Constitution de ressources linguistiques multilingues à partir de corpus de textes parallèles et comparables

Dhouha Bouamor

UNIVERSITÉ PARIS SUD
ÉCOLE DOCTORALE D'INFORMATIQUE
CEA-LIST et LIMSI-CNRS

THÈSE

présentée pour obtenir le grade de
DOCTEUR DE L'UNIVERSITÉ DE PARIS SUD
Spécialité : Informatique
par :

Dhouha BOUAMOR

Titre :

Constitution de ressources linguistiques multilingues à partir de corpus de textes parallèles et comparables.

JURY

Rapporteur	Reinhard RAPP	Professeur, Université de Mainz
Rapporteur	Éric GAUSSIER	Professeur, Université J. Fourier
Examinateur	Philippe LANGLAIS	Professeur, Université de Montréal
Examinateur	François YVON	Professeur, Université Paris Sud
Directeur	Pierre ZWEIGENBAUM	Directeur de recherches, CNRS
Encadrant	Nasredine SEMMAR	Chercheur, CEA-LIST

soutenue le 21/02/2014

Résumé

Les lexiques bilingues sont des ressources particulièrement utiles pour la Traduction Automatique et la Recherche d'Information Translingue. Leur construction manuelle nécessite une expertise forte dans les deux langues concernées et est un processus coûteux. Plusieurs méthodes automatiques ont été proposées comme une alternative, mais elles qui ne sont disponibles que dans un nombre limité de langues et leurs performances sont encore loin derrière la qualité des traductions manuelles. Notre travail porte sur l'extraction de ces lexiques bilingues à partir de corpus de textes *parallèles* et *comparables*, c'est à dire la reconnaissance et l'alignement d'un vocabulaire commun multilingue présent dans ces corpus.

En nous basant sur des corpus parallèles, nous présentons une approche qui porte sur le traitement d'expressions polylexicales, allant de leur acquisition automatique à leur intégration dans un système de traduction automatique statistique. Notre intérêt se porte par ce type d'unités car, en plus du fait qu'elles soient fréquemment utilisées dans le langage oral et écrit de tous les jours ainsi que dans les communications spécialisées techniques et scientifiques, leur identification est fondamentale pour les applications faisant intervenir les aspects sémantiques de la langue et surtout la traduction automatique.

Pour les corpus comparables, nous proposons deux approches innovantes dont le but est d'extraire des lexiques bilingues spécialisés dans les domaines de la *finance des entreprises*, du *cancer du sein*, de *l'énergie éolienne* et de la *technologie mobile*. La première approche étend l'approche distributionnelle par un processus de désambiguïsation lexicale. Le but de cette approche est de ne garder que les éléments du contexte les plus susceptibles de donner la meilleure représentation du mot à traduire. Notre deuxième approche repose sur Wikipédia et l'analyse explicite sémantique. L'originalité de cette approche réside dans le fait que, au lieu de considérer l'*espace des mots d'un corpus* pour la représentation des mots que l'on souhaite traduire, ces derniers sont représentés dans l'espace des titres des articles de Wikipédia. Les approches nouvellement introduites se comparent favorablement aux méthodes existantes dans la plupart des configurations testées.

Mots clés : *extraction lexique bilingue, corpus parallèle, corpus comparable, alignement, traduction automatique statistique.*

Abstract

Bilingual lexicons are central components of machine translation and cross-lingual information retrieval systems. Their manual construction requires extensive expertise in both languages involved and it is a costly process. Several automatic methods were proposed as an alternative but they often rely of resources available in a limited number of languages and their performances are still far behind the quality of manual translations. Our work concerns bilingual lexicon extraction from multilingual parallel and comparable corpora, in other words, the process of finding translation pairs among the common multilingual vocabulary available in such corpora.

Based on parallel corpora, we present an approach that focuses on processing multiword expressions, ranging from their automatic acquisition to their integration into a statistical machine translation system. We focus on such units because, besides the fact that they are commonly used in the oral and written language as well as the specialized technical and scientific communications, their identification is crucial for applications involving semantic aspects of language and especially for machine translation.

For comparable corpora, we propose two innovative approaches that aims at extracting bilingual lexicons specialized in the *corporate finance, breast cancer, wind energy* and *mobile technology* domains. The first approach augments the distributional approach by proposing a word sense disambiguation process that keeps only the words that are more likely to give the best representation of a word to be translated. Our second approach is based on Wikipedia and its explicit semantic analysis. The main originality of this approach is in the way words are represented : instead of representing the words in a corpus words space, they are represented in Wikipedia titles' space. The newly introduced methods compare favorably to existing methods in almost configurations tested.

Keywords : *bilingual lexicon extraction, parallel corpus, comparable corpus, alignment, statistical machine translation.*

Remerciements

Je tiens à remercier Pierre Zweigenbaum, mon directeur de thèse, pour sa bonne humeur, sa disponibilité, ses encouragements, son soutien moral et son encadrement.

En parlant de l'encadrement, je remercie Nasredine Semmar, mon co-directeur, pour ses conseils utiles. Ses conseils m'ont permis de prendre les bonnes décisions et mener à bien ce travail.

Je voudrais remercier mes rapporteurs Éric Gaussier et Reinhard Rapp pour l'intérêt qu'ils ont porté à mon travail ainsi que les remarques et les suggestions qu'ils m'ont faites.

Je souhaite aussi remercier Philippe Langlais et François Yvon qui m'a fait l'honneur de présider mon jury de thèse.

Mais cette thèse s'est également inscrite dans un environnement humain particulièrement chaleureux. Je pense à toutes ces personnes grâce à qui je ne me suis jamais senti seule, personnes qui m'ont accompagné au long de ces années de travail et particulièrement Adrian Popescu avec qui j'ai eu l'occasion de travailler.

Je conclurai donc en remerciant tous amis. Merci à toutes et à tous.

J'aimerais aussi remercier tous les membres du LVIC dans leur ensemble pour leur bonne humeur et les discussions agréables qu'on a eu.

Last and not least, je tiens à remercier mes parents Messaoud et Mélika pour tout ce qu'ils m'ont inculqué et appris et sans qui je ne serai pas arrivé là. Un grand merci du fond du coeur à mon fiancé Khaled pour ses encouragements et son soutien moral

continu. Je remercie également chaleureusement mes frères Noureddine, Mohamed Ali et Aladin. Un très grand merci aussi à mes soeurs Wafa et Houda, à ma tante Yosra, à ma belle-soeur Manel et mon petit coeur Dorra. Un grand merci à toute ma famille.

Et je termine enfin en remerciant tous ceux que j'ai pu oublier.

Table des matières

Liste des tableaux

Table des figures

Introduction

So many languages, so few resources. How to bridge the gap ?
Mike Maxwell
Linguistic Data Consortium

Le langage naturel est le mode privilégié par les humains pour communiquer entre eux, de manière parlée, signée ou écrite, ne cesse d'évoluer depuis des dizaines de milliers d'années. Cette évolution a donné lieu à une grande variété de langues en usage aujourd'hui qui renferment un large éventail de phénomènes linguistiques et de caractéristiques faisant l'objet d'études très diverses. Afin de préserver toute forme de langues dans le monde, ces dernières doivent non seulement être apprises et utilisées constamment, mais elles doivent également être documentées et mises en relation.

Avec l'explosion d'Internet ces dernières décennies, des milliers de pages Web représentant tout autant de textes dans différentes langues exploitables et accessibles instantanément ont été mises en ligne. Parallèlement à cette inflation documentaire, des genres textuels nouveaux apparaissent et l'on voit émerger un *multilinguisme* dont l'ampleur est corrélée à *l'internetisation* grandissante du monde. L'information textuelle électronique disponible est donc volumineuse, diversifiée et multilingue.

Pour des raisons variées, diverses communautés s'intéressent de plus en plus aux données textuelles multilingues. Les historiens, les juristes, les philologues analysent les corpus multilingues avec des outils d'exploration permettant d'observer plus finement les correspondances entre différents volets de corpus. En linguistique computationnelle, ces données multilingues sont utilisées dans plusieurs applications à savoir la traduction automatique (TA), la recherche d'information interlingue, ou encore pour construire des *lexiques bilingues*. Ces lexiques rassemblent des unités lexicales (mots

simples, mots complexes) en relation de traduction. Ces ressources traductionnelles sont généralement utilisées avec profit dans différentes applications du Traitement Automatique des Langues (TAL) qui s'étendent de la linguistique contrastive à la lexicographie et de la TA à la recherche d'information interlingue.

La richesse terminologique d'une langue est liée aux activités de l'homme et est en perpétuelle évolution, avec un dynamisme parfois surprenant. Or, le rythme de création de ces ressources est beaucoup plus faible que le rythme de création des *néologismes* lui-même corrélé au rythme effréné de production des textes surtout pour des domaines de spécialité. Le but de cette thèse est donc de construire et d'actualiser ce type de ressources linguistiques qui constituent des enjeux majeurs, vu la divérsite inter et intra langues.

La recherche académique et industrielle est actuellement très dynamique, et de nombreux travaux se sont intéressés à cette problématique. La première solution envisagée consiste alors à observer la terminologie en situation, en se basant sur des *textes bilingues parallèles*. Ces textes parallèles représentent des textes dans une langue source et leurs traductions en langue cible. Les approches qui sont à l'origine des méthodes d'extraction d'équivalents de traduction à partir de ce type de corpus sont issues du domaine de la TA statistique (TAS). À la base de l'approche probabiliste on trouve des modèles statistiques élaborés dans le cadre de recherches sur la TA. À l'origine, le développement de ces modèles visait à générer d'un texte cible à partir d'un texte source. Par la suite, ces calculs probabilistes ont été utilisés pour l'alignement ou l'extraction de lexiques bilingues (Brown *et al.*, 1991; Brown *et al.*, 1993). La seule exigence de ces modèles probabilistes est que le corpus parallèle soit aligné au niveau de la phrase.

Si les méthodes les plus efficaces pour construire des lexiques bilingues s'appuient sur des corpus parallèles (Gale et Church, 1991; Koehn et Knight, 2001), la disponibilité de ces corpus reste un problème surtout pour des domaines de spécialité. Les ressources monolingues en domaine de spécialité sont par contre beaucoup plus courantes. Cependant, et contrairement aux corpus parallèles, aucun alignement n'est présent entre deux textes dans deux langues différentes. Extraire des équivalents de traduction à partir de ces corpus dits *corpus comparables* est toutefois possible en se basant sur d'autres critères que l'alignement au niveau des phrases.

L'extraction de lexiques bilingues à partir de corpus comparables fait l'objet d'une attention particulière depuis le milieu des années 1990. Les approches utilisées se basent sur l'hypothèse distributionnelle de Harris (Harris, 1954) 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. La caractérisation d'un mot est généralement enregistrée sous forme d'un vecteur de contexte. Ces vecteurs sont des structures de données contenant des informations sur l'environnement des mots. Ce sont ces informations contextuelles qui sont comparées entre les vecteurs sources et cibles pour obtenir des équivalents de traduction.

Nos contributions portent sur l'extraction de lexiques bilingues à partir de corpus de textes parallèles et comparables. Notre **première contribution** concerne l'extraction lexicale à partir de corpus parallèles. Comme souligné auparavant, cette tâche est bien maîtrisée et plusieurs outils d'alignement sont disponibles (Giza++, Berkeley Aligner). La limitation principale des lexiques extraits à partir de ce type de corpus est leur manque de couverture pour les expressions polylexicales ou expressions à mots multiples. Ce type d'expressions, dans le consensus actuel du domaine du TAL, forment des unités linguistiques complexes telles que *chemin de fer, cordon bleu, New York, etc* (Sag *et al.*, 2002). Leur identification est fondamentale pour les applications faisant intervenir les aspects sémantiques de la langue. Nous présentons une approche qui s'intéresse particulièrement aux expressions polylexicales. Notre approche identifie ce type d'unités lexicales et leurs équivalents de traduction dans un corpus parallèle français-anglais dans le but d'en construire un lexique bilingue.

De même, la prise en compte des expressions polylexicales dans les applications du TAL est cruciale. En TA, la non reconnaissance d'une expression constitue l'une des sources principales d'erreurs (Constant *et al.*, 2011). Par exemple, l'expression française *chemin de fer* se traduit en anglais par le mot *railway*. Un système de TA ne doit donc pas chercher à traduire cette expression par *way of iron*. Notre **deuxième contribution** consiste à étudier l'impact de l'utilisation du lexique bilingue d'expressions polylexicales construit pour un système de TAS : nous explorons différentes façons d'intégrer ce type de connaissances linguistiques à de tels systèmes.

Les contributions restantes portent sur l'acquisition automatique de lexiques terminologiques bilingues *spécialisés* à partir de corpus comparables. Notre intérêt se porte sur quatre domaines divers et variés qui s'étendent de la *finance des entreprises* au *cancer du sein* et de *l'énergie éolienne* à la *technologie mobile* et sur les deux paires de langues *français-anglais* et *roumain-anglais*. Cette variété de domaines et de langues nous permet d'étudier le comportement de différentes approches pour (1) divers domaines de spécialité ayant des vocabulaires très éloignés et pour (2) une paire de langues qui sont richement représentées et pour une autre paire de langues qui comprend le roumain, une langue peu dotée en ressources. Rappelons que comme il a été décrit dans la littérature, cette tâche repose sur la capacité à représenter le contexte d'un mot (Hazem et Morin, 2012a). Nous nous intéressons donc en grande partie à la notion de représentation contextuelle et tentons de répondre à la question suivante : Quelle information enregistrer effectivement dans les vecteurs de contexte ?

C'est dans nos deux dernières contributions que nous essayerons de répondre à cette question. Notre **troisième contribution** augmente l'approche distributionnelle par un processus de désambiguïsation lexicale. Ce processus permet de ne garder que les éléments du contexte les plus susceptibles de donner la meilleure représentation du mot à traduire. Dans la **quatrième contribution**, nous faisons appel à l'analyse explicite sémantique, où au lieu de considérer l'*espace des mots du corpus* pour la représentation des mots que l'on souhaite traduire, ces derniers sont représentés dans un espace de mots particuliers : les titres d'articles de Wikipédia. L'originalité de nos travaux se situe dans l'étude des niveaux de représentation des mots : *lexicale* et *sémantique*.

Organisation du manuscrit :

Ce manuscrit est organisé de la façon suivante :

- Dans le chapitre 1 nous définissons les notions de corpus parallèles et comparables et passons en revue les différents travaux existants concernant l'extraction de lexiques bilingues de différents niveaux de granularité, à partir de ces deux types de corpus.
- Le chapitre 2 est consacré à la présentation de l'approche d'extraction de lexique bilingue d'expressions polylexicales à partir d'un corpus parallèle français-anglais.

- Dans le chapitre 3, nous étudions l'impact de l'utilisation du lexique bilingue obtenu dans un système de traduction automatique statistique.
- Le chapitre 4 présente le contexte global de la tâche d'extraction lexicale à partir de corpus comparables. Nous introduisons également les ressources linguistiques utilisées, notamment les corpus comparables sur lesquels nous avons réalisé nos expériences.
- Le chapitre 5 est centré sur l'approche d'extraction de lexiques bilingues spécialisés à partir de corpus comparables étendant l'approche distributionnelle par l'utilisation de la désambiguïsation sémantique.
- Le chapitre 6 décrit l'approche utilisant l'analyse sémantique explicite pour extraire des lexiques bilingues.
- Enfin, nous présentons les conclusions et les perspectives de cette thèse.

Chapitre 1

État de l'art

1.1 Introduction

Depuis plus d'une décennie, l'extraction de lexiques bilingues constitue un champ d'investigation très animé. Les recherches menées dans ce cadre sont assez vastes et englobent un large éventail de contextes, allant de l'acquisition basée sur les fréquences à celle fondée sur des contextes de mots ou encore sur des dépendances syntaxiques. C'est avec la disponibilité de corpus multilingues constitués de collections de textes dans différentes langues que cette tâche a été rendue possible. Nous distinguons deux types de corpus multilingues : les *corpus parallèles* et les *corpus comparables*. Les corpus parallèles sont composés de paires de documents qui des traductions mutuelles alors que les corpus comparables sont composés de documents ayant des traits communs tels que le genre, la période, le domaine, etc, sans être des traductions. Une définition plus détaillée sera présentée dans la suite.

Ainsi, en fonction du type de corpus utilisé, nous considérons l'extraction de lexiques bilingues comme *supervisée* lorsqu'elle repose sur des données annotées comme les corpus parallèles, *moins supervisée* dans le cas où elle utilise beaucoup moins de données annotées comme les corpus comparables et non *non-supervisée* lorsque des corpus indépendants sont exploités. À cet égard, les approches d'extraction de lexiques bilingues différent en fonction du type de ressources linguistiques

utilisées (Koehn et Knight, 2001). Par exemple, les techniques employées en corpus parallèles ne sont généralement pas applicables aux corpus comparables.

Ce chapitre est consacré à la présentation des différentes approches proposées dans la littérature pour l'extraction de lexiques bilingues à partir de corpus multilingues. Nous présentons tout d'abord dans la section 1.2 les propriétés des différents types de corpus multilingues utilisés dans cette étude. Nous décrivons ensuite dans la section 1.3 différentes approches faisant appel aux corpus parallèles pour l'extraction des lexiques bilingues. La section 1.4 est enfin consacrée à la description des approches qui se basent sur des corpus comparables.

1.2 Corpus multilingues

Les corpus multilingues sont composés de documents dans des langues différentes. Les informations qui peuvent être mises à jour par l'investigation et l'analyse de ces corpus en font une ressource importante pour la traduction automatique, la désambiguïsation sémantique et la recherche d'informations interlingue. En extraction lexicale, ces corpus permettent de suivre automatiquement l'évolution d'une langue et sont utilisés pour créer ou enrichir des lexiques bilingues. Cette section est consacrée à la description d'une typologie des différents corpus multilingues et à la présentation du degré de comparabilité entre ces derniers. Trois types de corpus multilingues ont été définis dans la littérature. Nous distinguons *les corpus parallèles, les corpus comparables* et *les corpus indépendants*.

1.2.1 Corpus parallèles

Les corpus parallèles sont constitués par des paires de documents en relation de traduction. (Somers, 2001) définit les corpus parallèle en tant que textes disponibles dans deux ou plusieurs langues constitués d'un texte original et de sa traduction. À titre d'exemple, les actes du parlement européen (EuroParl), traduits dans 11 langues européennes, et ceux du parlement canadien (Hansard) traduits dans 3 langues font partie des corpus parallèles disponibles. Un exemple de texte parallèle est donné dans la figure 1.1. La *pierre de Rosette* (figure 1.1), constituée par

⋮	⋮
I would congratulate Mrs Peijs on her work. It is a very good report. There are still one or two small points of controversy in it, but she has done an excellent job. I apologise for the confusion we have had in committee over the voting and I am glad that it has all been resolved now.	*Je félicite Mme Peijs de son travail. C'est un très bon rapport, il contient encore un ou deux petits points de controverse, mais elle a réalisé un travail excellent. Je présente mes excuses pour la confusion que nous avons eue en commission concernant le vote et me réjouis que tout soit à présent réglé.*
⋮	⋮

TABLEAU 1.1: Paire de paragraphes parallèles extraite du corpus EUROPARL.

un fragment de stèle gravée de l'Égypte antique portant trois écritures d'un même texte (égyptien en hiéroglyphes, égyptien en écriture démotique et alphabet grec) est considérée comme un corpus parallèle. Cette œuvre a permis à Champollion de déchiffrer l'écriture hiéroglyphique en 1822. Selon (Fung, 1998), un corpus parallèle doit réunir l'ensemble des caractéristiques suivantes :

1. Un mot n'a qu'*un seul sens* dans le corpus.
2. Une *traduction unique* est associée à chaque mot.
3. Il n'y a *pas de traductions manquantes* entre un corpus source et un corpus cible.
4. Les *positions* et *fréquences* des mots en relation de traduction sont *comparables*.

Or, les deux premières caractéristiques ne sont en général pas satisfaites du fait que dans certains corpus parallèles comme EUROPARL, un mot peut se traduire par plusieurs mots et peut avoir plusieurs sens.

Les corpus parallèles constituent donc un élément moteur pour la construction de lexiques bilingues robustes, la traduction automatique et la recherche d'information interlingue. Ils sont généralement construits par des traducteurs humains, qui à leur tour font appel à un lexique bilingue existant pour guider la traduction des textes. Néanmoins, ces corpus sont par nature des ressources rares notamment pour des domaines spécialisés et pour des paires de langues ne faisant pas intervenir l'anglais. (Abdul-Rauf et Schwenk, 2009) constatent par ailleurs que les corpus parallèles les

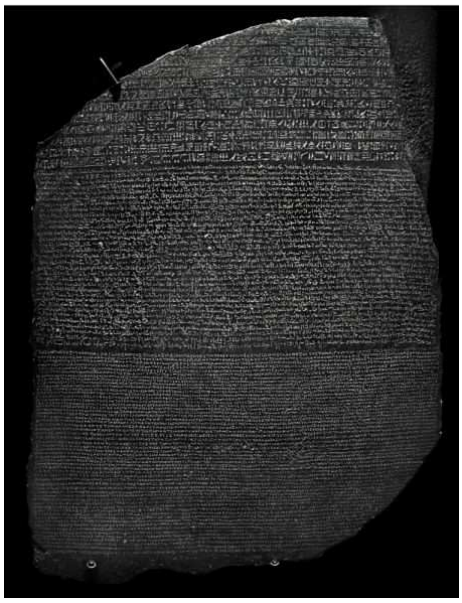


FIGURE 1.1: La pierre de Rosette, un corpus parallèle trilingue *Hiéroglyphe, Démotique* et *Grec*

plus exploités sont généralement caractérisés par un vocabulaire peu utilisé, comme par exemple les corpus HANSARD et EUROPARL.

1.2.2 Corpus comparables

Les corpus comparables rassemblent des documents multilingues n'étant pas en relation de traduction mais partageant des traits communs tels que le domaine, le type de discours, la période, etc. (Déjean et Gaussier, 2002) donnent la définition suivante de corpus comparable :

> « *Deux corpus de deux langues* l_1 *et* l_2 *sont dits comparables s'il existe une sous-partie non négligeable du vocabulaire du corpus de langue* l_1*, respectivement* l_2*, dont la traduction se trouve dans le corpus de langue* l_2*, respectivement* l_1*.* »

À son tour (Ji, 2009) définit les corpus comparables comme des collections de documents décrivant des sujets similaires. Dans la figure 1.2, nous présentons un schéma de la notion de comparabilité par rapport aux définitions attribuées à ces corpus. Intuitivement et selon la définition proposée par (Déjean et Gaussier, 2002),

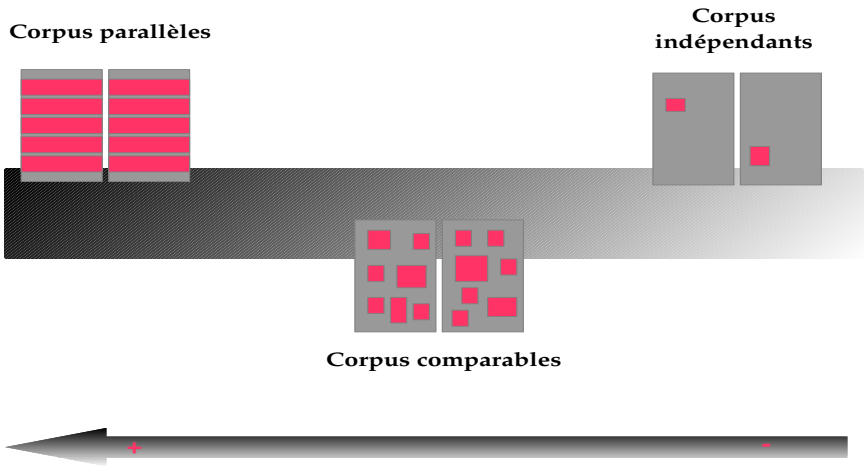


FIGURE 1.2: Comparabilité des corpus multilingues.

les corpus parallèles peuvent être considérés comme un cas particulier des corpus comparables. Il s'agit de corpus *parfaitement comparables* (Prochasson, 2009).

La catégorie des corpus indépendants comprend la grande majorité des textes sur Internet. Ils se composent de documents traitant des sujets similaires ou variés et utilisant un vocabulaire avec un usage différent au sein du même corpus ou entre les deux corpus source et cible. Les corpus comparables peuvent être vus comme toute collection de textes dans de différentes langues n'étant pas des traductions mutuelles (Bowker et Pearson, 2002).

La capacité des corpus comparables à améliorer la performance de différentes applications du TAL qui y ont recours serait fortement liée à leur *degré de comparabilité*. Cette notion ne constitue pas l'objet de notre étude mais nous présentons un bref aperçu de différentes approches étudiant ce phénomène. Plusieurs travaux ont mentionné le besoin d'une définition de la *comparabilité* et formulent leur compréhension de celle-ci. (Li et Gaussier, 2010) ont introduit une mesure qui permet d'indiquer le degré de comparabilité entre les deux parties source et cible d'un corpus comparable. Ils ont constaté que selon cette mesure, l'amélioration de la qualité du corpus comparable influence la qualité de l'extraction lexicale. (Su et Babych, 2012) quant à eux mesurent la comparabilité de textes à leur potentiel d'extraction de segments parallèles et d'amélioration de la performance des systèmes de traduction automatique. La conception de la comparabilité varierait donc d'une application à une autre

en Breast cancer is a type of cancer originating from breast tissue, most commonly from the inner lining of milk ducts or the lobules that supply the ducts with milk. Cancers originating from ducts are known as ductal carcinomas, while those originating from lobules are known as lobular carcinomas. Breast cancer occurs in humans and other mammals. While the overwhelming majority of human cases occur in women, male breast cancer can also occur

fr Le cancer du sein est une tumeur maligne de la glande mammaire. Autrement dit, c'est un cancer qui naît dans les unités cellulaires dont la fonction est de sécréter le lait, les unités ducto-lobulaires du sein, essentiellement chez la femme (le cancer du sein survient 200 fois moins souvent chez l'homme, qui possède lui aussi des seins, bien qu'atrophiés). Ce cancer est le plus fréquent chez la femme, avec 89 cas pour 100 000.

ro Cancerul mamar este o boală în care celulele maligne se formează în țesuturile sânului. Sânul este format din lobi și ducte. Fiecare sân are 15 până la 20 secțiuni numite lobi, lobi care au secțiuni mai mici numite lobuli. Cancerul de sân este uneori descoperit la femeile însărcinate sau care tocmai au născut. Majoritatea cazurilor de îmbolnăvire se petrec în jurul vârstei de 50 de ani. În ultimii 20 de ani, cazurile de îmbolnăvire s-au înmulțit foarte mult, iar răspândirea printre femeile tinere a început să ia amploare.[necesită citare] Mulți specialiști oncologi consideră că accidentul de la Cernobîl poate să fi influențat răspândirea acestei boli.necesită citare În cazul detectării timpurii, boala poate fi vindecabilă.

FIGURE 1.3: Premiers paragraphes des articles Wikipédia décrivant le mot *cancer du sein* en anglais, français et roumain

(Leturia *et al.*, 2009). Elle serait également influencée par le type de corpus qui peut être général ou de spécialité et par la source de collecte des documents.

Contrairement aux corpus parallèles, les corpus comparables sont largement disponibles et les textes qui les composent proviennent généralement de la même source mais sont écrits indépendamment dans chaque langue. Ils sont construits à partir de textes *originaux* plutôt que des textes traduits (corpus parallèles). Ceci permet de réduire le biais de traduction et d'éviter par conséquent l'effet de calque. L'exemple le plus significatif est celui des textes traitant d'une même actualité internationale et publiés par différentes agences de presse (Agence France-Presse (AFP), Reuters, etc). Comme le montre la figure 1.3, les articles de Wikipédia reliés par les liens interlingues constituent également une source de corpus fortement comparables. Nous remarquons que le contenu des premiers paragraphes des articles Wikipédia décrivant le mot *cancer du sein* en anglais, français et roumain est très comparable.

La particularité des corpus comparables est qu'ils ne respectent pas les contraintes imposées par les corpus parallèles. Selon (Fung, 1995), dans un corpus comparable :

1. Les mots ont *plusieurs sens* dans le même corpus.

2. De *multiples traductions* peuvent être associées à un mot.

3. Les traductions pourraient *ne pas exister* dans le document cible.

4. Les *positions* et *fréquences* des mots sont *incomparables*.

Comme souligné plus haut, nous considérons que les deux premières caractéristiques s'appliquent à la fois au corpus parallèles et comparables. La différence réside donc dans les deux derniers points. Ces caractéristiques montrent qu'en comparaison avec les corpus parallèles, la tâche d'extraction de lexiques bilingues à partir de ce type de corpus est *moins supervisée* du fait qu'elle requiert moins de données annotées. Une extraction moins supervisée permet donc (1) de compenser le manque de données parallèles pour des domaines génériques et spécialisés et, (2) de couvrir un large éventail de paires de langues.

1.3 Lexiques bilingues à partir de corpus parallèles

L'extraction de lexiques bilingues à partir de corpus parallèles est fortement liée à la traduction automatique statistique (Brown *et al.*, 1993). Conventionnellement, on se réfère à cette tâche en utilisant le terme *alignement*. La littérature concernant l'alignement en corpus parallèles est particulièrement riche et les travaux menés dans ce cadre comprennent des approches d'alignement d'unités linguistiques de différents niveaux de granularité (document, phrase, segment, mot, etc). Le niveau de granularité de ces unités est déterminé en fonction de l'application finale de celles-ci (traduction automatique, recherche d'information interlingue). Partant d'un corpus parallèle, des alignements de granularité de plus en plus fine peuvent être obtenus de manière séquentielle. Dans cette section nous décrivons les approches d'alignement de deux principaux niveaux de granularité : *l'alignement phrastique* (phrase) et *l'alignement sous-phrastique* (mot, segment, expression).

