et une connaissance parfaite des indices contextuels. Or, la complexité et les spécificités de chaque langue font qu'il n'est pas aisé d'énoncer une telle définition qui garantisse une extraction de couples de traductions, efficace dans tous les cas de figure. Toute la difficulté réside dans la manière de définir, d'extraire et de comparer ces contextes dans le but de construire des lexiques bilingues fiables. Nous nous efforcerons tout au long des différents chapitres de cette thèse à essayer de mieux comprendre cette notion de contexte, pour ensuite l'étendre et l'adapter afin d'améliorer la qualité des lexiques bilingues. Une première partie des contributions vise à améliorer l'approche directe qui fait office de référence dans la communauté. Nous proposerons plusieurs manières d'aborder le contexte des mots pour mieux les caractériser. Dans la deuxième partie des contributions, nous commencerons par présenter une approche qui vise à améliorer l'approche par similarité inter-langue. Ensuite, une méthode nommée Q-Align, directement inspirée des systèmes de question/réponse sera présentée. Enfin, nous présenterons plusieurs transformations mathématiques et donc plusieurs représentations vectorielles, pour nous concentrer essentiellement sur celles que nous aurons choisi pour développer une nouvelle méthode d'alignement.

Mots-clés

Corpus comparables, extraction terminologique bilingue, vecteurs de contexte.

Abstract

Most work in bilingual lexicon acquisition from comparable corpora are based on the distributional hypothesis that has been extended to the bilingual scenario. Hence, two words are more likely to be translation of each other if they appear in the same lexical contexts. This assumption presupposes a clear and rigorous definition of context and a thorough knowledge of contextual clues. However, the complexity and specificity of each language make the formulation of such a definition that ensures effective extraction of translation pairs in all cases not easy. All the difficulty lies in how to define, extract and compare these contexts in order to build reliable bilingual lexicons. We strive throughout the different chapters of this thesis to try to understand this notion of context, and then extend and adapt it to improve the quality of bilingual lexicons. The first part of contributions aims at improving the standard approach considered as a baseline in the community. Thus, we propose several ways to consider the context for better words characterization. In the second part of the contributions, we first present an approach that aims to improve the extended approach. Then, a method called Q-Align directly inspired from question/answering systems is presented. Finally, we present several mathematical transforms and thus multiple vector space representations to focus primarily on the ones we have chosen to develop a new alignment method.

Keywords

Comparable corpora, bilingual terminology extraction, context vectors.