1.3.1 Alignement phrastique

L'alignement de phrases présente une utilité sans cesse croissante pour de nombreuses applications de TAL. Les textes parallèles ne sont pas toujours traduits phrase à phrase. Les phrases longues peuvent être divisées et les phrases courtes peuvent être fusionnées. Il existe même des langues comme le thaï dont l'écriture ne comprend pas d'indicateur de fin de phrase. Un alignement au niveau de la phrase s'avère donc utile.

De nombreuses approches d'alignement de phrases ont été proposées dans la littérature. Les premiers travaux remontent à (Brown *et al.*, 1991), où l'alignement se base sur le nombre de mots dans chaque phrase. L'idée qui sous-tend leur approche est que plus les tailles de deux phrases se rapprochent, plus elles sont susceptibles de s'aligner. Citons également les travaux de (Gale et Church, 1991) qui, au lieu de considérer le nombre de mots, ils utilisent le nombre de caractères dans chaque phrase comme caractéristique de base. D'autres approches se basent sur l'alignement de mots simples pour extraire des paires de phrases en relation de traduction (Kay et Röscheisen, 1993; Chen, 1993). Une combinaison de ces deux approches a été présentée dans (Moore, 2002). Outre l'alignement de mots et la longueur des phrases, des approches géométriques (Melamed, 1996) et de reconnaissance de formes (Melamed, 1999) ont été également utilisées dans ce cadre.

En alignement phrastique, des caractéristiques telles que la corrélation entre les tailles de phrases, et les contraintes lexicales sont souvent suffisantes pour que l'alignement de phrase soit relativement bon. Cependant, il est bien souvent préférable d'aligner des unités textuelles plus larges (paragraphe, section, chapitre) avant de procéder à l'alignement phrastique. Un alignement sous-phrastique facilitera également l'alignement phrastique (Kay et Röscheisen, 1993), et réciproquement. Ce type d'alignement permet de réduire l'espace de recherche des mots en relation de traduction, et les mots en relation de traduction permettent de repérer les phrases en correspondance.

1.3.2 Alignement sous-phrastique

L'alignement sous-phrastique est une tâche de niveau de granularité plus fin que celui de l'alignement phrastique. Ce type d'alignement constitue une composante importante pour la traduction automatique statistique (Brown *et al.*, 1993; Och

et Ney, 2003). Cette section est consacrée à la description de différentes approches s'intéressant à l'alignement de mots simples, de segments et d'expressions polylexicales.

1.3.2.1 Alignement de mots et de segments

Pour mettre en relation des mots qui sont des traductions mutuelles, la plupart des travaux se basent sur des *approches purement statistiques* et partent du constat que la corrélation entre les distributions de mots en relation de traduction est élevée. La distribution d'un mot est généralement définie par sa position et sa fréquence dans le corpus parallèle. Dans ce type d'approches, un modèle du corpus est construit à partir des données parallèles alignées au niveau phrastique. Ce modèle doit permettre une maximisation globale de la relation de traduction dans son ensemble. Concrètement, pour chaque couple d'énoncés source-cible d'un corpus parallèle, il s'agit de chercher à déterminer les meilleurs liens entre les mots de l'énoncé source et ceux de la cible. Les plus connus de ces modèles sont ceux remontant aux travaux de (Brown *et al.*, 1993). Ces modèles sont au nombre de cinq (IBM1,..., IBM5), de complexité croissante, chacun introduisant des paramètres permettant d'affiner les résultats du précédent, tels la position des mots, leur déplacement au cours du processus de traduction ("distorsion") ou encore le nombre de mots correspondant au mot source ("fertilité"). Une extension de ces modèles a été introduite dans (Vogel *et al.*, 1996). Dans cette étude, une intégration d'un modèle d'alignement à base de Modèles de Markov Cachés (Hidden Markov Model, HMM) a été mise en place. Ce modèle a été utilisé dans GIZA++ (Och et Ney, 2003), le système d'alignement le plus utilisé. Un exemple d'alignement établi par cet outil est présenté dans la figure 1.4. Nous constatons que les alignements obtenus décrivent aussi bien des correspondances entre mots (*president* $\longleftrightarrow$ *président*) qu'entre *blocs de mots* (*cut off* $\longleftrightarrow$ *retire la parole à*). Ces blocs sont constitués par des groupes de mots ou n-grammes contigus nommés *segments* (*phrases* en anglais).

Bien que GIZA++ soit utilisé dans la plupart des systèmes de traduction statistique, sa performance n'est prouvée que pour l'alignement des phrases de petites tailles (de l'ordre de 50 mots) et pour des paires de mots apparaissant au moins 50 fois dans le corpus parallèle (Koehn et Knight, 2001). Pour produire des résultats satisfaisants, ces modèles purement statistiques dits *génératifs* nécessitent d'importantes

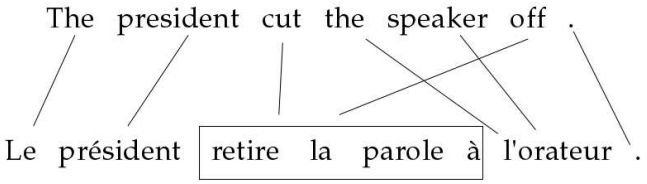


FIGURE 1.4: Alignement entre une phrase en anglais et sa traduction en français sous forme de liens. Le texte encadré correspond à un segment.

quantités de données parallèles pour l'apprentissage. La complexité mathématique et l'absence d'encodage explicite des connaissances linguistiques font de ces modèles des boites noires et rendent difficile l'analyse des résultats produits, notamment ceux liés à l'alignement. Il est impossible de savoir pourquoi certains alignement ont échoué tout comme il est parfois difficile de savoir pourquoi d'autres ont réussi.

Pour palier ces insuffisances, des modèles statistiques dit *discriminants* sont utilisés. (Allauzen et Wisniewski, 2009) proposent deux modèles discriminants d'alignement mot à mot. Le premier modèle formalise la tâche d'alignement comme une tâche de classification multiclasse et traite celle-ci avec un classifieur à maximum d'entropie. Ce modèle permet d'introduire aisément des caractéristiques arbitraires tout en présentant une complexité faible aussi bien en apprentissage qu'en inférence. Il prédit les alignements indépendamment les uns des autres, bien qu'il soit intuitivement plus pertinent de le faire conjointement afin de pouvoir choisir l'alignement d'un mot en tenant compte des alignements de ses voisins. C'est pour cette raison qu'ils ont considéré un modèle fondé sur les champs conditionnels aléatoires (CRFs). Ces modèles formalisent l'alignement de mots comme une tâche d'étiquetage de séquence dans laquelle chaque mot de la phrase source est associé à l'index d'un mot de la phrase cible. Toutefois, l'utilisation de ces modèles discriminants pour l'alignement se heurte à une difficulté majeure : leur apprentissage nécessite des corpus alignés mot à mot alors que la quasi-totalité des corpus disponibles aujourd'hui sont alignés phrase à phrase et que les rares corpus alignés mot à mot ne comportent généralement que peu d'exemples. (DeNero et Klein, 2010) présentent également un modèle discriminant dédié à l'alignement de segments. Ce modèle prédit directement quel ensemble de segments doivent être extraits à partir d'une phrase. (Haghighi *et al.*, 2009) ex-

ploitent les contraintes des *grammaires de transduction inversible* présentées dans (Wu, 1997). Ces grammaires fournissent des contraintes structurelles cohérentes sur la relation entre une phrase et sa traduction. Les parties source et cible d'une paire de phrases alignées sont analysées simultanément selon un arbre de dérivation binaire. La particularité de cette technique est qu'elle permet d'inverser les constituants d'une phrase d'une langue à l'autre à n'importe quel niveau de l'arbre. Les approches purement statistiques tentent d'analyser les textes bilingues en se basant sur des modèles probabilistes. Les phrases ne sont pas considérées comme des entités structurées et on n'a pas recours à des lexiques. Les résultats obtenus par ces approches sont suffisamment utiles pour des applications réelles telles que la traduction automatique, la recherche d'information interlingue, etc. Un reproche majeur adressé à ces approches est qu'elles ne fonctionnent que quand les corpus traités sont larges.

Même si le domaine a connu une grande activité au cours de ces dernières années, peu d'améliorations ont été adoptées au final par la communauté. Certaines recherches (Xu et Chen, 2011) tentent d'améliorer la performance de l'alignement par des connaissances obtenues à partir d'alignements fait par des humains. Ils montrent que, par rapport à GIZA++, les gains réalisés par des alignements humains sont inférieurs à un point Bleu et que plus la taille de corpus d'apprentissage est grande, moins cette amélioration l'est.

Citons également les travaux de (Lardilleux, 2010), qui présente l'outil d'alignement *anymalign*. Cet outil permet d'aligner les mots simples ou segments de faible fréquence ou rares, qui dans la littérature, sont habituellement rejetés et jugés peu fiables. La méthode consiste à rendre les mots fréquents rares dans des sous corpus constitués par échantillonnage et d'effectuer la tâche d'alignement : il s'agit de se placer dans un espace vectoriel dont le nombre de dimensions est constitué par le nombre d'énoncés (source, cible) du corpus parallèle. À chaque mot est associé un vecteur dont la i-ième composante est le nombre d'occurrences de ce mot dans la i-ème phrase, puis pour chaque couple de mots, une similarité entre les vecteurs correspondants est mise en place en se basant sur le cosinus. Comme le font remarquer (Och et Ney, 2003), d'une part, le choix d'une mesure de corrélation pour l'alignement bilingue est généralement assez arbitraire, car ces mesures produisent des résultats de qualité comparable, par exemple l'indice de Jaccard et le cosinus. D'autre part, ils considèrent que ces méthodes dites associatives qu'ils qualifient d'heuristiques, ne

peuvent pas rivaliser avec les méthodes estimatives qui reposent sur l'estimation de paramètres dans le cadre d'une théorie bien fondée.

Peu de travaux exploitent des connaissances linguistiques pour effectuer la tâche d'alignement sous phrastique. Citons par exemple les travaux de (Zribi, 1995) qui dans le cadre de sa thèse propose une solution formelle unique pour l'alignement de phrases, de paragraphes et de mots simples faisant intervenir un lexique bilingue. Il considère que ces problèmes sont analogues. Le principe de sa méthode consiste à comparer les unités source et cible et à retenir les couples pour lesquels la comparaison est concluante. Les approches à bases de connaissances linguistiques font quant à elles une analyse plus ou moins fine des textes traités et l'alignement se base sur des lexiques bilingues de transfert. En plus d'un lexique bilingue, (Semmar *et al.*, 2010) font appel aux entités nommées et aux cognats pour l'alignement de mots simples. Ces unités sont tout d'abord mises en correspondances. Ensuite, pour les unités non alignées encadrées par des mots alignés, le système recourt à la catégorie grammaticale des mots sources et cibles. Partant de l'idée que plus les catégories grammaticales et les relations syntaxiques sont proches dans les deux langues, plus les alignements qui en résultent ont la chance d'être bons (Zribi, 1995), (Ozdowska, 2006) présente une approche syntaxique pour l'alignement de mots simples. Dans cette approche, la technique de programmation logique inductive est utilisée pour apprendre des règles de propagation syntaxique. Le principe requiert un corpus parallèle aligné au niveau de la phrase, que ce corpus soit analysé syntaxiquement et que également des paires de mots amorces extraites d'un lexique bilingue soient présentes dans les paires de phrases. Ces approches sont plus précises et peuvent être appliquées à de très petits corpus. Tout le problème réside dans la couverture des dictionnaires bilingues.

L'alignement de mots simples constitue une tâche bien maîtrisée et la taille des lexiques bilingues qui en résultent est assez large. Or un des points faibles des lexiques est souvent le manque de couverture pour des unités complexes comme les collocations, mots composés et expressions idiomatiques (Sagot *et al.*, 2005). Dans la section suivante nous décrivons différentes approches s'intéressant à l'extraction de lexiques bilingues de ce type d'unités nommées *expressions polylexicales*.

1.3.2.2 Vers l'alignement d'expressions polylexicales

Une expression polylexicale (EPL, en anglais *MultiWord Expression (MWE)*) peut être définie comme une combinaison de mots pour laquelle les propriétés syntaxiques ou sémantiques de l'expression entière ne peuvent pas être obtenues à partir de ses parties (Sag *et al.*, 2002). Les EPL regroupent les expressions figées et semi-figées (ex. *cordon bleu*), les entités nommées (ex. *New York*), les verbes à particule (ex. *grow up*), les constructions à verbe support (ex. *faire face à*), etc.(Sag *et al.*, 2002; Constant *et al.*, 2011). Elles sont fréquemment employées dans les textes écrits car elles constituent une part significative du lexique d'une langue. (Jackendoff, 1997) estime que la fréquence de leur utilisation est équivalente à celle des mots simples. Bien qu'elles soient facilement employées et reconnues par les humains, leur identification pose un problème majeur pour diverses applications du traitement automatique des langues. Dans cette étude, nous nous intéressons à l'extraction de ce type d'unités.

Au cours des dernières années, de nombreux travaux de recherche ont été menés sur la tâche d'extraction d'EPL bilingues à partir de corpus parallèles. La traduction des EPL d'une langue à une autre exige que ce type d'unité soit reconnue. C'est pour cette raison que la plupart des travaux identifient tout d'abord les EPL dans chaque partie du corpus parallèle, et puis se basent sur différentes techniques d'alignements pour les mettre en correspondance.

Identification monolingue d'EPL :
Les techniques d'identification d'EPL tournent autour de trois approches : (1) des approches symboliques (2) des approches statistiques et (3) des approches hybrides combinant (1) et (2). Les approches symboliques se basent sur des patrons morpho-syntaxiques définis manuellement (N+Prep+N, N+N, Adj+N, ...). Ces approches font appel à des étiqueteurs morphosyntaxiques pour prendre en considération certaines catégories de mots et à des outils de lemmatisation pour reconnaître toutes les formes fléchies d'une unité lexicale. Le travail de (Kupiec, 1993) peut être considéré comme l'un des premiers travaux sur l'extraction d'EPL à partir de corpus parallèles. Ce travail était centré sur des groupes nominaux comme *"late spring" [la fin du printemps]*, identifiés sur la base de leur catégorie en utilisant un reconnaisseur à états finis. Plusieurs travaux se sont basés sur cette technique dont ceux de (Okita *et al.*, 2010).

(Dagan et Church, 1994) présentent *Termight*, un outil de création de lexiques bilingues de termes techniques. L'identification de ces termes se fait sur la base d'un étiqueteur morpho-syntaxique. Ensuite, la liste de candidats trouvée est filtrée manuellement. L'application de filtres à base de catégories grammaticales permet une réduction importante du bruit dans les sorties par l'exclusion de candidats constitués de mots vides. Malgré leur simplicité, les approches symboliques restent difficile à appliquer lorsque les données ne sont pas étiquetées morpho-syntaxiquement. Une autre limite de cette approche est que la définition de patrons d'extraction d'EPL est dépendante de la langue.

Les approches statistiques d'identification d'EPL se concentrent sur leur comportement collocationnel. Ce comportement est quantifié sur la base de mesures d'association lexicales. Le résultat d'extraction est représenté par la liste de paires candidates triée par ordre décroissant du score d'association obtenu. Les paires situées en haut de la liste sont les plus susceptibles de constituer de vraies EPL et de présenter un intérêt lexicographique. (Smadja *et al.*, 1996) proposent l'outil Xtract pour l'extraction de collocations à partir de textes à travers une combinaison de n-grammes et d'une mesure d'information mutuelle. (Pecina, 2008) compare 55 mesures d'association pour le classement d'EPL candidates. Cette étude montre que la combinaison de différentes mesures d'association par une technique de classification classique (réseaux de neurones) contribue à de meilleurs résultats que lorsque ces mesures sont employées individuellement. Une limite pratique de ces approches est l'importante combinatoire générée, en particulier si l'on cherche à extraire des EPL de plus de deux mots. En plus, dans ce type d'approche, la définition d'un seuil à partir duquel un segment extrait peut être considéré comme une EPL ou pas est nécessaire.

Il est devenu clair que la simple mesure d'association ne suffit pas à identifier les EPL et qu'il conviendrait de considérer en plus leurs propriétés linguistiques (Piao *et al.*, 2005). Les approches hybrides combinent les informations statistiques avec des informations linguistiques morphologiques, syntaxiques ou encore sémantiques pour l'identification des EPL. Par exemple, (Cook *et al.*, 2007) utilisent des connaissances à priori sur la structure syntaxique d'une expression idiomatique en vue de déterminer si une instance de l'expression est utilisée littéralement ou d'une façon idiomatique. Ils présument que, dans la plupart des cas, les usages idiomatiques d'une expression ont tendance à se produire dans un petit nombre de formes de cet idiome. (Seretan et Wehrli, 2007; Daille, 2001; Moirón et Tiedemann, 2006) utilisent des patrons mor-

phosyntaxiques pour identifier des EPL candidats dans un texte et les pondèrent par leur valeur d'association. Dans ce cadre, la valeur d'association permet de prédire si une expression candidate est une EPL ou pas.

Des propriétés sémantiques des EPL ont été récemment utilisées pour distinguer les EPL compositionnelles de celles non compositionnelles. En effet, (Katz et Giesbrecht, 2006; Baldwin *et al.*, 2003) utilisent l'analyse sémantique latente (LSA) et montrent que contrairement aux EPL non compositionnelles, les EPL compositionnelles apparaissent généralement dans des contextes similaires à leur constituants. La limite principale de ce type d'approche, faisant intervenir l'aspect sémantique, est que la distinction entre une utilisation idiomatique ou pas d'une EPL s'appuie sur des expressions idiomatiques connues et que cette information est généralement absente. En outre, cette approche ne fonctionne que lorsque l'expression en question est extrêmement fréquente.

Alignement d'EPL :

Pour identifier des correspondances entre expressions dans différentes langues, la plupart des travaux font appel à des outils d'alignement de mots simples pour guider l'alignement d'EPL (Dagan et Church, 1994; Moirón et Tiedemann, 2006; Ren *et al.*, 2009). D'autres se basent sur des algorithmes d'apprentissage statistique comme par exemple l'algorithme itératif de ré-estimation *Expectation Maximization* (Kupiec, 1993; Okita *et al.*, 2010). Une hypothèse largement suivie pour acquérir des EPL bilingues est qu'une expression dans une langue source garde la même structure syntaxique que son équivalente dans une langue cible donnée (Seretan et Wehrli, 2007; Tufis et Ion, 2007). Or, les EPL ne se traduisent pas forcément par des expressions ayant la même catégorie grammaticale (i.e « *insulaire en développement* » et « *small island developing* ») ou la même longueur[1] (i.e « *en ce qui concerne* » et « *as regards* »). Dans (Semmar *et al.*, 2010), l'alignement de mots composés consiste à établir des correspondances par des règles de formulation entre les mots composés de la phrase source et ceux de la phrase cible.

1. La longueur d'une EPL est calculée en nombre de mots.

1.4 Lexiques bilingues à partir de corpus comparables

Bien que les corpus parallèles bilingues se soient multipliés au cours des dernières années, ils sont encore relativement peu nombreux par rapport à la grande quantité de textes monolingues. Plus important encore, il est difficile de disposer de corpus parallèles spécialisés dans un domaine particulier et s'ils existent, ces ressources doivent avoir été construites par des traducteurs humains. Par conséquent, les lexiques bilingues construits à partir de ces corpus sont le résultat d'une rétro-ingénierie du lexique utilisé par les traducteurs. Cette rétro-ingénierie ajoute donc en plus un biais de traduction causé par les phénomènes de calques (c'est à dire une traduction mot à mot) et d'autres traductions influencées par la langue source. En outre, l'acquisition et le traitement de ces corpus sont des tâches coûteuses en temps. Pour pallier ces insuffisances, les recherches récentes se sont donc penchées sur l'exploitation de ressources diverses et plus disponibles : *les corpus comparables*.

L'alignement lexical à partir de corpus comparables est toutefois une opération délicate : il n'est plus possible de s'appuyer sur la distribution des mots dans le document. Les approches proposées cherchent plutôt à prendre en compte le contexte de chaque terme à aligner, c'est-à-dire la façon dont ils sont employés et les mots avec lesquels ils cooccurrent dans le texte. Cette section est consacrée à la présentation des différentes approches proposées. Dans la section 1.4.1, nous décrivons les premières approches s'intéressant à l'utilisation de corpus comparables pour l'extraction de lexiques bilingues. Nous présentons ensuite dans les sections 1.4.2 et 1.4.3 l'approche standard et les différentes améliorations qui lui sont apportées. Enfin, nous passons en revue des approches connexes dans la section 1.4.4.

1.4.1 Premières approches

La première approche utilisant les corpus comparables pour construire des lexiques bilingues a été proposée par (Rapp, 1995). Dans cette étude, l'auteur montre que l'extraction lexicale à partir de textes non parallèles, voire indépendants devrait être possible en exploitant des *motifs de cooccurrence*. Il part de l'hypothèse que si deux mots cooccurrent fréquemment dans un texte dans une langue source, leurs traduc-

(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		*	*			
grün	2	*				*	
Himmel	3	*					
Lehrer	4						*
Pflonze	5		*				
schule	6				*		

(c)

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

FIGURE 1.5: Matrices de cooccurrences d'un ensemble de mots anglais et allemand. Lorsque l'ordre des mots anglais et allemand correspond, les motifs de cooccurrence de leurs matrices désignés par (*) sont identiques.

tions devraient également cooccurrer fréquemment dans le texte de la langue cible. Ainsi, les motifs de cooccurrence sont formés par la force d'association (dans quelle mesure deux termes sont reliés l'un à l'autre) entre chaque terme source et cible et ses voisins. Ainsi, les comparaisons des différents motifs de la langue cible et de la langue source devraient dévoiler de potentielles relations de traduction. La figure 1.5 illustre un exemple présenté dans (Rapp, 1995) qui décrit les matrices d'association et les alignements de quelques mots en anglais et allemand. Dans ces matrices, une association entre deux mots est représentée par une étoile (*). L'algorithme d'alignement consiste à réorganiser *aléatoirement* les matrices de façon à trouver un motif similaire entre la langue source et cible (tableau (c)). L'efficacité de cette approche n'a été démontré que sur des corpus volumineux. Néanmoins, cette approche est très coûteuse en temps de calcul puisqu'elle nécessite un nombre exponentiel de réorganisations et de comparaisons de paires de matrices.

Une approche différente a été présentée dans (Fung, 1995). Cette approche pose l'hypothèse que les mots ayant des contextes productifs dans une langue se traduisent par des mots dont le contexte est productif dans la langue cible et qu'inversement, les mots avec des contextes rigides se traduisent par des mots ayant des contextes

rigides. La productivité du contexte d'un mot pour un domaine donné est calculée sur la base d'une mesure *d'hétérogénéité de contexte* indiquant pour un terme donné le nombre de mots différents le précédant immédiatement (hétérogénéité à gauche) et le suivant immédiatement (hétérogénéité à droite). Sur la base de cette mesure, les empreintes caractérisant l'usage d'un terme dans une langue source et cible sont comparées pour identifier des paires en relation de traduction.

Ces deux premières approches s'intéressant à la compilation de lexiques bilingues à partir de corpus comparables (non parallèles) ne s'appuient pas sur la fréquence d'un mot dans les corpus. L'alignement se base sur la comparaison des environnements lexicaux. Notons également que dans ces approches, aucune connaissance linguistique préalable n'a été utilisée. La limite principale de ces approches est qu'elle requièrent des corpus comparables très volumineux pour que la caractérisation des termes à traduire soit suffisamment discriminante. Suite à ces premières études, la tâche d'extraction de lexiques bilingues à partir de corpus comparables a attiré l'attention de plusieurs chercheurs et différentes approches ont été mises en place que nous présentons dans les sections suivantes.

1.4.2 Approche standard

Les principaux travaux s'intéressant à la création de lexiques bilingues à partir de corpus comparables pourraient être vus comme une extension de l'hypothèse distributionnelle de Harris (Harris, 1954) et reposent sur sur la simple observation que si dans une langue source deux mots cooccurrent plus souvent que par hasard, alors dans un texte de langue cible, leurs traductions doivent également cooccurrer plus souvent (Rapp, 1995). Sur la base de ces hypothèses, *l'approche standard* ou dite encore *approche par traduction directe* a vu le jour. Cette approche se base sur la caractérisation et la comparaison d'environnements lexicaux des termes sources et cibles, représentés par des *vecteurs de contexte*. Ces vecteurs stockent un ensemble d'unités lexicales représentatif de leur voisinage. Dans la pratique, afin de pouvoir comparer les vecteurs de contexte de langues différentes, le passage d'une langue à une autre est nécessaire et s'effectue généralement par l'intermédiaire d'un dictionnaire bilingue amorce. Un aperçu général de cette approche est illustré dans la figure 1.6. Cette approche se compose des trois étapes suivantes :

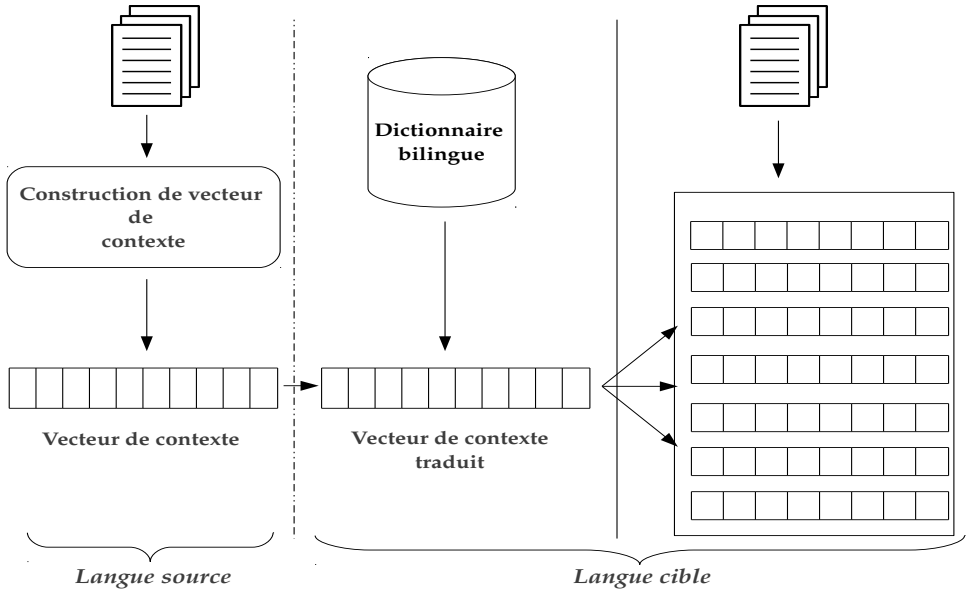


FIGURE 1.6: Aperçu général de l'approche standard d'extraction de lexiques bilingues à partir de corpus comparables.

1. Construire un vecteur de contexte pour chaque terme à traduire et tous les candidats à la traduction de la langue cible. Cela fournit une représentation distributionnelle de chacun de ces termes.
2. Utiliser un dictionnaire bilingue amorce pour traduire le vecteur de contexte du terme à traduire.
3. Comparer le vecteur de contexte de l'unité à traduire avec tous les vecteurs de contextes cibles à l'aide d'une mesure de similarité. Nous obtenons une liste ordonnée de traductions candidates pour le terme à traduire selon leur similarité distributionnelle.

1.4.2.1 Constitution des vecteurs de contexte

Les vecteurs de contexte servent à représenter les termes à traduire. Ils sont extraits en repérant les mots qui apparaissent autour du terme à traduire. Idéalement, ces termes doivent entretenir des relations de dépendance syntaxique avec le terme à traduire. Or, comme dans la plupart des cas cette analyse ne fournit que deux termes (opérateur-opérande), la caractérisation de ce contexte risque de ne pas être suffisamment riche. En outre et pour éviter les erreurs de ce type d'analyse, les recherches actuels définissent le contexte par les mots qui apparaissent simplement autour du

terme à traduire dans une fenêtre contextuelle de n mots. Ces mots sont généralement constitués de mots pleins (noms, verbes, adjectifs, adverbes).

Habituellement, des mesures d'associations comme l'information mutuelle (Morin et Daille, 2006), le rapport de vraisemblance (Morin et Prochasson, 2011) ou encore le rapport des chances (odds-Ratio) (Laroche et Langlais, 2010) sont utilisées pour définir les entrées du vecteur de contexte. Les mesures d'association comme le rapport des chances ou l'information mutuelle évaluent la dépendance statistique de deux grandeurs mesurées. Plus ces grandeurs sont dépendantes, plus leur valeur d'association sera importante. Le rapport de vraisemblance consiste par contre à calculer le ratio des vraisemblances de deux configurations correspondants aux hypothèses à confronter : que les deux mots cooccurrent et qu'ils apparaissent de manière indépendante. En extraction lexicale, la mesure d'association indique la degré de corrélation entre les un mot et un mot avec lequel il cooccurre. Elle est aussi utilisée pour indiquer la force d'association entre un terme à traduire et les éléments de son vecteur de contexte. Les vecteurs de contexte sont ainsi construit pour chaque terme à traduire et tous les candidats à la traduction du corpus de la langue cible.

1.4.2.2 Transfert des vecteurs de contexte

Afin de rendre possible la comparaison des vecteurs sources et cibles, et contrairement aux méthodes introduites par (Rapp, 1995) et (Fung, 1995), les vecteurs des termes sources sont traduits par le biais d'un dictionnaire bilingue amorce. Ce dictionnaire sert de pont entre la langue source et cible. Il constitue l'élément clé de l'approche standard. S'il propose plusieurs traductions pour un mot, l'ensemble de traductions proposées sont ajoutées. Par contre, les mots qui n'y figurent pas sont simplement ignorés. Les résultats d'extraction sont donc influencés par la couverture du dictionnaire bilingue.

1.4.2.3 Comparaison des vecteurs sources et cibles

Une fois traduits dans la langue cible, les vecteurs des termes à traduire sont comparés à l'ensemble des vecteurs de contexte des candidats à la traduction à l'aide d'une mesure de similarité vectorielle. La plus populaire est le cosinus, mais de nombreux

auteurs ont étudié des métriques alternatives comme l'indice de Jaccard pondérée ou encore la distance de Manhattan. En fonction des valeurs de similarité, nous obtenons une liste ordonnée de traductions candidates pour chaque terme à traduire.

1.4.2.4 Résultats de l'approche standard

Les recherches exploitant l'approche standard se sont intéressées à la construction et à l'extension de lexiques bilingues par des mots du domaine général (Rapp, 1995), de termes issue d'un domaine de spécialité (Chiao et Zweigenbaum, 2002; Déjean *et al.*, 2002; Prochasson *et al.*, 2009) ou encore de termes complexes (Morin et Daille, 2006; Laroche et Langlais, 2010). Comme il a été mentionné précédemment, le résultat d'alignement obtenu par cette approche est une liste ordonnée de candidats à la traduction pour chaque terme à traduire classée en fonction des valeurs de similarité entre leur vecteurs de contexte respectifs. Les résultats de cette approche sont évalués en comptant le nombre de candidats corrects trouvés dans les N premiers candidats renvoyés (succès au rang N ou $succès_N$). Cette méthode d'évaluation a été originellement utilisée dans la une conférence pour l'évaluation des systèmes de recherche d'information *TrecEval*.

La qualité des traductions obtenues par l'approche standard dépend du domaine auquel on s'intéresse, de la taille du corpus, de la taille de la fenêtre contextuelle et des mesures d'association et de similarité adoptées. Par exemple, (Rapp, 1999a) obtient une précision de 72 % au $succès_1$ pour un très large corpus comparable composé d'articles de journaux anglais-allemand. Dans le domaine médical, (Chiao et Zweigenbaum, 2002) obtiennent une précision de 20 % pour le $succès_1$ avec un corpus français-anglais d'environ 600 000 mots. Dans (Morin *et al.*, 2008), les auteurs utilisent un corpus français-japonais lié à la thématique du diabète et de l'alimentation. Pour les $succès_{10}$, ils portent la précision à 49 %. En pratique, et comme il a été noté dans (Prochasson, 2009), il est difficile de comparer les résultats de différents travaux en extraction de lexiques bilingues à partir de corpus comparables, en raison de différences entre les corpus, les domaines d'étude ou encore les ressources linguistiques utilisées.

1.4.3 Améliorations de l'approche standard

La couverture du dictionnaire bilingue assurant le transfert des vecteurs de contexte en langue cible demeure le noyau de l'approche standard. Si trop peu de mots sont traduits, la comparaison de vecteurs traduits et de vecteurs cibles ne donnera pas une bonne représentation de leur similarité distributionnelle puisque réalisée sur un échantillon trop faible de vocabulaire. La valeur des éléments non traduits des vecteurs de contextes disparaîtra lorsque ce vecteur sera transféré en langue cible. Pour limiter cet effet, des techniques visant à améliorer les résultats de l'approche standard ont vu le jour par *l'adjonction de ressources linguistiques supplémentaires*. Ainsi, en associant un dictionnaire de langue générale à un dictionnaire spécialisé, dans le but d'aligner des termes simples, (Chiao et Zweigenbaum, 2003) obtiennent une amélioration significative des performances d'alignement en faisant passer la précision de 61 à 94 % pour les $succès_{20}$. Dans (Morin et Prochasson, 2011), les auteurs combinent un lexique général avec un lexique de spécialité. Ce lexique est extrait à partir de segments parallèles identifiés dans le corpus comparable. Un gain en précision de +9 points pour le $succès_{20}$ a été rapporté. Une approche similaire a été proposée par (Vulić et Moens, 2012) où les traductions *sures* sont d'abord extraites pour en construire un lexique bilingue servant après d'amorce pour transférer les vecteurs de contexte du reste des candidats du corpus comparable. (Déjean *et al.*, 2002) s'appuient sur des propriétés hiérarchiques d'un thésaurus spécialisé pour améliorer les rangs des traductions candidates. Avec cette ressource supplémentaire, ils ont passé la précision de 57 à 63 % pour les $succès_{20}$. (Li et Gaussier, 2010) proposent une approche qui tente d'améliorer la précision de la méthode standard en introduisant une mesure de comparabilité du corpus comparable considéré et en améliorant le corpus selon cette mesure avant d'extraire le lexique bilingue.

Récemment, des recherches fondées sur l'hypothèse que plus les vecteurs de contextes sont représentatifs, meilleure est la mise en correspondance bilingue ont été menées. (Prochasson *et al.*, 2009) introduisent la notion de *points d'ancrage* constitués de translittérations et de mots composés scientifiques. L'hypothèse proposée repose sur le fait de donner plus d'importance à ces unités lorsque l'on compare les vecteurs de contexte. Pour un corpus de textes issus du domaine médical anglais/japonaise, une amélioration de la précision de 18 % en utilisant les translittérations et les mots composés scientifiques pour les $succès_{10}$ mais demeure nulle pour le $succès_1$. (Rubino et Linarès, 2011) combinent la représentation contextuelle avec une représentation

thématique et graphique (translittérations et cognats) de termes médicaux. Ils émettent l'hypothèse qu'un terme et sa traduction partagent des similarités d'un point de vue thématique et qu'en domaine de spécialité beaucoup de termes sont portés d'une langue à une autre sans subir de modification. Leur méthode atteint une précision de 83 % et un faible rappel de 26 % pour les traductions au $succès_1$. (Hazem et Morin, 2012a) proposent deux critères de filtrage du dictionnaire bilingue dans le but de ne garder que les mots qui donnent la meilleure représentation du vecteur de contexte dans la langue cible. Le premier critère se base sur les catégories grammaticales des mots du contexte mais aucune amélioration n'a été démontrée. Le deuxième critère est basé sur une mesure de pertinence d'un mot pour un domaine donné. Contrairement au premier critère, celui ci rapporte une petite amélioration (4 % en précision) par rapport à la méthode standard.

Les ambiguïtés révélées par le dictionnaire bilingue amorce ont été prises en compte plus récemment. (Gaussier *et al.*, 2004) utilisent une vue géométrique et décomposent le vecteur d'un mot en fonction de ses sens par l'utilisation de plusieurs méthodes comme l'analyse canonique de corrélation et l'analyse sémantique latente. Les meilleurs résultats sont obtenus par l'utilisation d'une approche mixte avec une amélioration de la précision moyenne (Mean Average Precision, MAP) de 10 % au $succès_{500}$. (Apidianaki *et al.*, 2013) proposent une approche basée sur une méthode de désambiguïsation lexicale trans-lingue. Dans leur approche, les sens candidats de chaque élément du vecteur de contexte correspondent aux clusters de sens de ses traductions qui sont trouvées dans un corpus parallèle. La désambiguïsation des clusters de traduction se fait sur la base des éléments du même vecteur de contexte. La désambiguïsation permet ainsi de ne garder que le cluster des traductions les plus pertinentes pour la description du terme à traduire. La limite principale de cette méthode est qu'elle requiert un corpus parallèle pour construire le lexique bilingue. Or, comme il a été noté auparavant, ce type de ressource est très rare surtout pour des domaines de spécialité.

1.4.4 Approches connexes

Dans cette section, nous présentons les différentes approches connexes à l'approche standard. La première approche divergeant de cette dernière est l'*approche par similarité interlingue* (Déjean et Gaussier, 2002). L'objectif principal de cette approche

est d'être le moins dépendant que possible du dictionnaire bilingue amorce. Elle se base sur l'idée que si deux mots ont des distributions similaires alors ils sont reliés sémantiquement et repose sur l'identification d'affinités du second ordre : «*Les affinités du second ordre dévoilent quels mots partagent les mêmes environnements. Les mots partageant des affinités du second ordre n'ont pas besoin d'apparaître ensemble, mais leurs environnements sont semblables.*»

Le principe de cette approche consiste à identifier les vecteurs de contexte du dictionnaire bilingue qui sont similaires au vecteur de contexte du mot à traduire. Le dictionnaire va permettre de traduire les vecteurs de contexte dans leur globalité et non élément par élément. De cette manière, les vecteurs de contexte transférés perdent moins de leur potentiel de discrimination en langue cible. Le dictionnaire bilingue est alors mieux exploité puisque des traductions candidates peuvent être proposées pour un mot à traduire même si aucun élément de son vecteur de contexte ne peut être traduit. Sur la base de cette méthode, (Déjean et Gaussier, 2002) obtiennent pour des termes simples français-allemand une précision, pour les $succès_{10}$ et $succès_{20}$, de 43 % et 51 % pour un corpus médical de 100 000 mots (respectivement 44 % et 57 % avec l'approche standard).

Nous distinguons également l'*approche syntaxique*, qui se base sur l'observation qu'un mot et sa traduction ont tendance à partager les mêmes relations de dépendance syntaxique. Dans cette approche, au lieu de représenter un terme à traduire par les mots qui les entourent, on considère les relations de dépendance syntaxique qu'entretient ce terme dans le corpus (Yu et Tsujii, 2009). Dans (Garera *et al.*, 2009), une amélioration de la précision de 16 % a été réalisée pour la paire de langues anglais-espagnol. Dans la même veine, (Gamallo, 2007) introduit une méthode dans laquelle les contextes sont représentés par des paires de patrons lexico-syntaxiques extraits d'un corpus parallèle. Pour le domaine général, leur méthode atteint une précision 74 % et surpasse les résultats obtenues par l'approche standard (32 %) (Gamallo, 2008). L'avantage principal de ces méthodes est qu'elles ne font appel à aucun dictionnaire bilingue, qui dans l'approche standard assure le transfert de mots du contexte. Dernièrement, (Hazem et Morin, 2013) combinent la représentation par fenêtre ou sac de mots et celle par relations de dépendances syntaxiques de deux manières. La première manière est une combinaison *a posteriori* des contextes, qui combine les scores renvoyés par l'approche standard selon les deux représentations. La seconde manière consiste en une combinaison *a priori* qui utilise les deux informations contex-

tuelles a priori dans un même vecteur pour ensuite appliquer l'approche standard une seule fois sur l'ensemble du corpus. Ces deux méthodes de combinaisons contextuelles ont montré des résultats supérieurs à l'utilisation de chaque représentation séparément, pour la plupart des configurations. Bien que la représentation syntaxique des termes à traduire soit plus intéressante d'un point de vue sémantique, leur efficacité est très sensible à la taille des corpus. Elle atteint ses limites lorsqu'il s'agit de traiter des corpus de petite taille.

Une *approche à base de graphe* a été récemment présentée dans (Dorow *et al.*, 2009). Dans cette étude, chaque partie (source et cible) du corpus comparable est représentée par un graphe dont les nœuds sont les différents mots du corpus et les arcs représentent des relations grammaticales, comme la coordination, modification et sous-catégorisation entre les mots. Cette approche est basée sur l'algorithme récursif SimRank et part de l'intuition que deux nœuds sont similaires s'ils établissent des relations grammaticales similaires avec des nœuds voisins similaires.

Un certain nombre de tentatives d'extraction des lexiques bilingues à partir de corpus comparable sans utilisation de dictionnaires bilingues amorces ont été mises en place. (Diab et Finch, 2000) utilisent une approche d'amorçage ne nécessitant que peu de traductions amorces. Leur méthode de recherche de nouvelles traductions est basée sur l'algorithme d'optimisation de descente de gradient. Ils modifient de manière itérative l'alignement d'un terme à traduire jusqu'à ce qu'ils atteignent un minimum local de la somme des différences au carré entre la mesure d'association de toutes les paires de mots dans une langue et celle des paires des mots traduits. En fonction du nombre des termes à traduire, une précision variant de 92 à 98 % est obtenue. La limitation principale que l'on peut associer à cette méthode est liée à sa complexité algorithmique. (Haghighi *et al.*, 2008) utilisent un modèle génératif basé sur l'analyse de corrélation canonique pour induire de nouvelles paires de traduction. L'apprentissage de ce modèle se base sur un bon nombre de caractéristiques à savoir les cooccurrences des mots et les chaînes orthographiques. Pour un ensemble varié de corpus et paires de langues, ils montrent que l'on peut extraire des lexiques bilingues de haute précision même en l'absence de dictionnaire amorce.

(Hazem et Morin, 2012b) décrivent le système *QAlign* qui considère la tâche d'extraction lexicale comme un problème de question réponse. Ce système considère le terme à traduire comme une partie d'une question dans la langue source et tente de

trouver la traduction correcte en émettant l'hypothèse qu'elle devrait figurer dans la réponse (dans la langue cible) à cette question. Les résultats obtenus par cette méthode sont en deçà de ceux de l'approche standard.

L'étude présentée dans (Shezaf et Rappoport, 2010) introduit une méthode de création de lexiques bilingues (hébreu-espagnol) de très bonne qualité à partir d'un lexique bruité construit à partir de deux lexiques pivots. Cette méthode se base sur deux corpus monolingues. La partie la plus importante de l'algorithme est nommée la signature des contextes non-alignés, qui ne sont autres que les mots qui cooccurrent le plus souvent avec le terme à traduire dans le corpus.

D'autres études portant sur des problématiques plus spécifiques à la tâche d'extraction de lexiques bilingues à partir de corpus comparables ont été introduites. Par exemple, comme les mots rares apparaissant avec une faible fréquence dans le corpus ne disposent pas d'information contextuelle suffisante, il est difficile de trouver leur traduction dans le corpus. Dans (Pekar *et al.*, 2006), avant de comparer les espaces vectoriels interlingues, la méthode modélise les vecteurs de contexte des termes rares en utilisant leur similarité distributionnelle aux mots de la même langue. Cette modélisation permet de prédire de nouvelles cooccurrences et de lisser les cooccurrences rares, qui sont peu fiables. (Prochasson et Fung, 2011) s'intéressent également aux mots rares. Ils considèrent deux traits, à savoir une similarité de vecteur de contexte et un modèle de cooccurrence de mots dans des documents alignés dans une approche d'apprentissage automatique.

Toutes les approches présentées se concentrent sur l'extraction de lexiques bilingues de termes simples. Quelques travaux se sont néanmoins intéressés aux termes complexes. Ce type d'unité ne fait pas l'objet de notre étude. Pour cette raison, nous présentons brièvement les approches proposées pour les traiter. Comme la couverture des dictionnaires bilingues utilisés pour traduire les contextes pour les termes complexes est généralement très faible surtout pour des domaines de spécialité, l'approche proposée pour aligner ce type d'unités se base sur leur *propriété compositionnelle* et émet l'hypothèse que la signification de l'ensemble peut être appréhendée par la signification de ses parties (Daille, 2001; Robitaille *et al.*, 2006). Ainsi, cette approche dite compositionnelle consiste à traduire un terme complexe mot à mot à l'aide d'un dictionnaire bilingue. Ensuite, elle combine ces traductions individuelles selon des formes appropriées pour produire des traductions candidates du terme complexe. La traduc-

tion compositionnelle est un processus complexe, en raison de plusieurs phénomènes linguistiques décrits dans (Morin et Daille, 2006) comme par exemple :

- **La fertilité** : les termes ne se traduisent pas systématiquement par des termes de même longueur. Par exemple le terme complexe anglais *computer science* se traduit en français par le terme simple *informatique*. Ce problème a été bien décrit dans (Brown *et al.*, 1993).
- **La non-compositionalité** : Lorsqu'un terme complexe est traduit par un terme de même longueur, la traduction ne s'obtient pas par la traduction mot à mot de ses constituants.
- **La variation** : Un même terme peut se présenter sous différentes formes suite à des variations lexicales morphologiques ou syntaxiques. Par exemple, les termes français *cancer du poumon* et *cancer pulmonaire* se traduisent en anglais par le terme *lung cancer*. Ce critère est pris en compte dans (Morin et Daille, 2006), où des règles morphologiques sont ajoutées pour améliorer la traduction compositionnelle.

Les résultats de l'alignement de termes complexes sont généralement en deçà de ceux obtenus en utilisant les termes simples. Par exemple, (Morin et Daille, 2006) qui obtenaient 49 % de résultats corrects avec des termes simples n'obtiennent plus que 18 % de résultats corrects dans le $succès_{10}$. Bien qu'elle soit complexe, la tâche d'alignement de termes complexe n'est pas inutile du fait qu'en domaine de spécialité ou général, ces unités sont généralement plus précises et spécifiques.

L'exploitation des corpus comparables demeure un champ d'investigation qui évolue rapidement. L'atelier sur la construction et l'exploitation des corpus comparables (Building and Using Comparable Corpora workshop, BUCC) initié en 2008, est un lieu de rencontre annuel pour étudier les nouvelles techniques d'extraction lexicale à partir de ce type de corpus.

1.5 Conclusion

Nous avons présenté dans ce chapitre un tour d'horizon de l'extraction de lexiques bilingues à partir de corpus de textes multilingues parallèles et comparables. Un premier constat est que les approches utilisées pour l'acquisition de lexiques à partir de corpus parallèles ne sont pas valables avec ceux utilisant des corpus comparables

et vise versa. Plus intéressant encore est qu'il existe un large éventail de techniques manipulant ces deux types de corpus, dont certaines sont suffisamment simples, tandis que d'autres requièrent plutôt des techniques complexes. Néanmoins, la désignation de la technique la plus efficace reste ouverte à un débat, ce qui souligne l'impératif de poursuivre les recherches et l'exploration de nouvelles approches.

Au cours des premières années, apprendre à partir de *corpus parallèles* était la tâche primordiale de l'extraction de lexiques bilingues. Le grand nombre de publications consacrées aux textes parallèles montrent le chemin parcouru en à peine plus d'une vingtaine d'années. De nos jours et surtout en extraction de lexiques bilingues de mots simples, les travaux s'appuyant sur ce type de corpus ont déjà enregistré des succès significatifs. C'est pour cette raison que les recherches actuelles se sont penchées vers l'extraction d'unités lexicales complexes comme les expressions polylexicales, où de nombreuses difficultés subsistent, et que le champ de recherche dans ce domaine est encore largement ouvert. Dans cette étude, notre intérêt se porte sur cet objet linguistique. Nous présentons dans la **deuxième partie** de cette thèse une étude sur le traitement d'expressions polylexicales bilingues, allant de leur acquisition automatique à partir de corpus parallèles à leur intégration dans une application clé du TAL : la traduction automatique.

Nous avons également présenté les approches utilisées pour extraire des lexiques bilingues à partir de *corpus comparables*. Actuellement, les approches contextuelles comme l'approche standard sont les plus souvent utilisées. Pour construire des lexiques bilingues, ces approches reposent sur les cooccurrences des mots dans chaque langue. Néanmoins, leur principale différence réside dans les informations de cooccurrence qu'elles acquièrent du contexte. Toutes les recherches menées dans ce cadre possèdent un objectif spécifique. Certaines visent à traduire des mots issus du domaine général (Rapp, 1999a), tandis que d'autres se sont concentrés sur des termes spécifiques à des domaines particuliers (Chiao et Zweigenbaum, 2002; Morin *et al.*, 2008). En outre, un grand nombre de ces études sont axées sur la traduction de mots simples (Haghighi *et al.*, 2008; Diab et Finch, 2000) et de mots composés (Daille, 2001; Robitaille *et al.*, 2006). Dans ce cadre, nous présentons dans la **troisième partie** de ce manuscrit de nouvelles approches d'extraction de lexiques bilingues à partir de corpus comparables, qui s'intéressent à des termes simples.

Première partie

Extraction de lexiques bilingues à partir de corpus paralèlles

Introduction générale

Cette partie est consacrée à la présentation de nos contributions qui se portent sur l'extraction de lexiques bilingues à partir de corpus parallèles. Comme il a été mentionné dans le chapitre 1, l'extraction de lexiques bilingues de mots simple à partir de ce type de corpus peut être considérée comme une tâche bien maîtrisée. Or, la faiblesse de ces lexiques est leur manque de couverture pour les expressions polylexicales. Dans cette partie, notre intérêt se porte par ce type d'unités car, en plus du fait qu'elles soient fréquemment utilisées dans le langage oral et écrit, leur identification est fondamentale pour les applications faisant intervenir les aspects sémantiques de la langue et surtout la traduction automatique.

Cette partie est organisée de la manière suivante : nous présentons dans le chapitre 2 une approche qui identifie et aligne les expressions polylexicales dans un texte parallèle, et une approche qui étudie l'apport de ce type d'unités pour un système de traduction automatique (chapitre 3).

Chapitre 2

Lexique bilingue d'expressions polylexicales

2.1 Introduction

Depuis les années 90, les recherches en traitement automatique des langues (TAL) se sont intéressées aux expressions polylexicales (EPL, en anglais *MultiWord Expressions*) et aux problèmes qu'elles soulèvent. Une EPL peut être définie comme une combinaison de mots pour laquelle les propriétés syntaxiques ou sémantiques de l'expression entière ne peuvent pas être obtenues à partir de ses parties (Sag *et al.*, 2002). Les EPL sont fréquemment employées dans les textes écrits car elles constituent une part significative du lexique d'une langue. (Jackendoff, 1997) estime que la fréquence de leur utilisation est équivalente à celle des mots simples. Bien qu'elles soient facilement employées et reconnues par les humains, leur identification pose un problème majeur pour diverses applications du traitement automatique des langues, à savoir l'analyse syntaxique (Nivre et Nilsson, 2004; Constant *et al.*, 2011), le résumé automatique (Hogan *et al.*, 2007), l'extraction d'information (Vechtomova, 2005) et en particulier pour la traduction automatique (Carpuat et Diab, 2010; Ren *et al.*, 2009).

Dans cette étude, notre intérêt se porte sur la constitution d'un lexique bilingue dont les entrées sont constituées d'EPL en relation de traduction. Ces EPL sont extraites à partir d'un corpus parallèle français-anglais. En extraction lexicale, acquérir

des lexiques à partir de corpus parallèles est une tâche bien maîtrisée. Or un des points faibles de ces lexiques est souvent leur manque de couverture pour les EPL (Sagot *et al.*, 2005). L'alimentation de lexiques bilingues par ce type d'unités lexicales s'avère donc important pour les applications faisant intervenir l'aspect bilingue (traduction automatique et recherche d'information interlingue).

Ce chapitre est organisé de la manière suivante : nous présentons une méthode qui identifie tout d'abord les EPL dans chaque partie du corpus parallèle pour les mettre en relation de traduction en un second lieu. Avant de présenter cette approche (section 2.3), nous discutons dans la section 2.2 de la définition, des propriétés et de différentes typologies d'EPL. La section 2.4 est consacrée à la présentation du cadre expérimental et à l'évaluation menée dans ce cadre.

2.2 Expressions polylexicales

2.2.1 Définition

Les expressions polylexicales, dans le consensus actuel du domaine du TAL, forment des unités linguistiques qui contiennent un certain degré de non-compositionalité lexicale, syntaxique, sémantique et/ou pragmatique. Les EPL sont généralement composées d'un groupe de deux ou plusieurs mots dans une langue dont le sens ne peut pas être déduit de ses constituants. Dans la littérature, les termes *idiome, collocation et expression polylexicale ou encore expression multi-mot* sont couramment employés de manière interchangeable. (Calzolari *et al.*, 2002) définissent les EPL comme suit :

[. . .] Des phénomènes différents mais liés [. . .]. Généralement et à un certain niveau d'analyse linguistique, l'ensemble de ces phénomènes peuvent être décrits comme une séquence de mots qui agit comme une seule unité.

Dans quelques travaux (Ramisch *et al.*, 2013), les auteurs ne donnent pas une définition des EPL mais énumèrent des exemples. Dans cette étude, nous utilisons la définition proposée par (Calzolari *et al.*, 2002) pour caractériser une EPL.

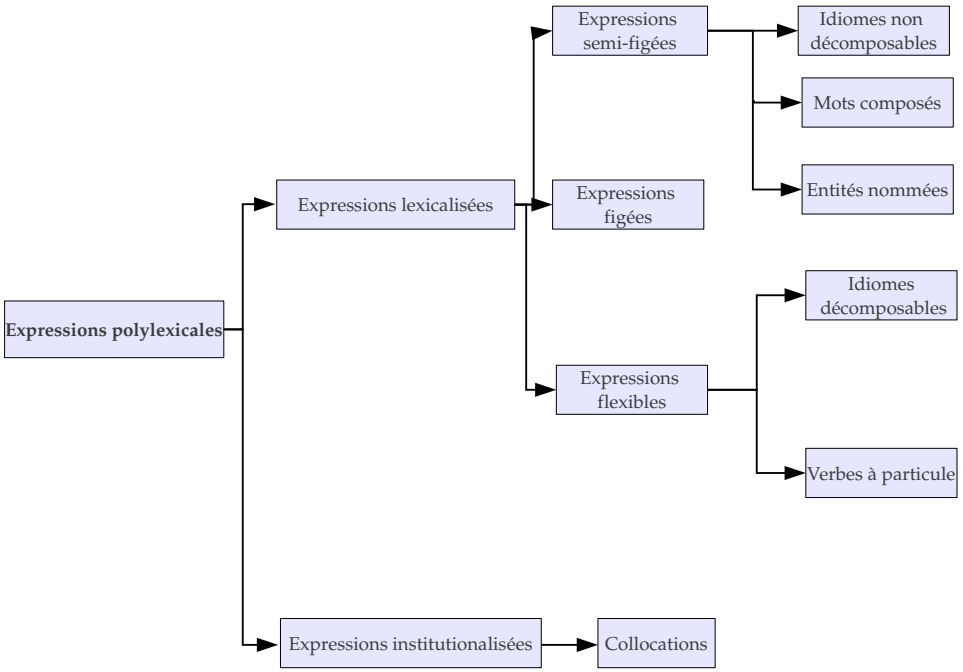


FIGURE 2.1: Typologie des expressions polylexicales selon (Sag *et al.*, 2002).

2.2.2 Typologie d'EPL

Les classifications et terminologies des EPL sont très nombreuses et variées dans la littérature linguistique. La classification la plus populaire actuellement dans la communauté internationale du TAL est celle décrite dans (Sag *et al.*, 2002). Ces derniers proposent de découper les EPL en deux classes : *les expressions lexicalisées* et les *expressions institutionalisées* (Figure 2.1). Les expressions lexicalisées possèdent un certain degré de figement syntaxique et/ou sémantique, qui peut être détecté par des critères linguistiques formels. Les expressions institutionalisées sont compositionnelles syntaxiquement et sémantiquement, mais sont statistiquement idiosyncratiques : les mots des expressions apparaissent ensemble soit par convention soit de manière habituelle comme par exemple l'EPL *"traffic jam"*. Nous présentons dans ce qui suit les types d'EPL de ces deux classes et détaillons ceux qui nous seront utiles dans la suite : *les collocations, les mots composés* et *les entités nommées*.

2.2.2.1 Les expressions lexicalisées

Expressions figées

Les expressions figées sont des combinaisons de plusieurs mots, non-compositionnelles du point de vue sémantique comme par exemple, l'expression *cul de sac* désignant une *impasse*. Les critères linguistiques pour déterminer si une combinaison de mots est une expression figée sont basés sur des tests syntaxiques et sémantiques. Par exemple, l'expression *boîte noire* est une expression figée car elle n'accepte pas de variations lexicales *(boîte sombre, caisse noire)* et elle n'autorise pas d'insertions *(boîte très noire)*.

Expressions semi-figées

Ces expressions respectent également les contraintes d'ordre des mots et de la non-compositionalité, mais elles sont soumises à un certain degré de variation lexicale, par exemple dans la forme de flexion. Il est ainsi possible de les considérer comme une unité complexe ayant une seule partie de discours mais qui est lexicalement variable à des positions particulières, comme par exemple la terminaison. Selon (Sag *et al.*, 2002), ces expressions prennent diverses formes, notamment des *idiomes non décomposables*, *des mots composés* et *des entités nommées*. Les idiomes non décomposables sont des expressions dont les composantes ne contribuent pas à la signification figurée de l'ensemble (par exemple, *kick the bucket* ou *shoot the breeze*). Les mots composés sont construits par une juxtaposition de deux mots permettant d'en former un troisième qui soit un mot à part entière et dont le sens ne se laisse pas forcément deviner par celui des deux constituants. Ainsi, un *garde-fou* est, en français, un lemme indépendant de *garde* et de *fou* dont le sens de « près d'un fossé, empêchant de tomber » ne peut être deviné. Les mots composés comme *car park, part of speech* sont similaire au idiomes non décomposables puisqu'ils sont également des unités non modifiable syntaxiquement. Les entités nommées sont des phénomènes qui ont été largement étudiés dans le TAL car ce sont des unités fondamentales pour plusieurs applications comme l'extraction d'information ou la traduction automatique. Les entités nommées comprennent de nombreux phénomènes linguistiques comme les noms propres (noms de personne, d'organisation, etc.), les expressions numériques ou les expressions de temps. Dans cette étude, nous nous intéressons plus particulièrement aux mots composés et aux entités nommées vu qu'ils apparaissent avec une fréquence élevé dans un texte et dans les textes du parlement européen, la plupart d'expressions sont constitués de mots composés et d'entités nommées.

Expressions syntaxiquement flexibles

Alors que les expressions semi-figées conservent le même ordre des mots, les expressions syntaxiquement flexibles présentent un éventail beaucoup plus large de variabilité syntaxique. Ce type d'expression se compose des *verbes à particule* et des *idiomes décomposables*. Les verbes à particule sont constitués d'un verbe plus une ou plusieurs particules comme par exemple *write up, look up*. Les idiomes décomposables ont tendance à être syntaxiquement souples dans une certaine mesure. Des idiomes comme *pop the question*, ou *spill the beans* sont décomposables, car chaque composant contribue à l'interprétation figurée de l'ensemble. Ce qui importe pour qu'un idiome soit considéré comme décomposable c'est que ses parties possèdent de la signification, littérale ou figurée, contribuant de façon indépendante à l'interprétation figurée de l'expression dans son ensemble.

2.2.2.2 Les expressions institutionalisées

Comme elles ont été définies plus haut, les expressions institutionalisées sont constituées essentiellement de *collocations*. Les collocations sont décrites comme des combinaisons de mots qui présentent des affinités et tendent à apparaître ensemble (pas forcément de manière contigüe) (Tutin et Grossmann, 2002), comme par exemple, *argument de poids*, *amour fou*. Il existe deux approches principales pour définir les collocations. Tout d'abord, en linguistique de corpus, les collocations sont considérées comme des combinaisons habituelles de mots au sens fréquentiel (Sinclair, 1991). Cette définition est celle utilisée le plus souvent par les chercheurs en TAL qui spécifient les collocations à l'aide de mesures associatives statistiques (Smadja *et al.*, 1996; Pecina, 2008). Elle est assez large et couvre toutes les EPL. Certaines collocations sont relativement figées comme par exemple *peur bleue*. Avec les critères utilisés dans le cadre du lexique-grammaire qui constitue à la fois une méthode est une pratique effective de description formelle des langues, ce type d'expressions serait considéré comme un mot composé : *peur (bleue ou rouge ou orange)*.

Dans cette étude, nous nous intéressons à l'identification d'expressions institutionalisées et de certaines expressions lexicalisées, plus particulièrement, nous donnons plus d'attention aux mots composés, collocation, noms propres et certaines expressions figées prépositionnelles (*en ce qui concerne, par rapport à, ...*) puisqu'elle constituent des EPL dont la fréquence est très élevée dans les textes.

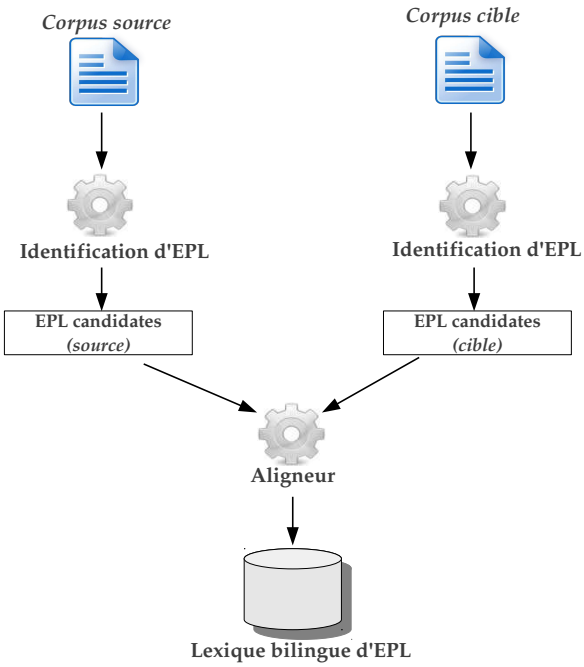


FIGURE 2.2: Vue d'ensemble du système d'extraction de lexique bilingue d'EPL.

2.3 Extraction de lexique bilingue

Dans cette section, nous décrivons l'approche proposée pour extraire un lexique bilingue d'EPL à partir d'un corpus parallèle français-anglais aligné au niveau de la phrase. Comme le montre la figure 2.2, notre approche d'extraction de lexique se compose de deux étapes. Dans la première étape, nous identifions les EPL candidates dans chaque partie du corpus parallèle. La deuxième étape consiste à mettre en relation les EPL de la langue source avec celles de la langue cible.

2.3.1 Identification monolingue d'EPL

2.3.1.1 EPL candidates

La méthode d'identification monolingue d'EPL est fondée sur une approche symbolique, très similaire à celle présenté dans (Okita *et al.*, 2010). S'ils définissent des patrons pour extraire seulement des mots composés, notre approche identifie à la fois des mots composés, des collocations, des expressions figées et des entités nommées.

Configuration	Anglais	Français
Adj-Noun	*plenary meeting*	*libre circulation*
Noun-Adj	*oil tanker*	*parlement européen*
Noun-Noun	*member state*	*état membre*
Past_Participle-Noun	*developped country*	...
Noun-Past_Participle	*parliament adopted*	*pays developpé*
Adj-Adj-Noun	*european public prosecutor*	...
Adj-Noun-Adj	*social market economy*	*bon conduite administratif*
Adj-Noun-Noun	*renewable energy source*	...
Noun-Noun-Adj	...	*industrie automobile allemand*
Noun-Adj-Adj	...	*ministère public européen*
Adj-Noun-Adj	*social fund assistance*	*important débat politique*
Noun-Prep-Noun	*point of view*	*chemin de fer*
Noun-Prep-Adj-Noun	*court of first instance*	*court de premier instance*
Noun-Prep-Noun-Adj	...	*source d'énergie renouvelable*
Adj-Noun-Prep-Noun	*european court of justice*	...
Noun-Adj-Prep-Noun	...	*politique européen de concurrence*

TABLEAU 2.1: Configurations morphosyntaxiques permises. Les EPL candidates extraites sont sous formes de lemmes. Les (...) correspondent à des patrons ne relevant aucune EPL dans notre corpus.

La méthode proposée requiert simplement une analyse morphosyntaxique des textes source et cible comme étape préliminaire à la procédure de construction d'expressions. À cet effet, nous faisons appel à la plateforme d'analyse multilingue LIMA du CEA-LIST (Besançon *et al.*, 2010). LIMA est composée d'un ensemble de modules dont le nombre et la nature varient selon la langue traitée et d'un ensemble de ressources linguistiques. Cet analyseur produit pour chaque texte une liste de lemmes étiquetés par leurs catégories grammaticales.

Comme la plupart des EPL sont constituées de combinaisons de noms, d'adjectifs ou encore de prépositions, nous produisons une liste de n-grammes candidats ($2 \leq n \leq 4$) dont la structure morphosyntaxique respecte une configuration prédéfinie, telle que celles décrites dans le tableau 2.1. Ces configurations ou patrons d'extraction , au nombre de seize, sont définies manuellement. Dans certains cas et pour certaines configurations aucune EPL n'est produite. Tel est le cas pour la configuration *Past_Participle-Noun* pour le français. L'utilisation de configurations morpho-

syntaxiques permet de ne garder que des n-grammes jugés pertinents et d'écarter ceux constitués de mots agrammaticaux comme par exemple *"is a, of the, de la"*.

Il est également important de noter que les EPL candidates sont constituées de lemmes plutôt que de formes de surface. Parmi les avantages de la lemmatisation, on peut citer la réduction du nombre de formes à considérer et l'augmentation des occurrences de chaque forme dans le corpus. En outre, cette opération linguistique permet de remédier au problème de dispersion de données. On notera aussi qu'en extraction lexicale, il est toujours préférable que les entrées d'un lexique soient lemmatisées.

Nous ajoutons à cet ensemble de candidats produits par les patrons d'extraction des expressions figées et des entités nommées reconnues par la plateforme LIMA. La reconnaissance d'expressions figées se fait à l'aide d'un processus de reconnaissance de formes basé sur des automates à états finis, permettant ainsi de reconnaitre reconnaître par exemple, les expressions *"in the light of, with regard to, en ce qui concerne ..."* comme des unités uniques. En ce qui concerne les entités nommées, cette étape de l'analyse utilise des fichiers de listes ainsi que des automates à états finis. Ainsi, un énoncé comme *"Le Moyen-Orient"* est reconnu comme un nom de lieu.

2.3.1.2 Heuristiques de filtrage

Le résultat de l'identification des EPL est représenté par une liste d'EPL candidates ordonnées en fonction de leur fréquence dans le corpus. Plusieurs candidats parmi ceux produits apparaissent imbriqués dans d'autres. Afin d'éviter un effet de surgénération, où nous identifions par exemple, des candidats qui sont imbriqués dans d'autres, nous proposons de filtrer la liste des candidats obtenue. Dans la littérature, le filtrage se fait par l'utilisation de mesures d'association (Daille, 2001; Seretan et Wehrli, 2007; Vintar et Fisier, 2008). Contrairement à ces travaux, notre système n'applique pas de filtre fondé sur des mesures d'association ou sur la fréquence. Nous proposons par contre deux heuristique de filtrage visant à ne garder que les expressions qui sont susceptibles de constituer des EPL. Ces heuristiques se basent plutôt sur la taille des EPL et considèrent que :

- Si une expression est imbriquée dans une autre et qu'elles apparaissent avec la même fréquence (par exemple les EPL *"first instance"* et *"court of first instance"*), on ne garde que la plus couvrante (plus longue).
- Si une expression apparaît dans un grand nombre d'autres expressions, nous suivons l'approche proposée par (Frantzi *et al.*, 2000) et éliminons toutes les expressions plus longues. Par exemple l'EPL *"member state"* apparaît dans les EPL candidates *"each member state, member state exercise, member state national, all member state"*. Dans ce cas, nous considérons que les EPL plus longues ne sont pas assez pertinentes et ne gardons que l'EPL *"member state"*.

Nous prenons en considération toutes les expressions extraites, aussi bien fréquentes que non fréquentes et celles dont les constituants ont un degré de corrélation élevé ou faible. À notre connaissance, aucune approche d'extraction d'EPL n'a aussi pris en considération l'ensemble des EPL trouvées.

2.3.2 Alignement d'EPL : approche par comparaison de distributions

Dans cette section, nous présentons la méthode d'alignement que nous proposons pour construire le lexique bilingue d'EPL. Cette méthode tente de trouver, pour chaque EPL de la langue source, la traduction qui lui correspond dans l'ensemble d'EPL de la langue cible. Cette tâche pose de sérieux problèmes en l'absence de ressources externes. C'est la raison pour laquelle la plupart des travaux de recherche portant sur l'alignement d'EPL utilisent des dictionnaires bilingues de mots simples et des règles de traduction compositionnelle (Semmar *et al.*, 2010) ou des outils d'alignement de mots simples (Ren *et al.*, 2009; Deleger *et al.*, 2009) pour mener à bien la tâche l'alignement. Dans la présente étude, la méthode d'alignement que nous proposons est indépendante de toute ressource externe, elle requiert simplement un corpus parallèle et la liste des EPL candidates dans les langues source et cible.

Notre approche hérite de la sémantique distributionnelle, où nous associons à chaque EPL source et cible une représentation spécifique qui servira par la suite de base pour l'établissement d'une relation de traduction entre chaque paire d'EPL (source, cible). Elle consiste à construire pour chaque EPL source (respectivement cible), l'empreinte de sa distribution dans la partie source (respectivement cible) du corpus parallèle. Cette approche s'appuie sur l'hypothèse qu'il n'y a pas de traduc-

Id.Phrase	Phrase
2	*. . . semblerait être* ***à nouveau*** *mis en accusation, le ministère public . . .*
55	*. . . vous demande donc* ***à nouveau*** *de faire le nécessaire . . .*
$n-1$	*. . . aussi de promouvoir* ***à nouveau*** *l'activité des femmes . . .*
n	*. . . que le règlement soit* ***à nouveau*** *modifié en collaboration . . .*

⇓

	1	2	3	4	. . . . .	55	. . . . .	$n-1$	n
à nouveau	0	1	0	0	. . . . .	1	. . . . .	1	1

Figure 2.3: Représentation vectorielle de l'expression *" à nouveau "*. Id.Phrase correspond à un identifiant unique de la phrase contenant l'expression dans notre corpus.

tions manquantes dans le corpus parallèle (Fung, 1995). Puisqu'aucune traduction ne manque, à chaque fois qu'une EPL apparaît à un endroit dans le corpus source, sa traduction apparaîtra à une position comparable dans le corpus cible. Ceci nous conduit par conséquent à émettre l'hypothèse que *la distribution d'une EPL et de sa traduction sont similaires.*

L'algorithme d'alignement que nous proposons enregistre cette distribution dans un vecteur de booléens, en notant la présence ou non des EPL source (respectivement cible) dans les phrases du corpus source (respectivement cible). À titre d'exemple, nous présentons le vecteur représentant l'EPL française *" à nouveau "* dans la figure 2.3. Reste donc à comparer les vecteurs sources avec les vecteurs cibles pour repérer les distributions les plus semblables. Plus en détail, notre algorithme d'alignement est *itératif* et opère de la façon suivante :

1. Trouver l'EPL la plus fréquente dans chaque phrase source.

2. Extraire les EPL cibles qui apparaissent dans toutes les phrases parallèles à celles où figure l'expression source.

3. Calculer un score de confiance pour chaque couple (source, cible). Ce score est calculé sur la base de l'indice de Jaccard (équation 2.1).

$$\text{Jaccard} = \frac{|V_s \cdot V_t|}{|V_s|^2 + |V_t|^2 - |V_s \cdot V_t|} \tag{2.1}$$

   Cet indice est fondé principalement sur le calcul du rapport entre le cardinal de l'intersection $|V_s \cdot V_t|$, qui est représentée ici par le nombre de phrase partagées

par chaque paire d'expression source et cible et le cardinal de l'union de ces expressions V_s et V_t.

4. Considérer l'expression cible qui maximise ce score comme la meilleure traduction.
5. Supprimer la paire de traductions du processus et retourner vers 1.

Cet algorithme qu'on qualifie de *glouton* cherche à chaque itération à établir une relation de traduction obtenant un maximum local, dans l'espoir d'obtenir un résultat optimum global. Ce maximum local est obtenu en favorisant les EPL fréquentes (première itération). Ainsi, outre l'hypothèse qu'une EPL et sa traduction partagent la même distribution, notre algorithme émet l'hypothèse qu'une EPL source et cible sont des traductions mutuelles si elles apparaissent avec des fréquences comparables dans le corpus parallèle. Une approche d'alignement itérative a été également présentée dans (Smadja *et al.*, 1996). Cette approche est différente de la nôtre dans la façon dont la traduction d'une EPL est extraite. Si nous utilisons une liste d'EPL candidates extraites au préalable, dans (Smadja *et al.*, 1996), les auteurs construisent la traduction en ajoutant un mot à chaque itération à un mot simple extrait à la première itération. La limite principale de cette approche est que, puisque basée sur l'ajout incrémental de mots à la traduction candidate, il est est difficile de savoir quant est ce que on arrête.

2.4 Evaluation

Deux évaluations ont été menées dans ce cadre. Nous évaluons à la fois la qualité de la méthode d'identification d'EPL et celle du lexique bilingue d'EPL extrait par notre aligneur. Dans ces évaluations, nous comparons les résultats d'identification et d'alignement d'EPL à des références créées manuellement. Nous avons réalisé une annotation manuelle des EPL dans un corpus de test constitué de 300 paires de phrases extraites aléatoirement du corpus Europarl (Koehn, 2005). Trois références ont été créées à partir de ce corpus à l'aide de l'outil d'annotation en ligne Yawat (Germann, 2008), à savoir deux références pour l'identification d'EPL en français et en anglais et une référence pour l'alignement français-anglais.

La performance des méthodes d'identification et d'alignement proposées est évaluée en utilisant les mesures usuelles de *précision* (P), *rappel* (R) et *F-mesure*

	P (%)	R (%)	F_1 (%)
Identification			
anglais	92,2	32,7	48,2
français	84,4	30,4	44,6
Alignement			
français-anglais	41,2	13,9	20,7

TABLEAU 2.2: Résultats d'identification et d'alignement d'EPL en termes de précision (P), rappel (R) et F-mesure (F_1).

(F_1). Ces mesures sont définies respectivement comme la proportion des candidates proposées appartenant à la référence et la proportion des éléments de la référence proposés parmi les candidates. La F_1 combine les deux mesures avec une importance égale. Le tableau 2.2 regroupe les résultats obtenus. Pour l'identification des EPL, Nous rapportons une valeur de F-mesure de 48,2 % pour l'anglais et de 44,6 % pour le français. Nous constatons également que les valeurs de précision sont très élevées par rapport à celles du rappel. La précision de l'identification des EPL atteint $92,2$ % pour l'anglais pour un rappel de $32,7$ %. Pour le français, les résultats vont dans la même direction ($84,4$ % de précision et $30,4$ % de rappel).

En ce qui concerne l'alignement d'EPL, et à partir des 300 paires de phrases alignées, un lexique bilingue composé de 116 paires d'expressions est extrait. Une faible valeur de F-mesure est rapportée : $20,7$ %. Ceci s'explique par le fait que, notre méthode d'identification ne capture que des EPL contiguës, alors que dans le corpus de test, plusieurs EPL se traduisent par des EPL non contiguës. En plus, nous avons posé l'hypothèse qu'une EPL se traduit généralement par une EPL. Néanmoins, nous avons pu identifier plusieurs cas dans lesquels une EPL se traduit par un mot simple. Par exemple, l'expression française *chemin de fer* se traduisant dans la référence par le mot simple anglais *railway*, est mise en correspondance par notre aligneur avec l'expression *railway sector*.

Il est également intéressant de rappeler que comme, notre méthode d'identification d'EPL se base sur des patrons morphosyntaxiques, deux facteurs qui lui sont indispensables pourraient affecter les résultats. D'une part, l'outil d'étiquetage morphosyntaxique apporte un taux d'erreurs qui est sans doute faible mais qui peut néanmoins influencer les résultats. D'autre part, il est fort probable que les patrons

Français	→	Anglais
parlement européen	→	european parliament
état par état	→	amount of state
coup d'état	→	military coup
zone non fumeur	→	no smoking area
insulaire en développement	→	small island developing
de bonne foi	→	good faith
politique de concurrence	→	competition policy
chemin de fer	→	railway sector
en ce qui concerne	→	in regard to
en ce qui concerne	→	as regards
en ce qui concerne	→	with reference to
en ce qui concerne	→	with respect to
coupe forestier	→	cut in forestation

TABLEAU 2.3: Exemples d'EPL bilingues alignées par notre aligneur.

morphosyntaxiques définis manuellement ne réussissent pas à couvrir la totalité des EPL du corpus parallèle.

En observant certaines paires d'expressions du tableau 2.3, nous remarquons que malgré les faibles scores du rappel obtenus, notre méthode présente plusieurs avantages. Premièrement, pour trouver la traduction adéquate pour chaque EPL et contrairement à la plupart des travaux antérieurs (Dagan et Church, 1994; Ren *et al.*, 2009) qui reposent sur la traduction mot à mot des composantes d'une EPL, notre méthode capture l'équivalence sémantique entre les EPL en n'ayant recours à aucune information préalable sur l'alignement des mots. Elle permet aussi d'aligner des expressions à caractère idiomatique telles que *à nouveau* → *once more* ou encore *état par état* → *amount of state* et peut trouver de multiples correspondances bilingues possibles pour les EPL source pour lesquelles plusieurs EPL cible correctes existent. Par exemple, notre méthode fournit pour l'EPL *en ce qui concerne* les traductions suivantes : *in regard to, with reference to, with respect to, as regards.*

2.5 Conclusion

Nous avons présenté dans ce chapitre une approche qui identifie et aligne des EPL bilingues à partir d'un corpus parallèle français-anglais. La méthode d'identification

d'EPL repose sur une approche symbolique et permet d'extraire des mots composés, des collocations, des expressions figées prépositionnelles et des entités nommées. Les expérimentations menées dans ce cadre montrent que cette méthode rapporte un haut niveau de précision et un faible niveau de rappel.

Nous avons également proposé un algorithme d'alignement qui se base sur la comparaison des distributions des EPL sources et cibles. Cet algorithme effectue des alignements de type $m-n$ avec $m > 1$ et $n > 1$ et prend en considération des EPL dont les constituants ont un degré de corrélation élevé ou faible. Une précision de $41,2\%$ et un très faible rappel de $13,9\%$ ont été obtenus. Plusieurs facteurs sont à la base de ces faibles scores comme notamment la définition des patrons morphosyntaxiques qui ne couvrent pas la totalité d'expressions du corpus, le choix de la taille des expressions et le fait que l'on pose l'hypothèse qu'une EPL se traduit forcément par une EPL. Même s'il est difficile de comparer les résultats de différentes approches d'extraction de lexiques d'EPL, nos résultats se rapprochent de ceux de (Lambert et Banchs, 2005), qui ont aussi rapporté un très faible rappel de $13,4\%$.

Il n'existe à ce jour aucun protocole commun pour l'évaluation des résultats d'alignement d'EPL. Ce manque de ressource rend impossible la comparaison des différentes approches en extraction de lexiques bilingues d'EPL. Outre l'évaluation intrinsèque menée dans ce cadre, il serait également intéressant de conduire une évaluation extrinsèque de la qualité du lexique bilingue d'EPL : c'est ce qui sera décrit dans le chapitre suivant.

Chapitre 3

Application des expressions polylexicales à un système de traduction statistique

3.1 Introduction

Comme il a été mentionné dans le chapitre , les lexiques bilingues constituent un outil précieux pour différentes applications du TAL. Néanmoins, le but principal de la majorité des travaux de recherche menés sur cet objet linguistique était l'acquisition *a priori* de correspondances entre des paires d'unités textuelles pour l'enrichissement de ressources lexicales. En revanche, relativement peu de travaux ont été réalisés sur l'exploitation de telles ressources, afin de rendre possible leur intégration dans des applications clés, comme la traduction automatique ou la recherche d'information interlingue.

Dans le chapitre précédent, nous avons décrit l'approche suivie pour acquérir un lexique bilingue d'EPL pour la paire de langues français-anglais. Pour évaluer sa qualité, nous avons mené une évaluation intrinsèque dans laquelle nous comparons les paires d'EPL bilingues acquises à un alignement de référence créé manuellement. Intuitivement, les EPL bilingues sont utiles pour améliorer les résultats de la Traduction Automatique Statistique (TAS). Avec l'émergence des systèmes de TAS à

base de segments (*phrase based approaches* en anglais) (Koehn *et al.*, 2003), diverses améliorations ont été obtenues. Ces segments sont définis comme de simples n-grammes systématiquement traduits dans un corpus parallèle sans motivation linguistique particulière. Dans de tels systèmes, le manque d'un traitement adéquat des EPL pourrait affecter la qualité de la traduction. En effet, la traduction littérale d'une expression non reconnue par le système de traduction comme une EPL constitue une cause principale de traduction erronée et incompréhensible (Ren *et al.*, 2009). Par exemple, un tel système proposera « *way of iron* » comme traduction pour « *chemin de fer* » au lieu de « *railway* » . Pour pallier ce manque, il est important d'utiliser un lexique dans lequel les EPL sont prises en compte.

Cependant, une difficulté consiste à trouver la meilleure façon d'intégrer les EPL dans de tels systèmes. (Lambert et Banchs, 2005) introduisent une méthode dans laquelle les EPL sont considérées comme un élément unique dans le corpus d'apprentissage. Lors de l'estimation du modèle de traduction, ces unités sont traitées comme des mots simples. En exploitant un corpus de petite taille, ils ont montré que la qualité de l'alignement et la précision de traduction ont été améliorées. Cependant, ils ont obtenu, dans des études plus récentes (Lambert et Banchs, 2006), basées sur un corpus de taille importante, un score BLEU (Papineni *et al.*, 2002) plus bas. Nous citons notamment les travaux de (Ren *et al.*, 2009) qui implémentent une méthode permettant d'intégrer des termes multi-mots issus du domaine médical dans Moses (Koehn *et al.*, 2007). Leur méthode a permis de gagner $0,17$ points de score BLEU par rapport au système de référence.

Dans ce chapitre, notre objectif est double : d'une part, nous considérons la TAS comme un mode d'évaluation extrinsèque de l'utilité des EPL. D'autre part, nous explorons différentes stratégies d'intégration de ces unités dans un système de TAS. Avant de décrire les différentes stratégies proposées (section 3.3), nous présentons dans la section 3.2 les principes du modèle standard utilisée en TAS. La section 3.4 est consacrée à la présentation du cadre expérimental, des différentes évaluations menées dans ce cadre et des résultats obtenus.

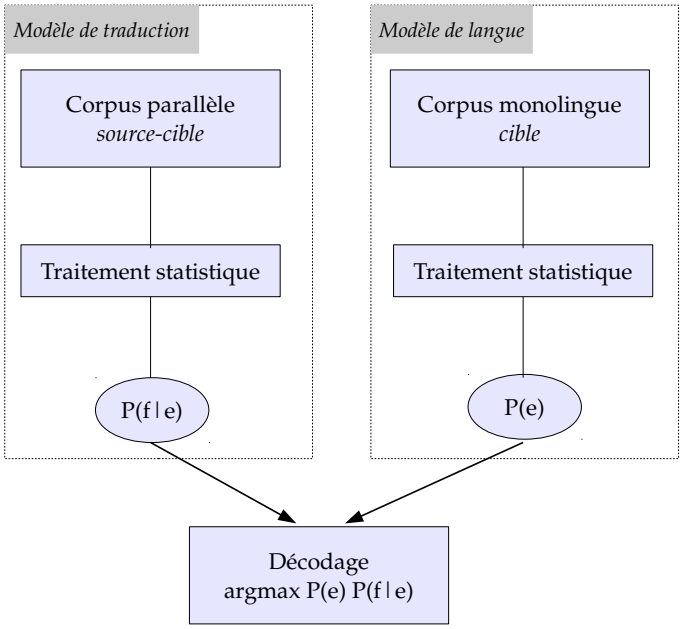


FIGURE 3.1: Vue d'ensemble d'un système de traduction statistique.

3.2 Traduction automatique statistique

3.2.1 Traduction statistique : modèle standard

Nous présentons dans cette section une vue d'ensemble des systèmes de traduction statistique. En fonction d'une distribution des probabilités $P(e|f)$, une phrase en langue source notée f et composée de i mots $f_1 \dots f_i$ est traduite en une phrase en langue cible e et contenant j mots $e_1 \dots e_j$. Historiquement, ce problème a été abordé en le transformant par l'application de la règle de Bayes de la manière suivante :

$$e^* = \arg\max_e P(e|f) \tag{3.1}$$

$$= \arg\max_e P(f|e)P(e) \tag{3.2}$$

Ce modèle est connu dans la littérature sous le nom de modèle de canal bruité (noisy chanel model) et décompose le problème en deux sous-problèmes plus simples

(équation 3.2). D'un côté le développement d'un *modèle de traduction* $P(f|e)$ estimé sur des corpus bilingues parallèles alignés au niveau de la phrase. Ce modèle de traduction sert de pont entre les langues source et cible. Son rôle est de guider la construction, pour chaque phrase source, d'un ensemble d'hypothèses de traduction en langue cible. De l'autre coté, le développement d'un *modèle de langue* $P(e)$ dont le rôle est de guider la recherche des séquences de mots les plus probables en se basant sur des connaissances extraites d'un corpus monolingue de la langue cible. Cette modélisation doit concentrer sa probabilité sur les phrases grammaticales indépendamment de la phrase source. Un aperçu général du développement d'un système de TAS est représenté dans la figure 3.1.

Pour inverser l'équation 3.1, (Brown *et al.*, 1993) supposent que la phrase à traduire f est grammaticalement bien formée et l'on souhaite construire une traduction e qui soit également bien formée. Le modèle de probabilité impliqué dans l'équation 3.1 doit être tel que pour toute phrase source f, il concentre la masse de probabilité sur des phrases en langue cible qui sont à la fois bien formées et qui sont des traductions de f. En plus de la justification théorique, le découplage réalisé par l'équation 3.2 présente un intérêt pratique puisqu'il sépare le problème de modélisation en deux sous problèmes indépendants (Gaussier et Yvon, 2011, chapitre 6, page 277).

3.2.2 Moses : TAS à base de segments

Moses (Koehn *et al.*, 2007) est un système de traduction libre, implémentant l'approche de traduction décrite dans la section précédente. Alors que les premiers systèmes de TAS travaillaient sur des mots, l'unité de traduction utilisée dans Moses est le *segment* (*phrase* en anglais), qui correspond à un groupe de mots contigus qui n'est pas forcément un syntagme au sens linguistique du terme. Le modèle de traduction rassemble donc un ensemble de bisegments en relation de traduction estimés à partir d'un corpus parallèle. Cette tâche est non-triviale et la constitution d'alignements sous-phrastiques à partir de phrases en relation de traduction nécessite des connaissances sur la traduction des unités qui composent ces phrases. La première étape consiste alors à construire des alignements de mots pour chaque paire de phrases, à l'aide d'un modèle d'alignement mot-à-mot. Cet alignement est construit pour le corpus parallèle dans les deux directions (source/cible et cible/source). En-

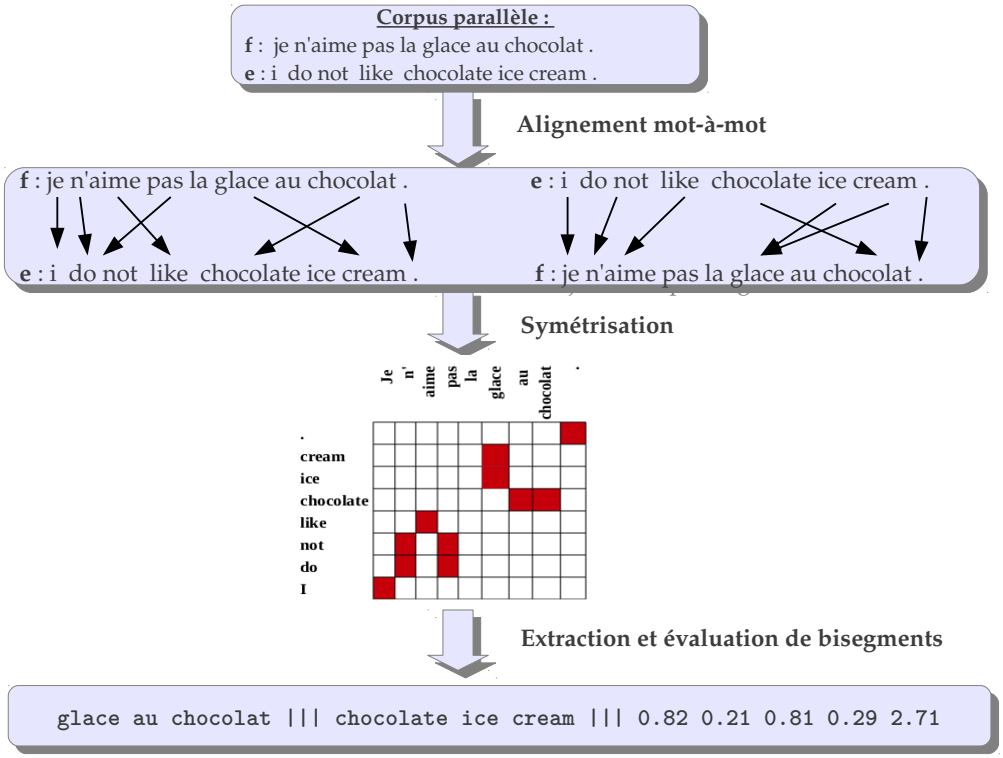


FIGURE 3.2: Vue d'ensemble du processus de construction de table de traduction. Cette figure est extraite de l'étude présentée dans (Gaussier et Yvon, 2011).

suite, pour les prendre simultanément en compte, il est courant d'utiliser des *heuristiques de symétrisation*, afin de produire un alignement unique.

Une première heuristique simple consiste à prendre l'*union* des deux alignements source/cible et cible/source. L'alignement résultant exploite au maximum les deux directions d'alignement, au risque de proposer des liens peu sûrs qui n'existent que pour une direction. Une deuxième heuristique consiste à sélectionner *l'intersection* des deux alignements d'entrée. Dans ce cas, les alignements obtenus sont plus fiables puisqu'ils sont identifiés dans les deux directions. (Och, 2003) propose différentes heuristiques dans le but de compléter l'alignement construit par l'intersection avec certains alignements figurant dans l'union.

Une fois que les deux alignements sont symétrisés, il reste à extraire et à évaluer l'ensemble de bisegments. L'extraction de bisegments repose sur des heuristiques d'extraction fondées sur la notion de cohérence d'un bisegment. L'évaluation d'un bisegment repose à la fois sur des statistiques accumulées sur l'ensemble du corpus parallèle

et sur l'exploitation du modèle d'alignement décrit dans la section précédente. Lors de la phase de décodage, ces hypothèses de traduction sont sélectionnées à partir d'un inventaire constitué d'un ensemble d'appariements entre des segments de longueur variable. Ces associations et les scores qui les accompagnent constituent la table de traduction *(phrase table)*. La figure 3.2 décrit le processus de construction d'une table de traduction présenté dans (Gaussier et Yvon, 2011). Les cinq scores calculés pour chaque bisegments correspondent à :

1. la probabilité de traduction $p(f|e)$
2. la probabilité lexicale $lex(f|e)$, qui évalue la qualité des alignements de mots qui constituent l'alignement d'un bi-segments.
3. la probabilité de traduction $p(e|f)$
4. la probabilité lexicale $lex(e|f)$
5. la pénalité d'apparition d'un segment (toujours $\exp(1) = 2.71$). Ce chiffre fournit un moyen de s'assurer que les traductions ne sont pas trop longues ou trop courtes.

3.3 EPL dans Moses

Rappelons que l'objectif de cette étude consiste à étudier l'impact de l'utilisation du lexique bilingue d'EPL construit en utilisant l'approche présentée dans le chapitre 2 pour un système de TAS. À cette fin, nous utilisons Moses comme système de traduction de base. Cependant, comme il a été mentionné dans la section 3.1, des recherches et expérimentations sont nécessaires pour trouver la meilleure façon d'intégrer ces connaissances linguistiques dans ce type de système. Dans cette section, nous explorons différentes stratégies et présentons :

- Trois stratégies d'intégration *dynamiques* où nous cherchons à modifier le modèle de traduction de différentes façons pour une meilleure prise en considération des EPL bilingues.
- Une stratégie d'intégration *statique* dans laquelle nous incorporons ces unités sans changer le modèle de traduction.

3.3.1 Stratégies d'intégration dynamiques

3.3.1.1 Nouveau modèle de traduction

Les tables de traduction constituent la source principale de connaissance pour le décodeur. Le décodeur consulte ces tables pour déterminer comment traduire une phrase source en langue cible. Cependant, en raison d'erreurs dans l'alignement automatique de certains mots, des segments extraits peuvent être dénués de sens (Och, 2003). Pour remédier à ce problème, nous proposons de considérer les EPL comme des paires de phrases parallèles. Ces paires de phrases sont ajoutées au corpus d'apprentissage pour estimer un nouveau modèle de traduction. Dans cette méthode que l'on note TRAIN, nous espérons que par l'augmentation du nombre d'occurrences des paires d'EPL, considérées comme de bons segments, une modification de l'alignement et de la probabilité de la traduction sera enregistrée.

3.3.1.2 Extension de la table de traduction

Dans cette stratégie, nous étendons la table de traduction de Moses. Dans (Wu *et al.*, 2008), les auteurs proposent une méthode dans laquelle ils enrichissent la table de traduction par un dictionnaire bilingue crée manuellement. Plutôt que de le construire manuellement, nous proposons dans la présente stratégie d'incorporer le lexique bilingue d'EPL construit automatiquement dans la table de traduction. Rappelons que les entrées du lexique bilingue obtenu sont pondérées par une valeur de similarité *(indice de Jaccard)*. Cette valeur de similarité est utilisée ici pour définir la probabilité de traduction dans les deux directions. En ce qui concerne la probabilité lexicale, nous la fixons simplement à 1. L'intuition qui sous-tend cette stratégie notée TABLE est que lors de la phase de recherche de segments candidats, le décodeur prendra en considération aussi bien les segments de base que les EPL bilingues.

3.3.1.3 Trait additionnel pour les EPL

(Lopez et Resnik, 2006) ont souligné qu'une meilleure définition des traits utilisés peut conduire à un gain substantiel dans la qualité des traductions. Nous suivons cette hypothèse et étendons la stratégie TABLE. En plus de l'incorporation des EPL

dans la table de traduction, nous définissons un nouveau *trait binaire* indiquant pour chaque entrée de la table de traduction s'il s'agit d'une EPL ou pas. Le but de cette stratégie notée TRAIT est de guider le système pour choisir les EPL bilingues plutôt que les hypothèses initiales.

3.3.2 Stratégie d'intégration statique

Dans cette stratégie, notée FORCÉ, nous voulons que le décodeur prenne en considération des EPL bilingues tout en gardant le modèle de traduction de base. À cet égard, nous utilisons le *mode de décodage forcé* du décodeur de MOSES. Ce dernier comporte un schéma de balisage XML permettant de spécifier des traductions pour des parties des phrases à traduire. Nous pouvons ainsi indiquer au décodeur ce qu'il faut utiliser pour traduire certains mots ou segments dans les phrases à traduire.

Dans le cadre de notre étude, nous représentons chaque EPL apparaissant dans le corpus de test par la balise XML adéquate en se basant sur les traductions produites par notre aligneur. Un exemple de représentation de l'EPL « *à nouveau* » est présenté ci-dessous :

... sembler être **à nouveau** mis en accusation, le ministère public ...

⇓

... sembler être < ***mwe*** **translation=*"once more"*** >**à nouveau**< /***mwe***> mis en accusation, le ministère public ...

3.4 Expériences et résultats

3.4.1 Cadre expérimental

3.4.1.1 Corpus et outils

Le système de traduction de référence noté RÉF et utilisé dans nos expérimentations repose sur Moses. Les données d'apprentissage, de développement et de test pro-

	Français	Anglais
Données d'apprentissage parallèles		
Textes du parlement européen	100 000	100 000
Données d'apprentissage monolingue		
Textes du parlement européen	-	1,8 M
Lexique bilingue d'EPL		
Textes du parlement européen	3 640	
Données de développement et d'évaluation		
Dev	4 000	4 000
Tous_Test	1 000	1 000
EPL_*Test*	323	323

TABLEAU 3.1: Données d'apprentissage, de développement et d'évaluation en nombre de phrases et taille du lexique bilingue en nombre de paires d'EPL en relation de traduction.

viennent du corpus Europarl (Koehn, 2005). Ce corpus regroupe un ensemble de phrases parallèles extraites des actes du parlement européen traduits dans 11 langues européennes. Dans nos expérimentations, nous nous intéressons à la traduction de textes du *français* vers l'*anglais*. Le tableau 3.1 rassemble des informations sur la taille des données d'apprentissage, de développement et d'évaluation. Le corpus d'apprentissage utilisé pour estimer le modèle de traduction est composé de 100 000 paires de phrases parallèles après normalisation. La normalisation est établie à travers les traitements suivants : segmentation en mots, suppression de phrases de plus de 50 mots et lemmatisation à l'aide de l'étiqueteur morphosyntaxique TreeTagger[1]. Ce même corpus est utilisé pour construire le lexique bilingue d'EPL. Le lexique bilingue résultant est composé de 3 640 paires d'EPL en français et leurs traductions en anglais. Comme les entrées de ce lexique sont sous forme de lemmes et que le mode de décodage forcé de Moses n'est actuellement pas compatible avec les modèles à base de facteurs, le modèle de traduction a été estimé sur des lemmes plutôt que sur des formes de surface.

Outre le modèle de traduction, nous avons estimé un modèle de langue trigramme sur une version lemmatisée de la totalité du corpus Europarl (1,8 M) en utilisant la boite à outils de calcul des modèles de langue IRSTLM[2].

1. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
2. http://hlt.fbk.eu/en/irstlm

Les stratégies dynamiques et statique décrites précédemment sont ensuite appliquées. Dans TRAIN, les EPL bilingues sont ajoutées au corpus d'apprentissage pour estimer un nouveau modèle de traduction. En ce qui concerne TABLE, la table de traduction de RÉF est enrichie par les EPL bilingues. Dans TRAIT, un trait additionnel 1/0 est introduit dans la table de traduction de TABLE. Finalement, FORCÉ maintient le modèle de traduction de RÉF. Tous les modèles obtenus sont optimisés par minimisation du taux d'erreur (*MERT : Minimum Error Rate Training*) (Och, 2003) sur un corpus de développement constitué de 4 000 paires de phrases issues du même corpus (tableau 3.1).

Deux séries d'expériences ont été menées : *Tous_Test* et EPL_*Test*. Le premier corpus de test *Tous_Test* est constitué de 1 000 paires de phrases parallèles extraites aléatoirement du corpus Europarl. Pour mesurer l'apport réel du lexique bilingue d'EPL, nous avons constitué un corpus de test noté EPL_*Test* où nous ne conservons que les phrases du corpus *Tous_Test* contenant au moins une EPL. Ce corpus contient 323 paires de phrases parallèles.

3.4.1.2 Qualité d'une traduction

Déterminer la qualité d'une traduction est un problème difficile et ouvert. Étant donnés un texte source et une traduction candidate, seule une personne bilingue, voire connaissant les intentions de l'auteur du texte source, peut véritablement juger de la qualité de la traduction candidate. Les critères de qualité peuvent être multiples et inclure par exemple des critères de correction grammaticale (*fluidité*) et de fidélité au sens du texte (*adéquation*). L'ARPA (Advanced Research Projects Agency, agence pour les projets de recherche) y a ajouté un critère quantifiant l'information effectivement transmise, critère déterminé à l'aide de questions à choix multiples (White, 1995). D'autres critères peuvent intervenir selon la tâche considérée : aide à la traduction, traduction de pages Web, surveillance, etc.

Ces critères de qualité constituent la vraie mesure de l'adéquation du système de traduction à la tâche visée, mais requièrent une intervention humaine, qui est généralement très coûteuse. Par ailleurs, toute évaluation subjective souffre des problèmes de non-reproductibilité et de variabilité inter-annotateur. C'est en particulier le cas des critères de fluidité et d'adéquation cités plus haut, dont l'évaluation sur

une échelle absolue de 1 à 5 est longue et difficile (Callison-Burch *et al.*, 2008). C'est pourquoi plusieurs mesures automatiques ont été développées au fil des années. Leur objectif est d'être corrélées avec les scores que produirait une évaluation manuelle. Ceci est un problème difficile, car une même phrase peut être traduite de nombreuses façons possibles et également acceptables. Les mesures automatiques doivent autoriser les variations légitimes et pénaliser les erreurs (Babych et Hartley, 2004). Les mesures présentées ci-dessous et utilisées dans ce cadre sont parmi les plus courantes dans la communauté de la traduction automatique. Elles nécessitent une ou plusieurs traductions de référence pour chaque phrase source. Dans nos expérimentations, à chaque phrase des corpus d'évaluation correspond une seule traduction de référence. Parmi les différentes mesures d'évaluation, nous utilisons deux mesures appartenant à différentes classes : Une mesure qui calcule le nombre de modifications nécessaires à apporter à la sortie d'un système pour atteindre la référence (TER) et une mesure qui calcule au contraire une ressemblance (BLEU).

Score TER

Le score TER (*Translation Edit Rate*, taux d'édition de la traduction) correspond à la distance d'édition (Snover *et al.*, 2006). Lorsqu'une seule traduction de référence e_r est disponible, le score TER d'une traduction candidate e_c est calculé comme suit :

$$TER(e_c) = \frac{n_{ins} + n_{sup} + n_{sub} + n_{dec}}{|e_r|} \tag{3.3}$$

où n_{ins}, n_{sup}, n_{sub} et n_{dec} sont respectivement les nombres minimums d'insertions, de suppressions de substitutions et de décalage pour modifier e_c en e_r , et $|e_r|$ dénote la taille de e_r en nombre de mots.

Score BLEU

Le score Bleu (*Bilingual Evaluation Understudy*) mesure une ressemblance de la traduction candidate à la traduction de référence (Papineni *et al.*, 2002). Ce score est une moyenne géométrique des précisions n-grammes (p_n, généralement avec n entre 1 et 4), multipliée par une pénalité de brièveté. La précision représente le nombre de n-grammes de l'hypothèse de traduction présents également dans une ou plusieurs références, divisé par le nombre de n-grammes total de l'hypothèse de traduction. pénalité de brièveté est destinée à pénaliser les systèmes qui essaieraient d'augmenter artificiellement leurs scores de précisions en produisant des phrases délibérément

courtes. L'expression du score Bleu est ainsi :

$$Bleu(e_c, e_r) = PB \cdot \exp(\sum_{i=1}^{n} w_n \log p_n) \tag{3.4}$$

où w_n est constant et vaut 1 et la pénalité de brièveté PB est calculée comme suit. Soient $c = |e_c|$ la taille de la traduction candidate et r la taille de la référence la plus proche de c. Alors :

$$PB = \begin{cases} 1 & \text{si } c \geq r \\ e^{1-\frac{r}{c}} & \text{si } c < r \end{cases} \tag{3.5}$$

BLEU étant un score de précision, plus il est élevé, meilleure est la traduction. La bonne corrélation entre Bleu et l'évaluation humaine a été constatée a plusieurs reprises (Papineni *et al.*, 2002). Cette mesure a gagné le statut de mesure automatique de référence au sein de la communauté de la traduction automatique.

3.4.2 Résultats et discussion

La qualité de traduction du système RÉF et des différentes stratégies d'intégration dynamiques et statique est évaluée sur les deux corpus de test sur la base des mesures BLEU et TER. Les résultats de traduction pour les différentes configurations sont rassemblés dans le tableau 3.2. À première vue, nous remarquons que le score BLEU varie en fonction du type du jeu de test. Concernant le corpus de test *Tous_test*, la meilleure amélioration est obtenue par la stratégie dynamique TRAIT. Cette méthode rapporte un faible gain de $+0,1$ points BLEU et augmente légérement le taux d'erreurs $-0,04$ points TER par rapport au système RÉF. Le premier exemple de traduction présenté dans le tableau 3.3 souligne la contribution du trait introduit à l'amélioration de la qualité de traduction. Contrairement à RÉF, traduisant l'EPL « *initiative communautaire* » par simplement le mot simple « *initiative* », la stratégie TRAIT mène à bien la traduction de l'EPL « *initiative communautaire* » par « *community initiative* » et de son contexte immédiat (« *for africa* »). Des scores BLEU plus faibles que ceux rapportés par RÉF sont obtenus par les stratégies TABLE et FORCÉ.

Pour le corpus de test EPL_*Test*, qui ne considère que les phrases contenant des EPL du lexique bilingue, nous constatons que toutes les stratégies d'intégration dy-

	BLEU		TER	
	Tous_Test	EPL_*Test*	*Tous_Test*	EPL_*Test*
RÉF	28,85	30,83	55,44	53,59
Stratégies dynamiques				
TRAIN	**28,87**	**31,06**	**55,38**	**53,32**
TABLE	28,82	**30,88**	55,42	**53,46**
TRAIT	**28,95**	**31,06**	55,48	**53,56**
Stratégie statique				
FORCÉ	28,20	29,19	56,01	55,05

TABLEAU 3.2: Résultats de traduction des corpus de test *Tous_Test* et EPL_*Test* en termes de scores BLEU et TER.

namiques rapportent des scores BLEU plus élevés que ceux obtenus par RÉF et la stratégie statique FORCÉ. Un gain de $+0,23$ points BLEU et $+0,27$ points TER est obtenu par TRAIN. Le même gain en points BLEU a été également rapporté par TRAIT pour un taux d'erreur plus élevé que celui de TRAIN mais toujours au deça de celui de RÉF. La stratégie TABLE rapporte des scores légèrement améliorés montrant un gain de $+0,05$ points BLEU et $+0,13$ points TER. Contrairement aux stratégies d'intégration dynamiques, la stratégie FORCÉ obtient de faibles scores sur les deux corpus de test. Ceci peut être expliqué de la manière suivante : nous supposons qu'en forçant le décodeur à traduire une EPL par une unité donnée, ce dernier échoue à bien traduire le contexte immédiat gauche ou droit de l'EPL induisant ainsi une diminution de la valeur du score BLEU. Ainsi, dans le second exemple du tableau 3.3, les deux systèmes produisent une bonne traduction pour l'EPL « *aide international* » . Cependant, FORCÉ échoue dans la traduction du segment « *relever de* » . Il est important de noter que cette traduction pourrait être soutenue dans le cas où nous associons à chaque phrase source de multiples traductions de référence. Dans une étude antérieure, (Ren *et al.*, 2009) ont proposé une stratégie similaire à la stratégie TRAIT dans laquelle ils indiquent pour chaque entrée de la table de traduction si un segment contient une paire d'EPL bilingue spécialisée. Pour le domaine médical, leur méthode rapporte un gain de $+0,17$ points BLEU par rapport à MOSES, un gain plus faible que celui obtenu par la stratégie TRAIT.

La question que l'on peut se poser en observant les différents résultats obtenus est : est-il possible de prétendre que le système ayant les meilleurs scores est vraiment le meilleur système ? En d'autres termes, les résultats obtenus par différentes stratégies sont-ils *statistiquement significatifs* ?

SRC	*je entendre en effet lancer un initiative communautaire pour le afrique en étendre le ligne nepad ...*
RÉFERENCE	*indeed , i intend to launch a community initiative* **for africa** *, develop the nepad line...*
RÉF	*i hear be indeed launch an* initiative **for the eu africa** *by extend the nepad line ...*
TRAIT	*i hear in fact launch a community initiative* **for africa** *by extend the nepad line ...*
SRC	*le deuxième groupe de problème relever de le aide international et du prochain engagement de johannesburg.*
RÉFERENCE	*another series of problem* **mention be a matter of** international aid *and the forthcoming johannesburg summit.*
RÉF	*the second group of the problem* **be a matter of** *international aid and the forthcoming johannesburg commitment.*
FORCÉ	*the second group of the problem* **relate to the** *international aid and the forthcoming johannesburg commitment.*

TABLEAU 3.3: Exemples de traductions. Notons que le texte est lemmatisé. Nous soulignons les EPL et mettons en gras différentes suggestions pour le contexte immédiat gauche ou droite.

Significativité statistique des résultats

Pour évaluer la significativité statistique des résultats obtenus, nous utilisons la *méthode par ré-échantillonage par amorce* décrite par (Koehn, 2004). Cette méthode estime la probabilité (*p-valeur*) qu'une différence mesurée entre les scores BLEU surgisse par hasard, par la création à plusieurs reprises (10 fois) d'échantillons uniformes avec remise à partir des corpus de test. Nous nous appuyons sur cette méthode pour comparer les méthodes TRAIN, TABLE et TRAIT apportant des gains dans le score BLEU (Tableau 3.2) par rapport à RÉF. Les résultats obtenus sont présentés dans le tableau ci-dessous.

Méthode	*p-valeur* (95 % *IC*)	
	Tous_Test	EPL_*Test*
RÉF	-	-
TRAIN	0,1	0,05
TABLE	-	0,3
TRAIT	0,01	0,01

TABLEAU 3.4: Test de significativité statistique des résultats en termes de *p-valeur*

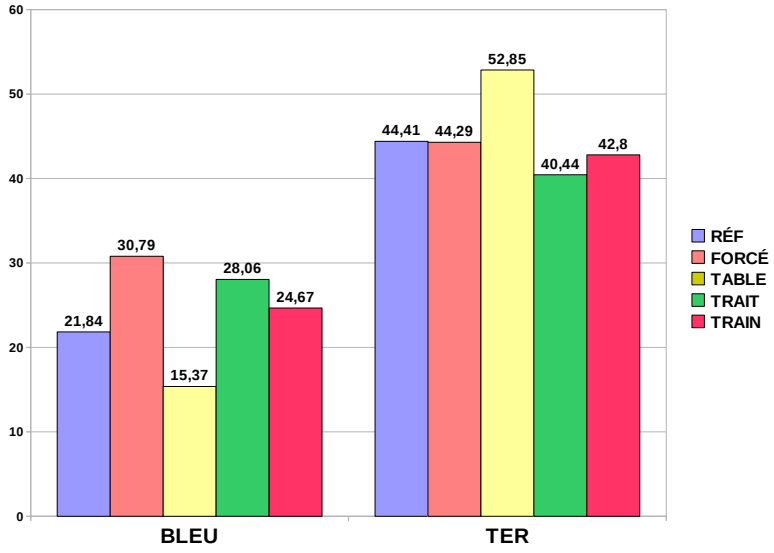


FIGURE 3.3: Évaluation lexicale des EPL en terme de scores BLEU et TER

Sur un intervalle de confiance (IC) de 95%, les résultats varient de non significatifs (quand $p > 0,05$) à hautement significatifs. Sur les deux corpus de test, nous remarquons que les améliorations apportées par la stratégie TRAIT, ayant une p-valeur de $0,05$, sont significatifs. Cependant, le faible gain en score BLEU obtenu par TABLE ($0,3$ de p-valeur) est non significatif. La cause est que nous utilisons la valeur de l'indice de Jaccard, une mesure utilisée pour comparer la similarité et la diversité entre des échantillons, pour définir la probabilité de traduction. Nous émettons l'hypothèse qu'un ajustement par la transformation des valeurs de l'indice de Jaccard obtenus pour chaque paire d'EPL en une probabilité de traduction pourrait assurer l'uniformité et la cohérence des probabilités dans la table de traduction.

Évaluation lexicale des résultats de traduction

Nous avons montré dans l'expérience précédente que les EPL bilingues améliorent significativement les résultats *globaux* de traduction et que la meilleure façon d'incorporer ces unités consiste d'abord à les ajouter dans la table de traduction du système de référence et d'ajouter ensuite un trait indiquant pour chaque entrée s'il s'agit d'une EPL ou pas (stratégie TRAIT). Le score BLEU est conçu pour relèver que les améliorations globales dans la traduction d'une ou plusieurs phrases. Dans certains cas, évaluer la qualité de traduction d'un segment individuel est aussi intéressant. Par exemple, lorsqu'il s'agit de traduire des textes contenant des EPL, il est intéressant

d'évaluer la qualité de traduction de ces unités. Ceci permet de mener une sorte d'évaluation lexicale fine des différentes hypothèses de traduction proposées pour les EPL par différents systèmes.

Dans ce cadre, nous évaluons les résultats de traductions d'un corpus source d'EPL, obtenues par le système de référence Moses (RÉF) et nos différentes stratégies d'intégration (TRAIN, TABLE, TRAIT et FORCÉ). Les EPL constituant la traduction de référence sont extraites manuellement du corpus EPL_*Test*. Ce corpus a été ensuite traduit par RÉF, TRAIN, TABLE, TRAIT et FORCÉ. Les résultats obtenus évalués par les mesures BLEU et TER sont présentés dans la figure 3.3. Le résultat le plus marquant est celui obtenu par la stratégie FORCÉ qui dans les expériences précédentes n'affiche aucune amélioration significative. Cette stratégie rapporte des gains respectifs de $+9,8$ et de $+0,12$ en point BLEU et TER. Ce constat vient confirmer que l'obtention d'un faible score BLEU dans les premières expériences, où nous évaluons la qualité globale de la traduction, n'est pas due à une mauvaise qualité du lexique bilingue d'EPL.

Nous constatons également que la stratégie TRAIT obtient des résultats qui dépassent de loin ceux de RÉF. Cette stratégie fait passer le score BLEU de $21,84$ à $28,06$. Plus intéressant encore, TRAIT obtient le score TER le plus faible et rapporte un gain de 4 points. La traduction des EPL par TRAIN améliore aussi les résultats de RÉF. Cette stratégie augmente le score BLEU d'environ 3 points. En conclusion, le lexique bilingue d'EPL affecte positivement les résultats de traduction du système de référence Moses.

3.5 Conclusion

Nous avons étudié dans ce chapitre l'apport d'un lexique bilingue d'EPL, construit automatiquement, pour un système de traduction statistique. Nous avons présenté trois stratégies d'intégration dynamiques où nous avons modifié le modèle de traduction de différentes façons pour une prise en considération des EPL bilingues et une stratégie d'intégration statique dans laquelle nous avons incorporé ces unités sans changer le modèle de traduction. Les expériences menées dans ce cadre montrent que la stratégie dynamique TRAIT, où un trait additionnel indique pour chaque entrée de la table de traduction s'il s'agit d'une EPL ou pas, améliore significativement

les résultats obtenus par Moses avec un gain allant jusqu'à $+0,23$ points BLEU. Une évaluation lexicale fine des résultats de traduction a aussi été mise en place. Cette évaluation a montré que la plupart des stratégies d'intégration améliorent d'une manière significative les résultats de traduction et que la qualité du lexique construit automatiquement est suffisamment bonne pour améliorer la qualité de traduction statistique.

L'évaluation faite en considérant une tâche de traduction automatique avec un système statistique montre une amélioration qui reste petite, dont on ne sait pas si elle resterait présente avec un corpus beaucoup plus grand. Un système de traduction à base de règles pourrait mieux profiter de l'extension de son lexique bilingue, dans la mesure où un système de ce type repose beaucoup plus sur son lexique. Or, à notre connaissance, il n'existe à ce jour aucun système de traduction à base de règles libre.

Deuxième partie

Extraction de lexiques bilingues : Vers l'exploitation de corpus comparables

Introduction générale

L'extraction de lexiques bilingues à partir de corpus parallèles est devenue de nos jours un classique en extraction terminologique et est une activité qui est passée du domaine de la recherche au domaine commercial (Déjean et Gaussier, 2002). L'extraction de lexiques bilingues à partir de corpus comparables est plus récente et reste confinée au domaine de la recherche. L'utilisation de ces corpus vient en outre pallier les insuffisances des corpus parallèles du fait que :

- les corpus parallèles sont par nature des ressources rares, et s'ils existent leur couverture pour des domaines de spécialité demeure très faible. En raison de ces limitations, de nombreux domaines génériques et spécifiques ne sont pas facilement accessibles.
- La plupart des corpus parallèles sont disponibles pour les langues riches en ressources comme l'anglais, le français, l'espagnol, etc. Pour des langues peu dotées de ressources comme le roumain, les problèmes d'acquisition de corpus parallèles sont plus problématiques.
- L'extraction de lexiques bilingues ne se base pas que sur l'utilisation des corpus parallèles. L'utilisation de corpus comparables de bonne qualité et de connaissances linguistiques supplémentaires peuvent être aussi efficaces que les corpus parallèles de petite taille (Koehn et Knight, 2000).

L'avantage principal des corpus comparables est leur disponibilité dans différentes langues et différents domaines de spécialité. C'est pour ces raisons que l'extraction de lexiques bilingues à partir de corpus comparables a attiré notre attention. Cette tâche fera l'objet de la troisième partie de ce manuscrit, où nous présenterons de nouvelles méthodes de création de lexiques bilingues.

Cette partie est organisée comme suit : dans le chapitre 4 nous présentons le contexte global de la tâche d'extraction lexicale à partir de corpus comparables et

introduisons les ressources linguistiques utilisées, notamment les corpus comparables sur lesquels nous avons réalisé nos expériences. Le chapitre 5 est centré sur l'approche d'extraction de lexiques bilingues spécialisés à partir de corpus comparables étendant l'approche distributionnelle par l'utilisation de la désambiguïsation sémantique. Enfin, nous décrivons dans le chapitre 6 l'approche utilisant l'analyse sémantique explicite pour extraire des lexiques bilingues.

Chapitre 4

Contexte et Matériel

4.1 Introduction

Depuis les années 90, l'extraction de lexique bilingues à partir de corpus comparables a attirée l'attention de plusieurs chercheurs (Rapp, 1995; Fung et Yee, 1998; Chiao et Zweigenbaum, 2002; Gamallo, 2007; Kun et Tsujii, 2009). Cette tâche est venue pallier les insuffisances des corpus parallèles citées plus haut. Cependant, l'alignement lexical à partir de ce type de corpus demeure une opération délicate du fait qu'il n'est plus possible de s'appuyer sur la distribution des mots dans le document, une caractéristique de base dans les approches utilisant les corpus parallèles. Face à cette difficulté et comme il a été mentionné dans le chapitre 1, les différentes approches liées à l'exploitation de corpus comparables héritent de la sémantique distributionnelle de (Harris, 1954) et reposent sur la simple observation que si dans une langue source deux mots cooccurrent plus souvent que par hasard, alors dans un texte de langue cible, leurs traductions doivent également cooccurrer plus souvent. La mise en œuvre de différentes approches reposant sur cette observation est conditionnée par la disponibilité de plusieurs ressources linguistiques, notamment les corpus comparables, les dictionnaires bilingues, les listes de références, différents paramètres expérimentaux pour la définition des environnements lexicaux et le protocole d'évaluation des résultats d'extraction.

Ce chapitre est destiné à la présentation de l'ensemble des ressources linguistiques et statistiques nécessaires à l'évaluation des différentes approches que nous présenterons dans les chapitres 5 et 6. Tout d'abord, nous présentons dans la section 4.2 les corpus comparables sur lesquels porte notre étude. La section 4.3 est consacrée à la description des dictionnaires bilingues servant de pont entre la langue source et la langue cible. Nous décrivons ensuite dans la section 4.4 les listes de paires de traductions de référence pour l'évaluation de différentes approches d'extraction lexicale. Dans la section 4.5, différents paramètres expérimentaux sont fixés pour le reste du manuscrit. Enfin, nous présentons le protocole d'évaluation dans la section 4.6.

4.2 Corpus Comparables

First, catch your corpus.
(Somers, 2001)

Les corpus de textes constituent une ressource primordiale pour toute tâche d'extraction de connaissances à partir de textes. En extraction lexicale à partir de corpus comparables, ces corpus sont utilisés pour fournir à la fois les informations lexicales et statistiques suivantes :

- *L'information contextuelle :* l'information sur les environnements lexicaux d'un terme source ou cible ainsi que leur fréquences.
- *Les listes de référence :* il s'agit des paires de traductions initialement sélectionnées pour évaluer un modèle d'extraction.

Dans cette étude, l'extraction lexicale porte sur des corpus comparables spécialisés dans les domaines de la *finance des entreprise, cancer du sein, énergie éolienne* et de la *technologie mobile*. Nous nous intéressons également à étudier différentes approches pour les paires de langues *français-anglais* et *roumain-anglais*. À cet égard, deux sources de données ont été utilisées pour extraire les corpus comparables. D'une part, nous nous basons sur Wikipédia[1] et proposons une technique permettant de construire des corpus comparables spécialisés. D'autre part, nous exploitons des corpus comparables disponibles sur le Web. Ceci permet d'étudier le comportements de

1. http://dumps.wikimedia.org/

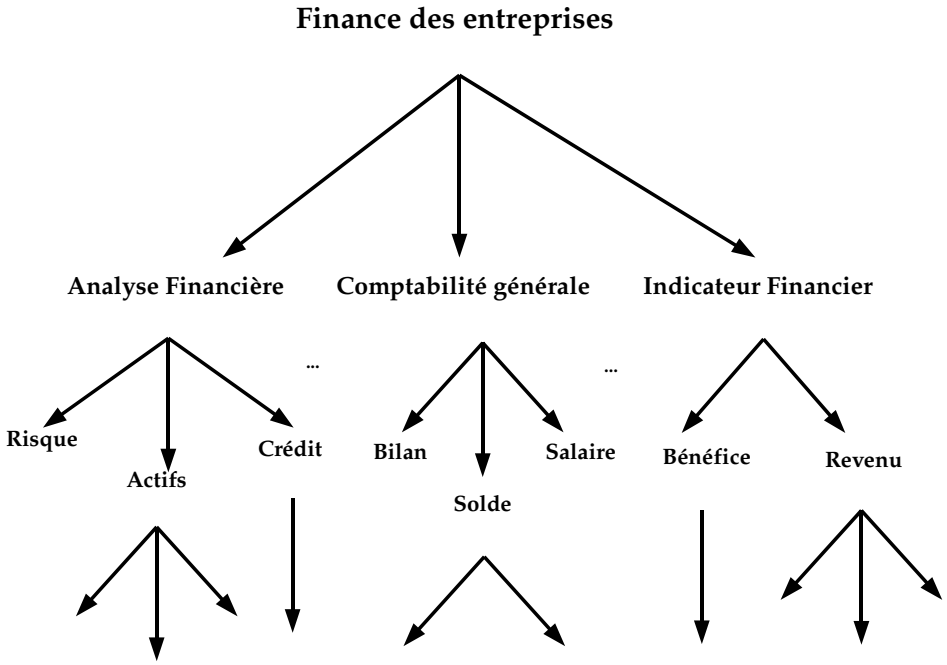


FIGURE 4.1: Arborescence de catégories de la thématique *Finance des entreprises*

différentes approches d'extraction de lexiques bilingues à partir de corpus comparables issus de différentes sources et traitant différentes thématiques. Cette section est alors consacrée à la description de ces deux sources de données.

4.2.1 Wikipédia comme corpus comparable

Wikipédia est une encyclopédie collective libre qui fournit de vastes collections de textes dans plusieurs langues et qui couvre aussi bien des domaines généraux et domaines de spécialité. Elle est composée d'articles qui font la synthèse des connaissances sur un sujet donné. Ce sujet se défini par un titre, un texte le décrivant et la thématique à laquelle il appartient (catégorie dans Wikipédia). Cette ressource a été largement exploitée pour extraire des corpus comparables. Nous distinguons les travaux qui partent d'une liste de mots amorces pour extraire les articles de Wikipédia qui lui correspondent. Dans (Laroche et Langlais, 2010), les auteurs utilisent les listes de référence comme amorce. Ceci permet de garder les articles où les mots sources et cibles apparaissent forcément. Citons également les les travaux qui se basent sur

les liens au sein de la même langue afin de chercher l'information dans la langue source et ainsi parcourir le corpus monolingue (Sadat et Terrasa, 2010) et sur les liens interlingues (Sadat et Terrasa, 2010; Rapp *et al.*, 2012) pour capturer l'information translinguistique. D'autres approches dont notamment (Gamallo et Garcia, 2012) se basent sur le système de classement thématique de Wikipédia par catégorie pour extraire les articles traitant une thématique donnée.

Nous proposons une technique d'extraction de corpus comparables qui relève des deux derniers types d'approches. Cette technique se base en premier lieu sur les catégories de Wikipédia pour extraire des articles spécialisés en langue source. Ensuite, les liens interlingues sont utilisés pour chercher l'information translinguistique et construire la partie cible du corpus comparable. Nous considérons que le domaine d'étude constitue une *catégorie* dans Wikipédia. Une requête composée du domaine d'étude en langue source (par exemple *finance des entreprises*) est donc construite pour extraire une arborescence de catégories ou de thèmes ayant pour catégorie mère le domaine de spécialité. Un exemple d'arborescence est présenté dans la figure 4.1.

Ensuite, nous collectons tous les articles associés à chacune des catégories de l'arborescence pour construire un corpus spécialisé monolingue (en langue source). Pour collecter les articles en langue cible, les *liens interlingues* au sein de chaque article du corpus monolingue sont ensuite utilisés pour extraire les articles correspondants en langue cible. Les articles Wikipédia collectés ont été convertis en texte brut et nettoyés. Sur la base de cette technique, nous avons construit :

- quatre corpus comparables **roumain-anglais**, spécialisés dans les domaines sur lesquels porte notre étude (*finance des entreprises, cancer du sein, énergie éolienne* et *technologie mobile*).
- deux corpus comparables **français-anglais** relevant des domaines de la *finance des entreprises* et du *cancer du sein*.

Il est intéressant que noter que la limite principale de cette technique est qu'elle ne permet d'extraire que des corpus comparables pour des paires de langues et domaines de spécialité couverts par Wikipédia.

4.2.2 Corpus du projet TTC

La disponibilité des corpus du projet européen TTC [2] nous a permis de dériver les corpus comparables français-anglais des domaines de *l'énergie éolienne* et celui de la *technologie mobile*. Ces corpus ont été construits à l'aide du crawler *Babouk* (de Groc, 2011). L'objectif principal de ce crawler est de rapatrier des documents pertinents pour un domaine défini. Il requiert simplement un ensemble de termes ou URL amorces en entrée et s'appuie sur un *catégoriseur* basé sur un lexique pondéré pour ordonner les documents à télécharger par ordre de pertinence. Un seuil pour la catégorisation est fixé automatiquement pour filtrer les documents non pertinents.

Ce crawler a permis la constitution de corpus comparables spécialisés dans les domaines de l'*énergie éolienne* et de la *technologie mobile* pour les différentes langues européennes (français, anglais, allemand, espagnol, italien, etc.). Dans notre étude, nous n'utilisons que les corpus français et anglais.

4.2.3 Normalisation des corpus

L'ensemble des corpus comparables ont été normalisés à travers les étapes de pré-traitement linguistique suivantes : la segmentation de mots, l'étiquetage morpho-syntaxique et la lemmatisation à l'ensemble des corpus. Pour le français et l'anglais, la normalisation a été effectuée à l'aide de l'étiqueteur morphosyntaxique TreeTagger (Schmid, 1995). En ce qui concerne la langue roumaine, nous utilisons l'étiqueteur décrit dans (Simionescu, 2011). Nous avons écarté les mots fonctionnels et n'avons gardé que les *noms, adjectifs, verbes* et *adverbes*. Des liste de mots vides prédéfinies ont été également utilisé à cet effet. Ces listes sont constituées de mots non significatifs figurants dans un texte. La signification d'un mot s'évalue à partir de sa distribution dans une collection de textes. Un mot dont la distribution est uniforme sur les textes de la collection est dit vide. En d'autres termes, un mot qui apparaît avec une fréquence semblable dans chacun des textes de la collection n'est pas discriminant, ne permet pas de distinguer les textes les uns par rapport aux autres. Elles sont constituées de respectivement 238, 426 et 264 mots en anglais, français et roumain extraits à partir du Web [3].

2. http://www.ttc-project.eu/index.php/releases-publications
3. https://sites.google.com/site/kevinbouge/stopwords-lists

Domaine	Français	Anglais
Finance des entreprises	402 486	756 840
Cancer du sein	396 524	524 805
Énergie éolienne	145 019	345 607
Technologie mobile	97 689	144 168
Domaine	**Roumain**	**Anglais**
Finance des entreprises	206 169	524 805
Cancer du sein	22 539	322 507
Énergie éolienne	121 118	298 165
Technologie mobile	200 670	124 149

TABLEAU 4.1: Taille des corpus comparables en nombre de *mots pleins*.

Ainsi, huit corpus comparables traitant différentes thématiques pour deux paires de langues ont été crées. Dans le tableau 4.1, nous présentons la taille des corpus résultants. Nous définissons la taille par le nombre de mots pleins composant les corpus monolingues spécialisés. La taille des corpus spécialisés varie au sein et entre différentes langues. Les corpus issus du domaine de la *finance des entreprise* sont les plus riches dans les deux paires de langues. Pour le roumain, le corpus relevant du domaine du *Cancer du sein* est de taille particulièrement réduite, avec approximativement 22 000 mots. Cette variabilité permettra de vérifier s'il existe une corrélation entre la taille du corpus et les résultats obtenus.

4.3 Dictionnaires bilingues

Un dictionnaire bilingue est généralement utilisé dans les différentes approches d'extraction de lexiques bilingues à partir de corpus comparables. Ce dictionnaire sert de pont entre les langues sources et cibles. Dans les approches contextuelles comme l'approche standard, ce type de dictionnaire est utilisée pour traduire les vecteurs de contextes. Dans le cadre de notre étude, nous utilisons un dictionnaire bilingue pour chaque paire de langues. Pour la paire de langues **français-anglais**, nous utilisons le dictionnaire bilingue SCI-FRAN-EurADic [4], comportant 243 539 paires de mots français-anglais issues du domaine général. Ce dictionnaire a été révisé manuellement pour garder 234 355 entrées dont environ 120 000 entrées sont des mots simples.

4. http://catalog.elra.info/product_info.php?products_id=666

Pour la paire des langues **roumain-anglais**, le dictionnaire bilingue utilisé est construit à partir de Wikipédia. Un graphe de traduction des titre d'articles de Wikipédia est tout d'abord extrait à partir des liens interlingues explicitement disponibles dans les articles de Wikipédia. Ce graphe est exploité pour relier l'espace conceptuel d'un mot dans la langue source à l'espace lui correspondant dans la langue cible. La taille résultante de l'espace conceptuel en langue cible est généralement plus petite que celle de la langue source, du fait que seule une partie des articles sources possède des traductions. Ce graphe de traduction est ensuite transformé en un dictionnaire bilingue en supprimant les marques d'homonymie des titres ambigus, ainsi que les listes, les catégories et d'autres pages d'administration. En outre, comme notre intérêt ne se porte que sur des mots simples, nous ne retenons que les paires de traduction composées d'unigrammes dans les deux langues. Les redirections des unigrammes sont également ajoutées lorsqu'elles existent. Le dictionnaire résultant est composé de 136 681 entrées. Nous considérons que ce dictionnaire est incomplet puisque : (1) les titres de Wikipédia sont le plus souvent constitués de noms, (2) les termes issus de domaines spécialisés n'ont souvent pas d'entrée dans Wikipédia et (3) le graphe de traduction ne couvre qu'une partie des mots disponibles dans une langue. Cependant, ce type de dictionnaire est particulièrement utile lorsqu'une paire de langues comme la paire roumain-anglais est couverte que par peu de ressources dictionnariques.

4.4 Listes de références

En extraction lexicale à partir de corpus comparables, rien ne garantit qu'un élément dont on cherche la traduction apparaisse effectivement dans le corpus cible. C'est l'une des raisons pour lesquelles une liste de référence est nécessaire pour évaluer la qualité des résultats d'alignement. En outre, cette liste est indispensable pour avoir une référence de ce que l'on considère comme une traduction correcte. Habituellement, la taille de cette liste tourne autour de 100 *termes* (Hazem et Morin, 2012a; Chiao et Zweigenbaum, 2002). Il convient à ce stade de faire la différence entre les termes *terme* et *mot*. Selon (Sager, 1990), les termes constituent la manifestation linguistique d'objets réels ou immatériels qui partagent des propriétés communes. (Daille, 2002) considère qu'un terme se définit par un nom ou un groupe nominal utilisé dans un

Domaine	Français-Anglais	Roumain-Anglais
Finance des entreprises	125	69
Cancer du sein	96	38
Énergie éolienne	89	38
Technologie mobile	142	94

TABLEAU 4.2: Taille des listes de référence par domaine et paire de langues.

contexte précis pour un sens précis. Une étude complète portant sur l'extraction terminologique a été présentée dans (L'Homme, 2004), où cette définition est donnée :

> « *La particularité du terme par rapport aux autres unités lexicales d'une langue, et d'avoir un sens spécialisé, c'est à dire un sens qui peut être mis en rapport avec un domaine de spécialité.*»

Sans entrer plus en détails dans le définition de *terme*, nous retiendrons cette dernière définition qui semble être la plus en adéquation avec le contexte de notre étude. Par ailleurs, la littérature distingue les termes simples, composés d'un seul mot, des termes complexes, c'est-à-dire des syntagmes nominaux constitués d'au moins deux unités lexicales pleines. Dans le reste du manuscrit, (1) nous employons *terme* lorsque nous nous référerons à un vocabulaire spécialisé et *mot* dans le cas général et (2) nous nous intéressons uniquement à l'extraction de *termes simples*.

Des listes de référence ont été crée pour les quatre domaines et pour les deux paires de langues :

- **Français-anglais** : pour le domaine de la *finance des entreprise* cette liste est extraite à partir du *glossaire bilingue de la micro-finance*[5]. En ce qui concerne le domaine du *cancer du sein*, les termes sont issus du thésaurus *UMLS*[6]. Dans le cas des domaines de l'*Énergie éolienne* et la *Technologie mobile*, ces listes sont extraites à partir de glossaires spécialisés trouvés sur le Web.
- **Roumain-anglais** : les listes de référence pour cette paire de langues sont toutes construites par un locuteur natif de la langue roumaine à partir des listes construites pour le français-anglais, avec l'aide d'outils de traduction automatique comme Google Translate. Dans le cas où plusieurs traductions sont proposées, le locuteur ne garde que la ou les traductions qui sont susceptibles de faire partie du vocabulaire du domaine de spécialité.

5. http://www.microfinance.lu/la-microfinance-cest-quoi/glossaire.html
6. http://www.nlm.nih.gov/research/umls/

Les tailles des différentes listes de référence obtenues sont présentées au tableau 4.2. Notons que les termes présents dans les listes de référence apparaissent au moins cinq fois dans chaque partie des corpus comparables.

4.5 Paramètres expérimentaux

En plus de ceux déjà mentionnés précédemment, trois autres paramètres doivent être fixés : la taille de la fenêtre contextuelle, la mesure d'association et la mesure de similarité.

4.5.1 Fenêtre contextuelle

Dans les approches contextuelles comme l'approche standard, la fenêtre contextuelle est utilisée pour déterminer combien de mots apparaissant autour du terme à traduire sont considérés comme pertinents pour sa représentation. Dans la littérature, la taille de la fenêtre contextuelle varie, allant d'une fenêtre de 2 mots (Rapp, 1999b; Diab et Finch, 2000; Gaussier *et al.*, 2004) à une fenêtre de taille 25 (Prochasson *et al.*, 2009). (Fung, 1998) propose que plus la taille de le fenêtre est large meilleurs sont les résultats d'extraction lexicale tandis que (Gaussier *et al.*, 2004) insistent pour qu'une fenêtre contextuelle de taille 2 soit utilisée car celle ci permet d'atténuer le bruit dans l'espace multidimensionnel du terme à traduire. Plusieurs travaux (Laroche et Langlais, 2010; Morin et Prochasson, 2011; Hazem et Morin, 2012a) ont également étudié l'impact de ce paramètre en le faisant varier et ont constaté que les meilleurs résultats sont obtenus lorsque la taille de la fenêtre est fixée à 7, partant de l'idée qu'elle approxime les dépendances syntaxiques. Dans notre étude, nous suivons ces travaux et fixons la taille de la fenêtre contextuelle à 7 (3 mots à droite et 3 mots à gauche du terme à traduire).

4.5.2 Mesure d'association

Le nombre de mots dans les vecteurs de contexte n'est pas le seul paramètre nécessaire à la représentation des termes à traduire, leur pondération constitue aussi

	i	$\neg i$	
j	O_{11}	O_{12}	
$\neg j$	O_{21}	O_{22}	

TABLEAU 4.3: Table de contingence pour deux mots i et j.

un paramètre clé (Fung et Yee, 1998). Par conséquent, une pondération des mots dans les vecteurs de contexte peut être préférable au simple dénombrement de cooccurrences. Selon (Rapp, 1995), cette pondération permet de renforcer la corrélation entre les cooccurrences des mots.

Pour calculer le degré ou la force d'association entre deux unités lexicales, une multitude de mesures d'association ont été décrites dans la littérature. (Fung et Yee, 1998) utilisent la mesure IDF plutôt que la fréquence. Contrairement à la fréquence, cette mesure réduit l'importance des mots les plus fréquents, par la prise en compte du contexte global (la totalité du corpus). Toutefois, la mesure IDF est souvent utilisée pour identifier des mots vides plutôt que de trouver des mots très liés. Assez souvent, le rapport de vraisemblance logarithmique (LLR, *Log Likelihood Ratio*) est utilisé pour trouver les mots de contexte les plus associés aux termes à traduire (Rapp, 1999b; Prochasson *et al.*, 2009; Morin et Prochasson, 2011; Hazem et Morin, 2012a). Le LLR mesure le ratio des vraisemblances de deux configurations correspondants aux hypothèses à confronter : que les deux mots cooccurrent et qu'ils apparaissent de manière indépendante. Nous relevons également la mesure d'information mutuelle (Morin et Daille, 2006) ou encore l'information mutuelle ponctuelle (PMI, *Pointwise Mutual Information*) (Shezaf et Rappoport, 2010; Andrade *et al.*, 2010). Le PMI utilise les fréquences relatives et représente la quantité d'information partagée par deux variables. Il est à noter que contrairement à l'information mutuelle, cette mesure favorise les liens entre les mots de faibles fréquence.

Récemment, une étude complète des différentes mesures d'association a été présentée dans (Laroche et Langlais, 2010). Pour le domaine médical, les meilleurs résultats sont obtenus en utilisant le rapport des chances ajusté (*Discounted Odds Ratio*). Dans notre étude, nous suivons (Laroche et Langlais, 2010) et utilisons le *rapport des chance ajusté* comme mesure d'association. La raison étant que, pour le domaine du *cancer du sein*, les ressources linguistiques utilisées dans cette études

Terme	Mot du contexte	OddsRatio_{disc}
liquidité	financier	15.626
	économique	14.754
	entreprise	14.605
	banque	14.371
	intérêt	14.078
	bancaire	12.314
	échange	12.211
	investisseur	10.807
	devise	10.059
	refinancement	8.480

TABLEAU 4.4: Force d'association entre le terme français *liquidité* et dix éléments de son vecteur de contexte.

sont similaires aux nôtres. Cette mesure est définie dans l'équation suivante :

$$OddsRatio_{disc} = \log \frac{(O_{11} + \frac{1}{2})(O_{22} + \frac{1}{2})}{(O_{12} + \frac{1}{2})(O_{21} + \frac{1}{2})} \tag{4.1}$$

Le rapport des chances ajustée se base sur une table de contingence (table 4.3), regroupant les fréquences d'observations de deux mots dans une fenêtre donnée. O_{11} représente le nombre de fois où i et j cooccurrent, O_{12} est le nombre d'occurrences de j sans i, O_{21} correspond au nombre d'occurrences de i sans j, et O_{22} est le nombre d'occurrences de deux mots autres que i et j. Puisqu'il s'agit d'une version ajustée du rapport des chances, le $+\frac{1}{2}$ n'est autre que le facteur de lissage ou d'ajustement sur les mots rares.

Nous mesurons ainsi pour chaque élément d'un vecteur de contexte son association par rapport au terme à traduire (pivot). Si un mot apparaît très fréquemment dans le corpus, son association avec le terme pivot sera faible, puisqu'elle sera affecté par le nombre de fois ou il apparaît sans celui-ci. En revanche si deux unités sont rares dans le corpus mais cooccurrent fréquemment, leur valeur d'association sera élevée. L'utilisation de cette mesure permet à la fois d'évaluer la corrélation entre le terme à traduire et les éléments de son vecteur de contexte et de normaliser les poids des éléments des vecteurs de contexte entre le corpus source et le corpus cible dans le cas où la fréquence des unités ne sont pas comparables d'un corpus à l'autre. Le tableau

4.4 décrit les valeurs du $OddsRatio_{disc}$ obtenues pour dix éléments du vecteurs de contexte du terme français *liquidité*.

4.5.3 Mesure de similarité

La mesure de similarité constitue une composante essentielle en extraction de lexiques bilingues à partir de corpus comparables. Cette mesure est utilisée pour évaluer la ressemblance entre un terme en langue source et un terme en langue cible. Différentes mesures de similarité ont été utilisées dans des études antérieures. La plus utilisée est le cosinus (Fung et Yee, 1998; Chiao et Zweigenbaum, 2003; Gaussier *et al.*, 2004; Prochasson *et al.*, 2009), mais de nombreux auteurs ont étudié des métriques alternatives comme la distance du Jaccard pondérée (Chiao et Zweigenbaum, 2002; Hazem et Morin, 2012a) ou encore la distance de Manhattan (Rapp, 1999b). Dans notre cadre, nous nous sommes basés comme la plupart des travaux précédents sur la mesure du *cosinus*. Le cosinus de l'angle formé par deux vecteurs source v_s et cible v_c est défini dans l'équation 4.2.

$$Cos(v_s, v_c) = \frac{v_s \cdot v_c}{|v_s| \cdot |v_c|} \tag{4.2}$$

Le cosinus de l'angle s'obtient par le produit scalaire divisé par le produit des normes des vecteurs source et cible. Cette mesure retourne des valeurs dans un intervalle de [0,1]. Une valeur de 0 indique que les deux vecteurs sont indépendants (c'est-à-dire orthogonaux – ce qui peut signifier qu'ils n'ont aucun critère en commun) et une valeur de 1 indique que les deux vecteurs sont identiques.

4.6 Paramètres d'évaluation

Comme dans la plupart des applications du TAL, une mesure d'évaluation est nécessaire à l'évaluation des modèles d'extraction de lexiques bilingues. Les méthodes d'évaluation actuelles varient, mais la plupart d'entre elles ont originellement été introduites pour évaluer les systèmes de recherche d'information comme le rappel et la précision. En extraction lexicale, le modèle d'alignement renvoie une liste ordonnée

de candidats à la traduction pour chaque terme à traduire, classée en fonction de la similarité entre les vecteurs des candidats et celui du terme à traduire. Les résultats sont évalués à partir d'une liste de traductions de référence, en comptant le nombre de candidats corrects trouvés dans les N premiers candidats renvoyés pour chaque terme à traduire, désignés dans la campagne d'évaluation TREC par le terme *succès* à N ($succès_N$) (Tomlinson, 2012).

Dans ce mémoire, nous évaluons les résultats d'extraction lexicale en utilisant la F-mesure (ou F_1 score) au $succès_1$ et $succès_{20}$ et la MAP (*Mean Average Precision*) (Manning *et al.*, 2008) au $succès_{20}$. La F-mesure combine la précision et le rappel et leur pondération pour obtenir une valeur indiquant la performance globale du modèle d'alignement. Elle se définit comme suit :

$$F_1 = \frac{2 * (Précision * Rappel)}{Précision + Rappel} \tag{4.3}$$

où, la précision indique le nombre de traductions correctes divisé par le nombre de termes pour lesquels le système propose au moins une traduction. Le rappel est égal au rapport entre les traductions correctes et le nombre total des termes.

Comme souligné dans (Gaussier *et al.*, 2004), il conviendrait de considérer aussi le rang de la bonne traduction comme étant une caractéristique importante lors de l'évaluation du modèle d'alignement. Cette caractéristique n'est prise en compte que par la mesure MAP. Cette mesure représente la qualité d'un système en fonction de différents niveaux de rappel :

$$MAP(Q) = \frac{1}{Q} \sum_{|Q|}^{j=1} \frac{1}{m_j} \sum_{m_j}^{k=1} Précision(R_{jk}) \tag{4.4}$$

où Q constitue le nombre de termes à traduire, m_j est le nombre de traductions de référence pour le $j^{ème}$ terme et $Précision(R_{jk})$ est égale à 0 si la traduction de référence n'est pas trouvée pour le $j^{ème}$ terme ou $\frac{1}{r}$ s'il y figure (r est le rang de la traduction de référence dans les traductions candidates).

4.7 Conclusion

Dans ce chapitre, nous avons présenté les ressources linguistiques et statistiques requises dans la tâche d'extraction de lexiques bilingues à partir de corpus comparables. Nous nous concentrons sur l'exploitation de huit corpus comparables, conçus pour deux paires de langues (français-anglais et roumain-anglais) et couvrant quatre domaines de spécialité : *finance des entreprises, cancer du sein, énergie éolienne* et *technologie mobile*. Les textes reflétant ces domaines tendent à employer une terminologie particulière, rarement disponible dans des dictionnaires de langue générale. En utilisant ces corpus, nous espérons améliorer l'acquisition automatique des paires de traductions par l'amélioration de leurs représentation contextuelle.

Le chapitre suivant est consacré à l'introduction d'une nouvelle approche qui vise à améliorer les résultats de l'approche standard utilisée pour l'extraction de lexiques bilingues à partir de corpus comparables spécialisés. Cette approche tente de résoudre un problème très peu abordé dans la littérature, le problème de l'ambiguïté des mots. Nous démontrons expérimentalement qu'avec les ressources linguistiques et statistiques décrites dans ce chapitre, cette méthode améliore les résultats des approches de l'état de l'art, plus particulièrement lorsque plusieurs mots du contexte sont ambigus.

Chapitre 5

Désambiguïsation lexicale des vecteurs de contexte

5.1 Introduction

La caractérisation du contexte des mots constitue le cœur de la plupart des méthodes d'extraction de lexiques bilingues à partir de corpus comparables. Nous présentons dans ce chapitre une approche dont l'objectif est de représenter aux mieux les contextes des mots à traduire. Cette approche aborde et vise à résoudre le problème de la *polysémie* des mots dans les vecteurs de contexte, un problème non traité dans l'approche standard utilisée pour l'extraction de lexiques bilingues à partir de corpus comparables. Dans un cadre multilingue, nous nous intéressons à une configuration particulière et fréquente de la polysémie : le cas d'une unité lexicale possédant plusieurs sens et emplois qui se traduisent de façon distincte.

Dans l'approche standard, ce problème intervient avec l'utilisation du dictionnaire bilingue amorce, considéré comme un pilier pour la traduction des vecteurs de contexte de la langue source vers la langue cible. Considérons par exemple le cas présenté dans la figure 5.1. Le terme français *"action"* possédant trois emplois se traduit en anglais par les termes distincts *"action, share, stock, lawsuit, deed"* et *"act"*. Ces traductions comprennent aussi bien des termes de sens différents comme *lawsuit* et *share* que des traductions synonymes comme *share* et *stock*. Or, lorsque l'étude porte sur un

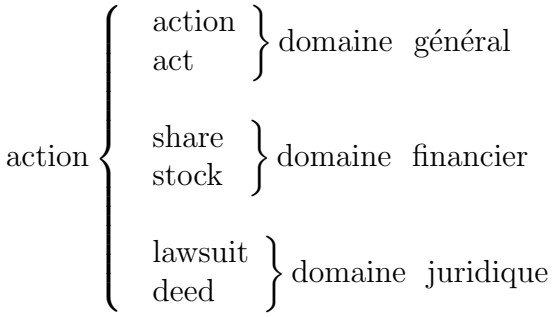


FIGURE 5.1: Exemple de terme dont les traductions sont polysémiques.

domaine de spécialité, ces traductions ne sont pas toutes pertinentes. Par exemple, dans le domaine de la *finance*, la prise en compte des termes *"lawsuit"* et *"deed"*, issus du domaine *juridique*, ne feront qu'introduire du bruit dans les vecteurs de contexte.

Dans de tels cas, il est difficile d'évaluer dans des ressources "plates" comme les dictionnaires bilingues quelles traductions sont les plus appropriées, car elles sont le plus souvent non ordonnées. De fait, peu d'approches ont pris en compte ce phénomène. L'approche standard présentée dans (Déjean *et al.*, 2002) prend en compte toutes les traductions disponibles et les conserve avec la même importance dans le vecteur traduit. En revanche, (Fung, 1998) propose de considérer les entrées des dictionnaires par ordre décroissant d'apparition. La première traduction proposée aura un poids plus important que la seconde. Cette démarche exige la disponibilité d'un dictionnaire bilingue probabiliste dans lequel les entrées sont classées par ordre d'importance. Récemment, (Morin et Prochasson, 2011) ont présenté une méthode dans laquelle si plusieurs traductions sont disponibles, elles sont prises en compte en fonction de leur fréquence dans le corpus cible. Leur méthode permet de faire ressortir les traductions les plus pertinentes pour un contexte donné.

Dans cette étude, nous proposons d'augmenter l'approche standard par un processus de désambiguïsation sémantique. L'intuition qui sous-tend cette approche est que, pour chaque mot polysémique du vecteur de contexte, au lieu de considérer toutes les traductions proposées par le dictionnaire bilingue, nous n'utilisons que les traductions susceptibles de donner la meilleure représentation du vecteur de contexte en langue cible. À notre connaissance, c'est une première application de la désambiguïsation sémantique en extraction de lexiques bilingues à partir de corpus comparables. Nous présentons dans un premier temps le processus de désambiguïsation proposé, qui re-

pose sur différentes mesures de similarité sémantique dérivées de la base de données lexicale WordNet (Fellbaum, 1998). Nous testons ensuite cette approche sur quatre corpus comparables spécialisés dans les domaines mentionnés au chapitre 4 (*finance des entreprise, cancer du sein, énergie éolienne* et *technologie mobile*) et pour deux paires de langues : français-anglais et roumain-anglais. Ceci permet d'étudier le comportement de différentes approches pour une paire de langues qui sont richement représentées et pour une paire qui comprend le roumain, une langue peu dotée en ressources.

Ce chapitre est organisé de la façon suivante : un aperçu général de notre approche est tout d'abord présenté dans la section 5.2. Dans la section 5.3, nous décrivons la ressource sémantique sur laquelle se base notre approche. Puis, nous présenterons le processus de désambiguïsation sémantique proposé (section 5.4). Enfin, nous détaillons et analysons les résultats obtenus dans la section 5.5.

5.2 Aperçu général de l'approche

L'approche d'acquisition de lexiques bilingues à partir de corpus comparables que nous proposons reprend les étapes de l'approche standard. Comme il a été mentionné dans la section 5.1, lorsque l'extraction lexicale porte sur un domaine particulier, les traductions du dictionnaire bilingue ne sont pas toutes pertinentes pour représenter des vecteurs de contexte des mots du corpus source avec des mots de la langue cible. Pour cette raison, nous introduisons un processus de désambiguïsation lexicale qui vise à améliorer l'adéquation des vecteurs de contexte et donc améliorer les résultats de l'approche standard. Un aperçu général de l'approche proposée pour l'extraction de lexiques bilingues est décrit dans la Figure 5.2. Cette approche se compose des étapes suivantes :

1. **Caractérisation du terme à traduire** (pivot) par la constitution d'un vecteur de contexte dans lequel nous collectons toutes les unités lexicales cooccurrentes dans une fenêtre contextuelle donnée. Des vecteurs de contexte sont également construits pour toutes les unités lexicales susceptibles d'être des traductions candidates à partir du corpus cible. Les relations entre les mots du contexte et l'élément pivot

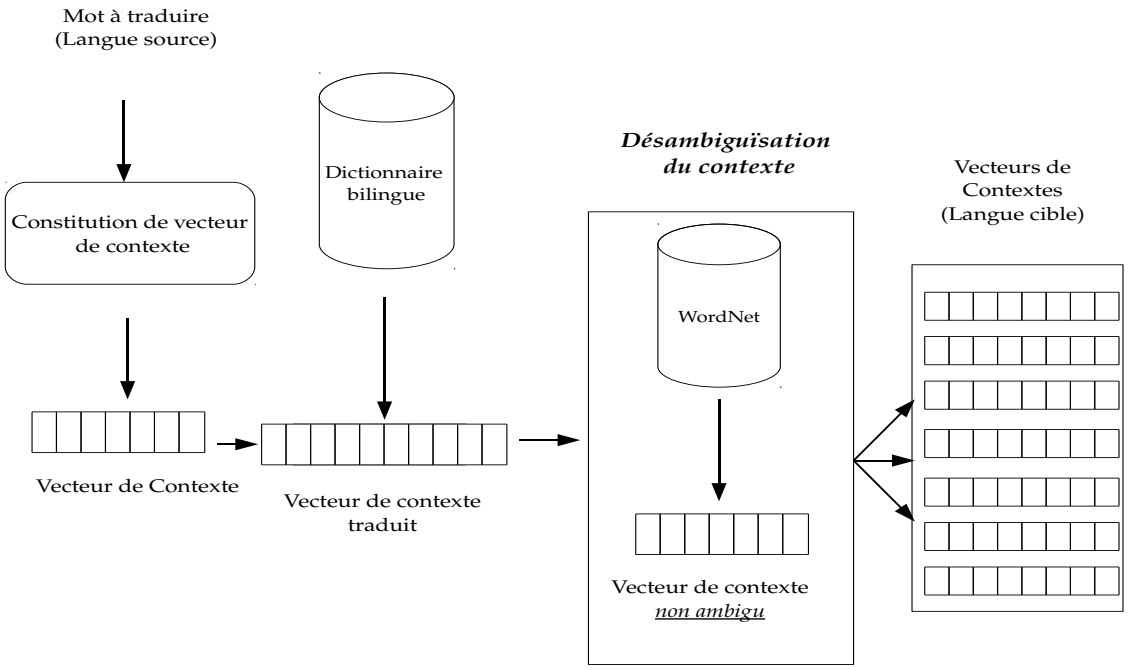


FIGURE 5.2: Aperçu général de l'approche d'extraction lexicale

du vecteur sont évaluées avec une mesure d'association. Les vecteurs enregistrent alors le motif d'association de chaque mot pivot avec ses voisins.

2. **Traduction du vecteur de contexte source** par le biais d'un dictionnaire bilingue amorce. Si le dictionnaire propose plusieurs traductions pour un élément, nous ajoutons l'ensemble des traductions proposées. Les mots ne figurant pas dans le dictionnaire sont tout simplement ignorés.

3. **Désambiguïsation lexicale des mots polysémiques** de chaque vecteur de contexte. Le processus de désambiguïsation proposé repose sur les mesures de similarité sémantique dérivées de la base de données lexicale WordNet. À la sortie de ce processus, le vecteur de contexte résultant synthétise et représente au mieux le terme à traduire. Cette étape sera décrite en détail dans le reste de ce chapitre.

4. **Comparaison des vecteurs source et cibles** en utilisant des mesures de similarité. Plus deux vecteurs de contexte sont similaires, plus il est probable qu'ils correspondent à des traductions.

5.3 Ressources sémantiques

Un grand nombre de techniques de désambiguïsation sémantique ont été présentées dans la littérature. Ces techniques se divisent en deux catégories principales : les *approches par apprentissage* et les *approches à base de connaissances à priori*. Les approches par apprentissage se divisent à leur tour en deux sous catégories. Nous distinguons les approches supervisées, qui nécessitent des corpus d'apprentissage étiquetés et les approches non supervisées classiques (clustering) qui viennent combler des limites des premières. La limite principale des approches supervisées est que l'obtention de grandes quantités de textes annotés en sens est très coûteux en temps et en argent et que l'on se heurte au goulot d'acquisition de données (Wagner, 2005). De plus, la qualité de la désambiguïsation de ces approches est restreinte par les exemples utilisés pour l'apprentissage. C'est pour cette raison que les méthodes non supervisées sont intéressantes puisqu'elles n'exploitent que des données non annotées.

Les approches à base de connaissances reposent sur l'utilisation de ressources lexicales comme les inventaires de sens, thésaurus, etc. C'est sous cette dernière catégorie que s'inscrit notre approche de désambiguïsation de vecteurs de contexte. Cette approche dérive une valeur de similarité sémantique entre différentes unités lexicales en utilisant WordNet et différentes mesures de similarité sémantique. Dans cette section, nous décrivons tout d'abord WordNet et ses différents domaines d'application. Ensuite nous présentons différentes les mesures de similarité sémantique proposées dans la littérature que nous allons utiliser.

5.3.1 WordNet

WordNet (Fellbaum, 1998) est une ressource lexicale structurée et joue le rôle d'un inventaire de sens et d'un dictionnaire. Une particularité de cette ressource est qu'elle donne accès à une hiérarchie de sens. Son but est de répertorier, classifier et mettre en relation de diverses manières le contenu sémantique et lexical de la langue anglaise. La composante principale sur laquelle repose le système entier est le *synset*, un groupe de mots interchangeables, dénotant un sens ou un usage particulier. À l'instar d'un dictionnaire traditionnel, WordNet offre, pour chaque mot une liste de

synsets correspondant à toutes ses acceptions répertoriées. WordNet définit ainsi le mot anglais *"share"* à l'aide des synsets suivants :

1. *N. share, portion, part, percentage (assets belonging to or due to or contributed by an individual person or group)*
2. *N. share (any of the equal portions into which the capital stock of a corporation is divided and ownership of which is evidenced by a stock certificate)*
3. *N. parcel, portion, share (the allotment of some amount by dividing something)*
4. *N. contribution, part, share (the effort contributed by a person in bringing about a result)*
5. *N. plowshare, ploughshare, share (a sharp steel wedge that cuts loose the top layer of soil)*
6. *V. share (have in common)*
7. *V. share (use jointly or in common)*
8. *V. partake, share, partake in (have, give, or receive a share of)*

Chaque synset dénote une acception différente du mot *share*, décrite par une courte définition. Une occurrence particulière de ce mot dénotant par exemple le premier sens (le plus courant), dans le contexte d'une phrase ou d'un énoncé, serait ainsi caractérisée par le fait qu'on pourrait remplacer le mot polysémique par l'un ou l'autre des mots du synset sans altérer la signification de l'ensemble. En outre, WordNet répertorie une grande variété de relations sémantiques lexicales comme *l'antonymie* ou taxinomiques comme par exemple *l'hyperonymie*, ou la *méronymie* permettant d'organiser le sens des mots en des systèmes de catégories que l'on peut consulter de manière cohérente et uniforme.

Cette ressource est largement utilisée dans des applications reposant sur le calcul de similarité des mots telles que la recherche de documents (Hwang *et al.*, 2011) ou d'images (Cho *et al.*, 2007; Choi *et al.*, 2012). Dans notre étude, cette ressource est utilisée pour dériver une similarité sémantique entre les éléments d'un vecteur de contexte. En sélectionnant les éléments les plus saillants, nous assurons une meilleure représentation du terme à traduire. Nous utilisons la version 3.0 de WordNet qui compte environ 117 597 synsets. À notre connaissance, c'est une première application de WordNet en extraction de lexiques bilingues à partir de corpus comparables.

5.3.2 Mesures de similarité sémantique

Globalement, il existe dans la littérature un bon nombre de mesures permettant de calculer, dans une hiérarchie de concepts comme WordNet, une valeur de similarité sémantique entre n'importe quel couple de mots. Parmi les mesures utilisant WordNet, nous distinguons :

1. Les mesures *à base de distance taxinomique* qui comptent simplement la distance entre deux mots dans la taxinomie de WordNet.
2. Les mesures *à base de traits*, qui considèrent la similarité entre deux mots comme le nombre d'unités lexicales en commun dans leurs définitions.
3. Les mesures se fondant sur le *contenu informationnel*, où un corpus annoté est nécessaire pour le calcul des fréquences des mots à comparer.

Dans cette étude, nous utilisons cinq mesures de similarité issues des types 1 et 2. Elles servent à sélectionner, parmi les différentes traductions candidates pour un terme donné, celle qui représente au mieux le sens du terme dans un contexte local. Le choix de ces mesures a été fait, (1) en sélectionnant les mesures connues dans la littérature comme étant les plus performantes, (2) de façon à varier le type de mesure utilisé et (3) à limiter au maximum la dépendance de la méthode à des ressources linguistiques, comme les corpus annotés dans les mesures à base de contenu informationnel.

5.3.2.1 À base de distance taxinomique

Le principe des mesures à base de distance taxinomique est de compter simplement le nombre de nœuds qui séparent deux mots dans une taxinomie. Plus la distance est faible, plus la similarité est élevée. Une limitation de ces approches est qu'elles ne reposent que sur les liens taxinomiques, qui ne représentent que des distances uniformes. Néanmoins, l'avantage principal de ce type de mesure est qu'elles ne requièrent aucun corpus, limitant ainsi leur exposition au problème des données éparses.

– **Path** : PATH constitue la mesure de similarité sémantique de référence. Elle est égale à l'inverse du chemin le plus court ($length(w_1, w_2)$) entre deux mots w_1 et w_2 (équation 5.1).

$$Sim_{path}(w_1, w_2) = \frac{1}{length(w_1, w_2)} \tag{5.1}$$

– **Leacock et Chodorow** (Leacock et Chodorow, 1998) : Dans cette mesure, que nous noterons LEACOCK, la similarité entre deux mots w_1 et w_2 est définie par la longueur du plus court chemin entre ces deux mots dans la *hiérarchie est-un* (is-a) de WordNet divisée par deux fois la profondeur maximale de la taxinomie, notée D. D représente la taille du plus long chemin possible d'une feuille au nœud racine dans la hiérarchie. Le plus court chemin entre deux nœuds est celui qui comprend le plus petit nombre de nœuds intermédiaires. Cette mesure est définie comme suit :

$$Sim_{lch}(w_1, w_2) = -\log(length(w_1, w_2)/(2 * D)) \tag{5.2}$$

où $length(w_1, w_2)$ est le plus court chemin entre deux nœuds et D la profondeur maximale dans la taxinomie. En utilisant la profondeur maximale de la taxinomie D, la valeur de similarité sera toujours supérieure à zéro, car il y aura toujours un chemin entre n'importe quelle paire de mots, à condition qu'ils soient trouvés dans WordNet.

– **Wu et Palmer** (Wu et Palmer, 1994) : Cette mesure notée ici par WUP a été initialement utilisée dans un système de traduction de verbes pour la paire de langues anglais-chinois. Cette mesure utilise la distance qui sépare deux synsets par rapport à leur ancêtre commun le plus spécifique (LCS) et par rapport à la racine de la taxinomie. La similarité entre deux mots w_1 et w_2 est définie par :

$$Sim_{wup}(w_1, w_2) = \frac{2 * depth(LCS)}{depth(w_1) + depth(w_2)} \tag{5.3}$$

où $depth(LCS)$ est le nombre de nœuds qui séparent LCS de la racine et $depth(w_1)$ (respectivement $depth(w_2)$) est le nombre de nœuds qui séparent w_1 (respectivement w_2) de la racine en passant par LCS. Selon (Lin, 1998), WUP, a l'avantage d'avoir de bonnes performances par rapport au autres mesures de similarité.

5.3.2.2 À base de traits

Les mesures de similarité à base de traits considèrent la similarité entre deux mots comme le nombre d'unités lexicales en commun dans leurs définitions. En outre, ce

type de mesure fait appel aux différentes relations définies dans WordNet. L'utilisation de ce nombre relativement élevé de relations a pour but de couvrir au maximum les différents types de liens que deux concepts peuvent partager. Deux mesures de ce type ont été utilisées dans ce cadre :

- **Lesk étendu** (Banerjee et Pedersen, 2002) : Cette mesure considère la similarité entre deux mots comme le nombre de mots en commun dans leurs définitions. Dans la version originale, on ne prend pas en compte de l'ordre des mots dans les définitions (sac de mots). Dans cette version étendu du Lesk, cette mesure est améliorée de deux façons. La première est l'incorporation des définitions des mots reliés par des relations taxinomiques de WordNet dans la définition d'un mot donné. La deuxième est une nouvelle manière de calculer le recouvrement entre les mots des définitions. Pour calculer le recouvrement entre deux mots, ils proposent de considérer non seulement le recouvrement entre les définitions des deux mots mais aussi des définitions issues de différentes relations : hyperonymes (has-kind), hyponymes (kind-of), méronymes (part-of), holonymes (has-part), etc. La mesure, notée LESK, est définie comme suit :

$$Sim_{lesk}(w_1, w_2) = \sum_{r \in R} overlap(gloss(r(w_1)), gloss(r(w_2))) \tag{5.4}$$

  où R est l'ensemble des relations dans WordNet considérées pour la mesure et $gloss()$ n'est autre que la définition associée aux synsets dans WordNet. Un problème important de LESK est qu'elle est très sensible aux mots présents dans la définition, et si certains mots importants manquent dans les définitions utilisées, les résultats obtenus seront de qualité moindre. De plus si les définitions sont trop concises (comme c'est souvent le cas) il est difficile d'obtenir des distinctions de similarité fines.
- **Vector** (Patwardhan, 2003) : Dans cette mesure que nous noterons VECTOR, chaque synset dans WordNet est représenté par un vecteur de contexte. Ce vecteur de contexte est simplement formé des unités lexicales qui constituent la définition du synset (cooccurrences de premier ordre) et de cooccurrences de second ordre. Ces cooccurrences ne sont autre que des mots déduits à partir de cooccurrences de premier ordre. Par exemple, si {*car* et *mechanic*} et, {*car* et *police*} sont des cooccurrences de premier ordre, *police* et *mechanic* seraient des cooccurrences de second ordre. La similarité entre deux synsets est ainsi déduite

en calculant le cosinus de l'angle entre les vecteurs correspondants. VECTOR est définie dans l'équation suivante :

$$Sim_{vector}(w_1, w_2) = \frac{v_1.v_2}{\|v_1\|.\|v_2\|} \tag{5.5}$$

où v_i est le vecteur de contexte du concept i.

5.3.3 Évaluation des mesures de similarité

Quand une mesure de similarité sémantique est évaluée, la référence (*gold standard*) le plus souvent utilisée est celle présentée dans les expériences présentées par (Miller et Charles, 1991). Dans (Seco *et al.*, 2004), toutes les mesures de similarité sémantique définies dans le module *WordNetSimilarity* [1] ont été évaluées en utilisant les paires de noms définies dans (Miller et Charles, 1991). Dans ces expériences, 30 paires de noms ont été classées par 38 humains en donnant une note de 0 (non similaire) à 4 (similarité parfaite) pour chaque paire. Le score final pour chaque paire n'est autre que la moyenne des scores obtenus par les 38 annotateurs. Selon (Seco *et al.*, 2004), les scores des cinq mesures de similarité employées dans notre approche de désambiguïsation sémantique sont présentées dans le tableau 5.1.

Mesure	Jugement humain
PATH	0.7
WUP	0.74
LEACOCK	0.82
LESK	0.67
VECTOR	0.87

TABLEAU 5.1: Jugement humain de différentes mesures de similarité sémantique utilisées.

Nous constatons que toutes les mesures de similarité employées sont en bonne corrélation avec le jugement humain et que VECTOR montre la corrélation la plus élevée avec une valeur de 0.87.

1. http://search.cpan.org/~tpederse/WordNet-Similarity/

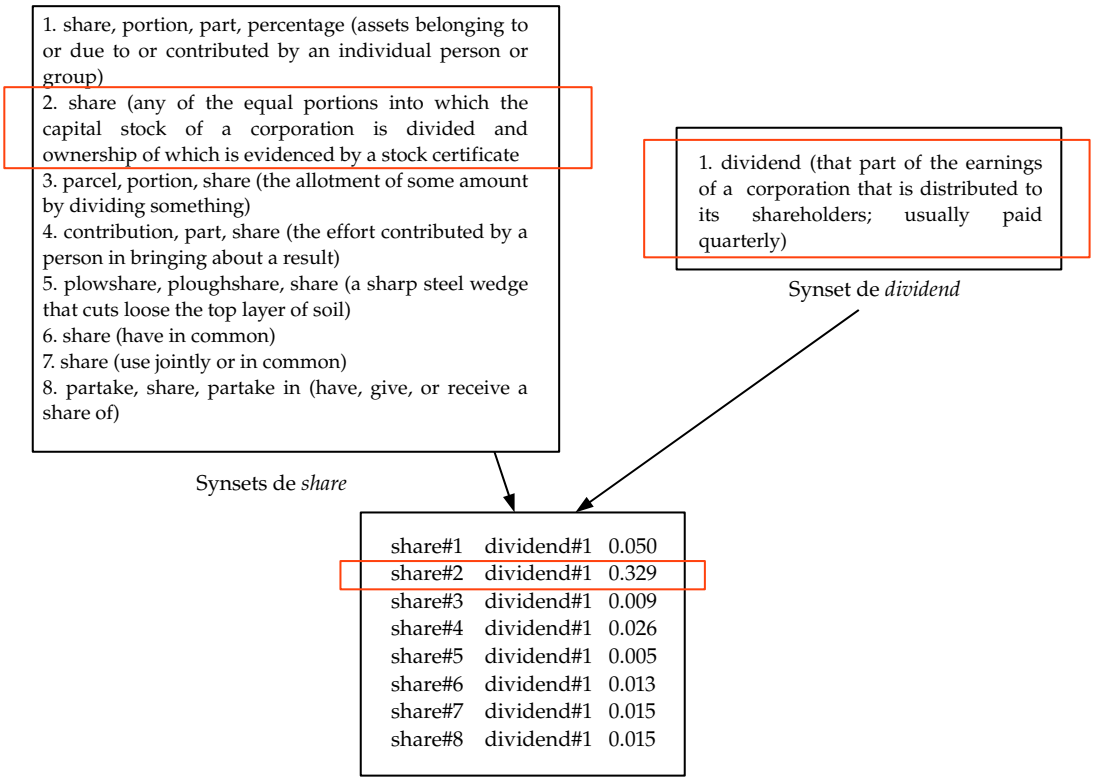


FIGURE 5.3: Valeurs de similarité obtenues pour tous les synsets associés à la paire de mots *(share,dividend)* en utilisant la mesure de similarité sémantique VECTOR

5.4 Algorithme de désambiguïsation

Maintenant que nous avons passé en revue les principales de mesures de similarité sémantique, nous allons présenter l'algorithme de désambiguïsation des vecteurs de contexte proposé. Comme il a été mentionné dans la section 5.2, une fois transférés dans la langue cible, la désambiguïsation des vecteurs de contexte intervient. L'algorithme proposé opère localement sur chaque vecteur de contexte, avec pour objectif de trouver la traduction la plus adéquate pour chacune des entrées polysémiques dans les vecteurs de contexte traduits. Nous utilisons à cet effet les mots monosémiques contenus dans le même vecteur de contexte, comme un ensemble d'unités lexicales non ambiguës pour inférer les sens de celles qui sont polysémiques. Nous émettons l'hypothèse qu'un mot est monosémique s'il ne possède qu'une seule traduction dans le dictionnaire bilingue. Cette hypothèse est vérifiée en sondant les

entrées monosémiques du dictionnaire bilingue contre celles de WordNet. Nous avons constaté que 95% des entrées sont monosémiques dans les deux ressources.

Conformément aux différentes mesures de similarité sémantique décrites dans la section 5.3.2, une valeur de similarité sémantique Sim est dérivée en comparant toutes les traductions fournies par le dictionnaire bilingue pour chaque mot polysémique et toutes les unités monosémiques apparaissant dans le même vecteur de contexte. Puisqu'un mot peut être défini par plus d'un synset dans WordNet, nous déterminons la similarité sémantique entre deux mots w_1 et w_2 comme le maximum de similarité sémantique Sim entre le ou les synsets parmi les $synsets(w_1)$ et les $synsets(w_2)$ selon la formule suivante :

$$Sem_{Sim}(w_1, w_2) = \max\{Sim(s_1, s_2); (s_1, s_2) \in synsets(w_1) \times synsets(w_2)\} \quad (5.6)$$

Par exemple, les valeurs de similarité sémantique obtenues par la mesure VECTOR pour la paire des mots anglais *share* et *dividend* sont décrites dans la figure 5.3. Ces mots sont représentés dans WordNet par respectivement huit et un synsets. Dans ce cas, tous les synsets associés à *share* sont comparés à l'unique synset de *dividend*. Par conséquent, la relation entre ces deux unités est représentée par la valeur de similarité maximale (*share*#2 *dividend*#1 0.329). Dans le cas où tous les mots du vecteur de contexte sont polysémiques, il est possible de prendre en compte toutes les combinaisons de mots possibles dans le calcul de similarité sémantique. Néanmoins, une augmentation de la complexité algorithmique et une détérioration des résultats d'extraction ont été constatés dans des expérimentations préliminaires. C'est pour cette raison que nous ne traitons que les vecteurs de contexte contenant au moins un mot monosémique. Nous obtenons ainsi une valeur de similarité entre chaque mot polysémique w_p et toutes les unités monosémiques. La valeur de similarité finale de chaque traduction w_p^j de w_p est sa *moyenne de similarité* à tous les mots monosémiques du vecteur de contexte, qui se définit ainsi :

$$Sim\_Ave(w_p^j) = \frac{\sum_{i=1}^{N} Sem_{Sim}(w_i, w_p^j)}{N} \quad (5.7)$$

où N est le nombre total des mots non polysémiques du vecteur traduit et Sem_{Sim} est la valeur de similarité entre w_p^j et le $i^{ème}$ mot monosémique. En fonction des moyennes de similarité, nous obtenons pour chaque w_p une liste ordonnée de traductions $w_p^1 \ldots w_p^n$.

Vecteur de contexte	Traductions	Comparaison	Sim_Ave
liquidité	liquidity	–	–
action	act	Sem_{Sim}(act,liquidity), Sem_{Sim}(act,dividend)	0.2139
	action	Sem_{Sim}(action,liquidity), Sem_{Sim}(action,dividend)	0.4256
	stock	Sem_{Sim}(stock,liquidity), Sem_{Sim}(stock,dividend)	***0.5236***
	deed	Sem_{Sim}(deed,liquidity), Sem_{Sim}(deed,dividend)	0.1594
	lawsuit	Sem_{Sim}(lawsuit,liquidity), Sem_{Sim}(lawsuit,dividend)	0.1212
	fact	Sem_{Sim}(fact,liquidity), Sem_{Sim}(fact,dividend)	0.1934
	share	Sem_{Sim}(share,liquidity), Sem_{Sim}(share,dividend)	***0.5236***
	plot	Sem_{Sim}(plot,liquidity), Sem_{Sim}(plot,dividend)	0.2011
dividende	dividend	–	–

Un exemple de désambiguïsation de vecteur de contexte du terme français « bénéfice » est décrit dans le tableau **??**. Ce vecteur est construit à partir de corpus comparable spécialisé et contient les mots *action*, *dividende*, *liquidité* et d'autres unités. Lors du transfert de ce vecteur de la langue source (français) à la langue cible (anglais), le dictionnaire bilingue propose les traductions « *act, stock, action, deed, lawsuit, fact, operation, plot, share* » , « *dividend* » et « *liquidity* » pour traduire respectivement les mots « *action* » , « *dividende* » et « *liquidité* » . Nous utilisons les unités lexicales non polysémiques « *dividende* » et « *liquidité* » pour désambiguiser le mot « *action* » . En observant la valeur de *Ave_Sim*, nous remarquons que dans ce contexte, les mots *share* et *stock* sont les traductions les plus appropriées au mot *action*. Nous remarquons aussi que les mots issus du domaine général se placent après pour retrouver à la fin les unités les moins proches (*deed et lawsuit*).

5.5 Évaluations

Cette section est consacrée à la présentation de l'ensemble des évaluations menées dans cette étude. Nous nous intéressons à l'extraction de lexiques bilingues à partir de corpus comparables spécialisés dans les domaines de la *finance des entreprise, cancer du sein, énergie éolienne* et de la *technologie mobile* et pour les paires de langues *français-anglais* et *roumain-anglais*. La majorité des paramètres sur lesquels se basent nos expériences, à savoir les corpus comparables, les dictionnaires bilingues amorces et les listes de références, ont été détaillés dans le chapitre 4.

L'approche proposée dans ce cadre consiste à augmenter l'approche standard par un processus de désambiguïsation lexicale qui se base sur différentes mesures de si-

milarité sémantiques. Deux types d'évaluations ont été conduites. Nous évaluons et comparons dans un premier temps les performances des cinq mesures de similarité sémantiques employées et utilisons dans un second temps une méthode de *fusion de données* qui combine les résultats obtenus par chacune des mesures. Cette section est organisée comme suit : nous décrivons tout d'abord les approches utilisées comme référence auxquelles sera comparée notre modèle d'extraction lexicale. Ensuite, nous étudions la polysémie dans les corpus comparables sur lesquels se basent nos expérimentations. Puis, nous décrivons la méthode de fusion de données que nous utilisons pour la combinaison des résultats. Enfin, nous présentons et analysons les résultats obtenus.

5.5.1 Approches de référence

Les résultats du modèle d'extraction lexicale intégrant le processus de désambiguïsation lexicale présenté dans la section 5.4 sont comparés aux approches de l'état de l'art suivantes :

- **L'approche standard** : lors du transfert des vecteurs de contexte, lorsque le dictionnaire propose plusieurs traductions pour un élément i, l'ensemble des traductions proposées j_x sont ajoutées avec la même importance dans le vecteur de contexte.
- **L'approche de pondération des vecteurs** : cette méthode a été proposée par (Morin et Prochasson, 2011). Dans le cas où plusieurs traductions sont disponibles, elles sont prises en compte en fonction de leur fréquence dans le corpus cible. C'est-à-dire que le score d'association de l'élément i sera réparti entre les différentes traductions j_x, cette répartition se faisant en fonction de la fréquence des éléments j_x dans le corpus cible. Le nouveau score d'association se calcule ainsi :

$$assoc(j_x) = assoc(i) * \frac{freq(j_x)}{\sum_k freq(j_k)}; \quad \forall j_x; j_x \in traductions(i) \tag{5.8}$$

  Cette méthode permet de faire ressortir les traductions les plus pertinentes pour un contexte donné.

Corpus	Taux de polysémie (%)	
	Francais	Roumain
Finance des entreprises	41	1
Cancer du sein	47	2.5
Énergie éolienne	51	2.1
Technologie mobile	37	1,9

TABLEAU 5.2: Taux de polysémie des corpus français et roumain pour les quatre domaines.

5.5.2 Polysémie dans les corpus comparables

La polysémie dans les corpus comparables sur lesquels porte notre étude constitue un paramètre important à évaluer. Ce paramètre informe sur l'importance de la désambiguïsation des vecteurs de contexte pour les différents corpus comparables et les différentes paires de langues. Vu qu'elle intervient lors de la traduction des vecteurs de contexte, nous mesurons un *taux de polysémie* pour les corpus en langue source (français et roumain). Pour les quatre domaines d'étude, les taux de polysémie sont consignés dans le tableau 5.2. Ce taux indique combien de mots dans les corpus français et roumain sont associés à plus d'une traduction dans respectivement les dictionnaires bilingues français-anglais et roumain-anglais.

Nous constatons que les corpus français sont hautement polysémiques, ce qui montre que la couverture du dictionnaire utilisé est suffisamment bonne pour couvrir différents domaines. Toutefois, l'ambiguïté notée dans les corpus en roumain n'est pas importante. La raison en est que les dictionnaires bilingues faisant intervenir des langues faiblement représentées comme le roumain sont par nature rares et s'ils existent, leur couverture n'est toujours pas suffisante. Dans le reste de nos expérimentations, ce paramètre sera pris en considération pour l'interprétation des résultats obtenus.

5.5.3 Fusion de données par système de vote

La désambiguïsation des vecteurs de contexte repose sur différentes mesures de similarité sémantiques dérivées de WordNet. Comme cela est rappelé dans (Zweigen-

baum et Habert, 2006) : « *Lorsque l'on dispose de plusieurs méthodes pour résoudre un problème, il est souvent plus productif de chercher à les combiner.* »

Partant de cette idée, nous proposons de combiner les résultats obtenus par chacune des mesures de similarité sémantique. Ceci devrait renforcer la confiance dans les traductions faisant consensus, tout en diminuant la confiance dans les traductions non consensuelles. Pour ce faire, nous utilisons une méthode de fusion de données largement utilisée dans la combinaison de différents systèmes de recherche d'information (Montague et Aslam, 2002; Nuray et Can, 2006), nommée la *méthode Condorcet*. Cette méthode est un système de vote qui accepte deux ou plusieurs listes ordonnées, les fusionne et propose une seule liste assurant une efficacité meilleure de tous les systèmes utilisés pour la fusion de données.

Le Condorcet compare chaque paire de candidats possibles pour décider de la préférence entre eux. Une matrice peut être utilisée pour présenter le processus de vote où chaque candidat apparaît dans la matrice aussi bien en ligne qu'en colonne. S'il existe m candidats, la matrice contiendra m^2 éléments. Initialement, la valeur 0 est attribuée à tous les éléments. Si d_i est préféré à d_j, nous ajoutons 1 à l'élément de la ligne i et colonne j (a_{ij}). Ce processus est répété jusqu'à ce que tous les bulletins de vote (matrice par paire de candidats) soient traités. Une fois tous les votes traités, on examine le contenu de la matrice finale, constituée en sommant tous les bulletins de vote. Pour chaque élément a_{ij}, si $a_{ij} > m/2$ alors d_i est préféré à d_j ; si $a_{ij} < m/2$ alors d_j est préféré à d_i ; sinon ($a_{ij} = m/2$), il y a égalité entre d_i et d_j. Le score final de chaque candidat est quantifié en additionnant les scores qu'il obtient dans tous les duels. Enfin, le classement final est réalisé en fonction des scores totaux calculés. Par exemple, considérons une assemblée de 60 votants ayant le choix entre trois propositions a, b et c, et supposons que les préférences se répartissent ainsi (en notant $a > b$, le fait que a est préféré à b) :

$$\begin{aligned}
23 \text{ votants préfèrent : } & a > c > b \\
19 \text{ votants préfèrent : } & b > c > a \\
16 \text{ votants préfèrent : } & c > b > a \\
2 \text{ votants préfèrent : } & c > a > b
\end{aligned}$$

Dans les comparaisons majoritaires par paires, on obtient :

35 préfèrent $b > a$ contre 25 pour$a > b$
41 préfèrent $c > b$ contre 19 pour $b > c$
37 préfèrent $c > a$ contre 23 pour $a > c$

Ce qui conduit à la préférence majoritaire $c > b > a$.

Dans notre étude, nous considérons le classement des résultats d'extraction lexicale provenant de différentes mesures de similarité comme un cas particulier du problème de vote où les candidats à la traduction trouvés dans les $succès_{20}$ correspondent aux candidats et les cinq mesures de similarité employés individuellement sont les votants. Nous nous appuyons sur cette méthode que nous nommons CONDORCET$_{Merge}$ pour reclasser les résultats d'extraction. L'évaluation de cette méthode se fait suivant l'ordre de la nouvelle liste de traductions fournie pour chaque terme à traduire.

5.5.4 Résultats expérimentaux et analyse

Les performances de notre approche et des approches de référence sont évaluées en utilisant les mesures de la F-Mesure et la MAP au $succès_{20}$ décrites dans le chapitre 4. Chaque mesure de similarité sémantique utilisée pour la désambiguïsation des vecteurs de contexte fournit, pour chaque unité polysémique, une liste de traductions ordonnée en fonction des valeurs de similarité obtenues. Á cet égard, il convient de s'interroger sur le nombre de traductions à considérer pour chaque mot polysémique : devrions nous ne considérer que l'élément pivot ayant la valeur de similarité maximale ou envisager un plus grand nombre de traductions, notamment lorsque la liste de traductions contient des synonymes (*share* et *stock* dans le tableau **??**). C'est précisément pour cette raison que dans nos expérimentations, nous prenons en considération pour chaque unité polysémiques différent nombres de traductions. Ce nombre est fixé en fonction du nombre de traductions moyen associé à chaque mot dans les corpus comparables. Pour le couple français-anglais, un mot de chaque corpus comparable possède en moyenne 7 traductions dans le dictionnaire bilingue français-anglais. Concernant la paire de langues roumain-anglais, un mot est associé en moyenne à 2 traductions dans le dictionnaire roumain-anglais. Ces méthodes sont notées par WN-T$_i$ avec i allant de la traduction pivot seule ($i = 1$) jusqu'au nombre moyen de traductions (7 pour le français-anglais et 2 pour le roumain-anglais).

Pour la paire de langues **français-anglais**, les valeurs de F-mesures obtenues par chaque méthode sont présentées dans le tableau 5.3. Pour le corpus traitant la thématique de la finance des entreprises, nous constatons que la désambiguïsation des vecteurs de contextes améliore considérablement les performances de l'approche standard pour toutes les configurations (mesures de similarité et nombre de traductions à considérer). La meilleure F-mesure est atteinte par la mesure VECTOR. La considération des deux traductions les plus similaires aux mots monosémiques (WN-T_2) dans les vecteurs de contexte fait passer la F-mesure de 0.17 à 0.31. Toutefois, aucune amélioration n'est constatée par rapport à MP11. Des résultats similaires sont obtenus pour le domaine de l'énergie éolienne. La meilleure F-mesure est obtenue par PATH lorsque les vecteurs de contexte sont totalement non ambigus (WN-T_1) avec un gain de +15% en F-mesure au $succès_{20}$. Par rapport à MP11, tout comme dans le domaine de la finance des entreprises, aucune amélioration n'est notée. En ce qui concerne la thématique de cancer du sein, les résultats montrent des améliorations pour la plupart des configurations par rapport à l'approche standard et celle de MP11. La F-mesure maximale est obtenue par la mesure de similarité LESK quand, pour chaque unité polysémique, un maximum de quatre traductions sont considérées dans les vecteurs de contexte. Cette méthode permet d'obtenir un gain respectif de +9% et +3% par rapport à l'approche standard et à l'approche de pondération de vecteurs MP11. Pour le domaine de la technologie mobile, les résultats obtenus montrent que la désambiguïsation des vecteurs de contextes apporte des valeurs élevés de F-mesures par rapport aux deux approches de l'état de l'art pour toutes les configurations.

Chacune des mesures de similarité utilisée offre une vision différente de la façon dont les candidats à la traduction d'un terme à traduire sont classés. Pour cette raison, nous utilisons la méthode CONDORCET_{Merge} présentée dans la section 5.5.3 pour fusionner et reclasser les résultats obtenus par les mesures de similarité utilisées individuellement. Les résultats obtenus par cette méthode (Tableaux 5.3a, 5.3b, 5.3c et 5.3d) dépassent ceux fournis par : (1) l'approche standard, (2) MP11 et (3) les mesures de similarité individuelles.

Bien que les quatre corpus soient assez différents (domaines et taux de polysémie), les résultats optimaux sont obtenus pour la plupart des corpus lorsque l'on considère pour chaque unité polysémique les deux traductions les plus similaires aux mots monosémiques du même vecteur de contexte. Les gains en F-mesure par rapport à

a) Finance des entreprises

Méthode		WN-T$_1$	WN-T$_2$	WN-T$_3$	WN-T$_4$	WN-T$_5$	WN-T$_6$	WN-T$_7$
Approche Standard (SA)					0.17			
MP11					0.33			
Mesures individuelles	WuP	0.24	0.28	0.30	0.27	0.25	0.21	0.22
	Path	0.25	0.28	0.30	0.28	0.25	0.21	0.21
	Leacock	0.25	0.29	0.30	0.27	0.27	0.24	0.23
	Lesk	0.27	0.29	0.29	0.27	0.25	0.25	0.21
	Vector	0.26	0.31	0.28	0.28	0.23	0.23	0.23
Condorcet$_{Merge}$		**0.36**	**<u>0.37</u>**	**0.35**	**0.36**	**0.33**	**0.27**	**0.26**

b) Cancer du sein

Méthode		WN-T$_1$	WN-T$_2$	WN-T$_3$	WN-T$_4$	WN-T$_5$	WN-T$_6$	WN-T$_7$
Approche Standard (SA)					0.49			
MP11					0.55			
Mesures individuelles	WuP	0.48	0.56	*0.56*	0.54	0.55	0.54	0.55
	Path	0.54	0.54	0.55	0.56	*0.57*	0.55	0.55
	Leacock	0.50	*0.57*	0.55	0.56	0.54	0.55	0.54
	Lesk	0.46	0.54	0.54	**0.59**	0.55	0.55	0.54
	Vector	0.51	0.56	0.53	0.56	0.54	0.56	0.55
Condorcet$_{Merge}$		**0.56**	**<u>0.61</u>**	**0.60**	**0.59**	**0.60**	**0.57**	**0.57**

c) Énergie éolienne

Méthode		WN-T$_1$	WN-T$_2$	WN-T$_3$	WN-T$_4$	WN-T$_5$	WN-T$_6$	WN-T$_7$
Approche Standard (SA)					0.08			
MP11					0.24			
Mesures individuelles	WuP	0.15	0.20	0.17	0.18	0.16	0.16	0.15
	Path	0.23	0.23	0.19	0.16	0.12	0.12	0.10
	Leacock	0.18	0.17	0.19	0.19	0.15	0.15	0.13
	Lesk	0.20	0.19	0.19	0.16	0.11	0.13	0.10
	Vector	0.21	0.20	0.21	0.16	0.13	0.11	0.12
Condorcet$_{Merge}$		**<u>0.34</u>**	**0.30**	**0.28**	0.21	0.20	0.18	0.17

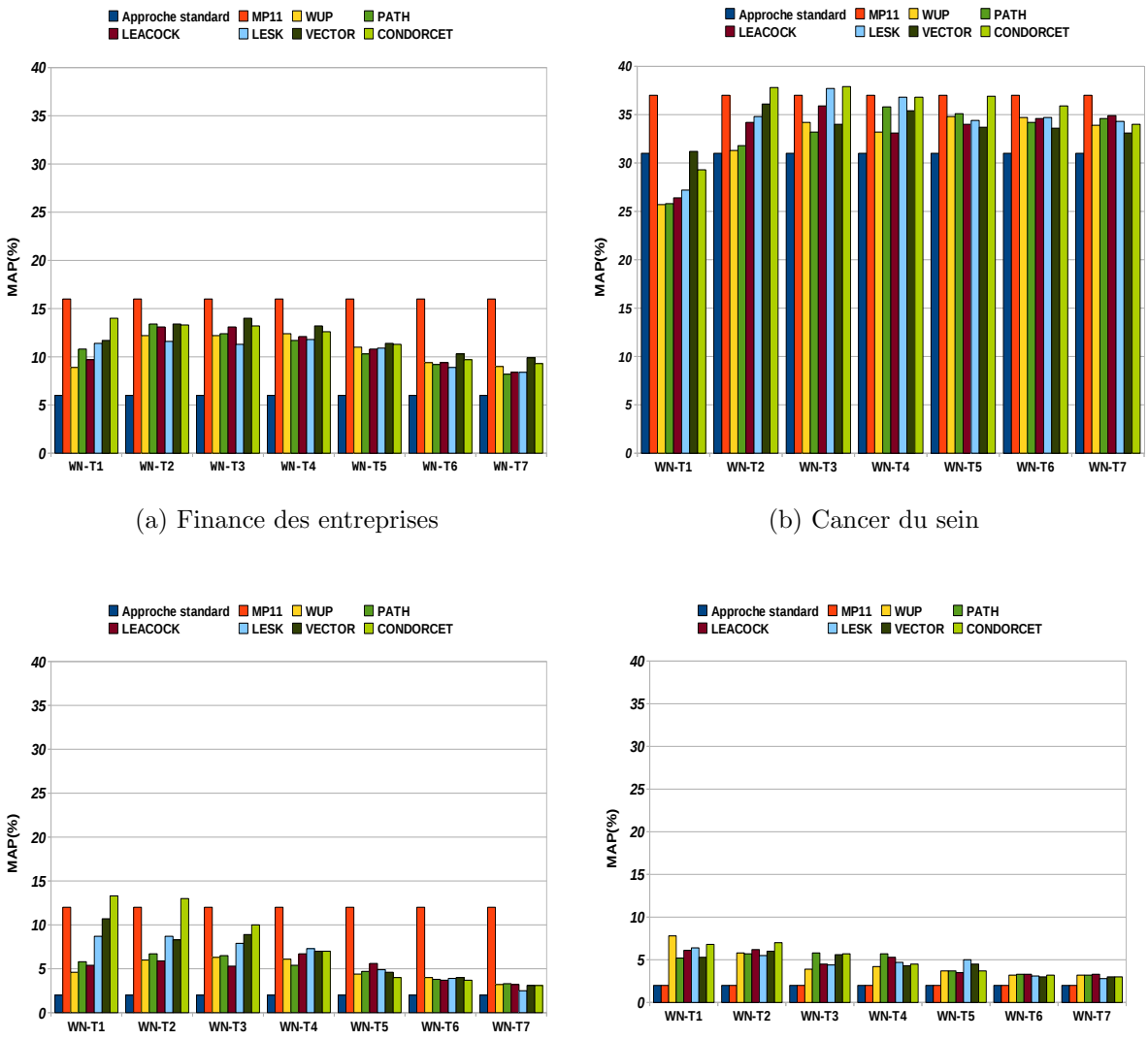
d) Technologie mobile

Méthode		WN-T$_1$	WN-T$_2$	WN-T$_3$	WN-T$_4$	WN-T$_5$	WN-T$_6$	WN-T$_7$
Approche Standard (SA)					0.06			
MP11					0.05			
Mesures individuelles	Wup	*0.15*	*0.17*	*0.13*	*0.10*	0.10	*0.08*	*0.08*
	Path	0.12	0.16	0.12	*0.10*	0.09	*0.08*	*0.08*
	Leacock	0.12	0.14	0.10	0.09	0.07	0.07	*0.08*
	Lesk	0.13	0.12	0.10	0.08	0.10	*0.08*	*0.08*
	Vector	0.14	0.15	0.11	*0.10*	*0.11*	*0.08*	0.07
Condorcet$_{Merge}$		**0.22**	**<u>0.24</u>**	**0.18**	**0.15**	**0.15**	**0.13**	**0.12**

Tableau 5.3: F-mesure au *succès*$_{20}$ pour les quatre domaines et la paire des langues français-anglais ; MP11=(Morin et Prochasson, 2011). Dans chaque colonne, le meilleur score obtenu par chaque mesure de similarité est en italique, les meilleurs résultats sont en gras, le meilleur résultat global est mis en gras et souligné.

l'approche standard sont de +20% pour la finance des entreprises, +12% pour le cancer du sein, +22% pour l'énergie éolienne et +18% pour la technologie mobile.

La figure 5.4 décrit les valeurs de MAP obtenues pour les quatre corpus comparables français-anglais. Nous remarquons que la méthode Condorcet$_{Merge}$ atteint aussi la meilleure MAP en considérant pour chaque mot polysémique les deux traductions les plus similaires aux éléments monosémiques des vecteurs de contexte (WN-



(a) Finance des entreprises

(b) Cancer du sein

(c) Énergie éolienne

(d) Technologie mobile

FIGURE 5.4: Valeurs de la MAP des quatre domaines pour la paire de langues *français-anglais*.

T_2). Pour les domaines de la *finance des entreprises, cancer du sein*, *énergie éolienne* et de la *technologie mobile* et par rapport à l'approche standard, cette méthode fait passer la MAP respectivement de 6 à 13%, 31 à 38%, de 6 à 13% et de 2 à 7%. Par rapport à MP11 de faibles gains en MAP de +1% pour les corpus du *cancer du sein* et *énergie éolienne* et de +5% pour le corpus de la *technologie mobile* ont été rapportés. Ceci montre que le reclassement des résultats par la méthode Condorcet est robuste au changement de domaine. Plus intéressant encore, la fusion des

a) Finance des entreprises

Méthode		WN-T_1	WN-T_2
Approche Standard (SA)		0.13	
MP11		0.13	
Mesures individuelles	WuP	0.13	0.13
	Path	0.13	0.13
	Leacock	0.13	0.13
	Lesk	0.13	0.13
	Vector	0.13	0.13
Condorcet$_{Merge}$		0.13	0.13

b) Cancer du sein

Méthode		WN-T_1	WN-T_2
Approche Standard (SA)		0.21	
MP11		0.21	
Mesures individuelles	WuP	0.18	0.21
	Path	0.18	0.21
	Leacock	0.15	0.18
	Lesk	0.21	0.21
	Vector	0.18	0.21
Condorcet$_{Merge}$		0.21	0.21

c) Énergie éolienne

Méthode		WN-T_1	WN-T_2
Approche Standard (SA)		0.08	
MP11		0.08	
Mesures individuelles	WuP	0.08	0.08
	Path	0.08	0.08
	Leacock	0.08	0.08
	Lesk	0.08	0.08
	Vector	0.08	0.08
Condorcet$_{Merge}$		0.08	0.08

d) Technologie mobile

Méthode		WN-T_1	WN-T_2
Approche Standard (SA)		0.16	
MP11		0.16	
Mesures individuelles	WuP	0.16	0.16
	Path	0.16	0.16
	Leacock	0.16	0.16
	Lesk	0.16	0.16
	Vector	0.16	0.16
Condorcet$_{Merge}$		0.16	0.16

Tableau 5.4: F-mesure au *succès*$_{20}$ pour les quatre domaines et pour la paire des langues roumain-anglais ; MP11=(Morin et Prochasson, 2011).

résultats d'extraction lexicale surpasse MP11 en F-mesure et MAP, montrant que la désambiguïsation est plus importante que la pondération des vecteurs de contexte. Nous constatons également que l'ajout progressif de traductions qui constituent du bruit pour le domaine d'étude dégrade les résultats.

Pour la paire de langues **roumain-anglais**, des résultats différents ont été obtenus. En observant le tableau 5.4, nous constatons que par rapport à l'approche standard et pour les quatre domaines, aucune amélioration n'a été remarquée. La raison en est que le dictionnaire bilingue utilisé pour la traduction des contextes est de faible couverture. Rappelons aussi que celui-ci a été construit à partir de titres d'articles de Wikipédia, qui dans la plupart des cas sont constitués d'entités nommées. Nous constatons donc que lorsque des corpus comparables comme ceux utilisés dans cette étude sont faiblement polysémiques (tableau 5.2), la désambiguïsation et la pondération des vecteurs de contextes semblent être inutile.

5.6 Conclusion

Nous avons présenté dans ce chapitre une approche qui vise à améliorer la représentativité des vecteurs de contexte. Nous introduisons un processus de désambiguïsation lexicale dans l'approche standard utilisée en extraction de lexiques bilingues à partir de corpus comparables. Ce processus se base sur des mesures de similarité sémantique dérivées de WordNet pour résoudre le problème de la polysémie des mots dans les vecteurs de contexte. Les évaluations ont été menées sur des corpus comparables spécialisés dans les domaines de la *finance des entreprises, cancer du sein, énergie éolienne* et de la *technologie mobile* pour les paires de langues français-anglais et roumain-anglais.

Dans nos expérimentations, nous avons tout d'abord évalué et comparé les performances des cinq mesures de similarité sémantique. Ensuite, nous avons utilisé la méthode Condorcet, une méthode de fusion de données pour fusionner et reclasser les résultats obtenus par toutes les mesures. Nous avons constaté que la fusion de données surpasse considérablement les résultats obtenus par deux approches de l'état de l'art. Plus intéressant encore, lorsque les corpus sont fortement polysémiques, la désambiguïsation des vecteurs de contexte affecte positivement les résultats globaux. Cela montre que l'ambiguïté présente dans les corpus comparables spécialisés entrave l'extraction de lexiques bilingues.

La principale faiblesse de cette approche est qu'elle s'appuie sur WordNet. Son application dépend donc de l'existence de cette ressource dans la langue cible. En outre, cette méthode est très dépendante de la couverture du dictionnaire bilingue utilisé pour traduire les vecteurs de contextes. En effet, les résultats obtenus pour la paire de langues roumain-anglais, une langue peu dotée en ressources, ne sont pas satisfaisants et la désambiguïsation s'avère alors inutile.

Chapitre 6

Analyse sémantique explicite pour l'extraction de lexiques bilingues

6.1 Introduction

Dans le chapitre précédent, nous avons présenté une approche qui propose d'ajouter un processus de désambiguïsation lexicale à l'approche standard utilisée en extraction de lexiques bilingues à partir de corpus comparables. Ce processus, qui repose globalement sur les mesures de similarité de WordNet, identifie la traduction la plus probable des mots polysémiques dans chaque vecteur de contexte de la langue source. Lorsque les corpus sur lesquels portent l'étude sont fortement polysémiques, la désambiguïsation des vecteurs de contexte affecte positivement les résultats globaux. Néanmoins, même avec l'essor de ce type d'approches, des problèmes comme l'écart entre la réalité sémantique de la langue et les traductions dérivées statistiquement, la rareté des ressources linguistiques pour la plupart des langues ou encore la qualité des ressources et traductions obtenues automatiquement restent ouverts. Le premier défi est d'ordre général et inhérent à toute approche automatique. Par contre, les deux autres défis peuvent être au moins partiellement résolus par une exploitation appropriée des ressources multilingues qui sont de plus en plus disponibles sur le Web.

Nous présentons une approche d'extraction de lexiques bilingue qui exploite Wikipédia de manière novatrice avec le but de s'attaquer aux problèmes mentionnées ci-dessus. Les principaux avantages d'utilisation de Wikipédia sont :

- La ressource est disponible dans des centaines de langues et est structurés en concepts ou articles non ambigüs.
- Les langues sont explicitement liées par des relations de traduction entre les entrées proposées par les contributeurs.
- Elle couvre un grand nombre de domaines et est donc potentiellement utile pour extraire un large éventail de lexiques spécialisés.

Un certain nombre de défis sont également associés à l'utilisation de Wikipédia notamment :

- La comparabilité des articles entre différentes langues est très variable.
- Le graphe de traductions est partiel puisque, si l'on considère toutes les paires de langues, seule une partie des articles est disponible dans les deux langues et explicitement connectées.
- Les domaines sont inégalement couverts dans Wikipédia (Halavais et Lackaffb, 2008), un ciblage de domaine est donc nécessaire.

L'approche proposée dans ce chapitre vise à exploiter les avantages de Wikipédia tout en répondant, adéquatement, aux enjeux associés. Parmi les techniques dédiées à l'extraction du contenu de Wikipédia, nous émettons l'hypothèse qu'une adaptation de l'analyse sémantique explicite (ESA) (Gabrilovich et Markovitch, 2007) serait adéquate à notre contexte d'application. L'ESA a été employée dans différentes applications du TAL comme par exemple l'estimation de similarité sémantique ou dans la classification de textes. Dans cette étude, cette technique est utilisée pour représenter les termes à traduire et extraire leurs équivalents de traduction. L'intuition qui sous tend notre approche est que au lieu de représenter les termes à traduire par analyse distributionnelle, qui consiste à les représenter dans l'espace des *mots du corpus*, elle les représente dans l'espace des *titres de Wikipédia*. En outre, pour passer de la représentation d'un terme de la langue source vers sa représentation dans la langue cible, nous remplaçons l'usage du dictionnaire bilingue par les liens interlingues de Wikipédia. Ceci permet de traiter aussi bien des langues pour lesquelles des ressources dictionnariques sont disponibles et des langues peu dotées de ressources. L'évaluation de cette approche est réalisée sur les quatre corpus comparables introduits dans le chapitre 4, pour les paires de langues français-anglais et roumain-anglais. Les résultats

expérimentaux montrent que les performances de l'approche nouvellement introduite dépassent de loin ceux obtenus par trois approches de l'état de l'art.

Ce chapitre est organisé de la manière suivante : dans la section 6.2, nous présentons tout d'abord l'analyse sémantique explicite et les différents travaux faisant appel à cette technique. Ensuite, nous présentons dans la section 6.3 une vue d'ensemble de notre approche d'extraction de lexiques bilingues. La section 6.4 est consacrée à l'étude de différentes représentation des termes à traduire. Nous décrivons par la suite notre approche de construction de graphe de traductions (section 6.5) et d'identification d'équivalents de traductions (section 6.6). Les évaluations menées dans ce cadre sont finalement présentés dans la section 6.7.

6.2 Analyse sémantique explicite (ESA)

L'analyse sémantique explicite (ESA) (Gabrilovich et Markovitch, 2007) est une méthode qui utilise Wikipédia comme une base de connaissance pour représenter des textes, allant de termes simples à des documents entiers, dans un espace sémantique structuré. Dans Wikipédia, chaque article associe un texte à une entrée de l'encyclopédie, entrée que l'on considère comme un *concept ou titre* t_j et chaque texte se compose de mots w_j. L'ESA calcule donc pour chaque mot w_j pris dans le vocabulaire du corpus des entrées de Wikipédia, une force d'association a_{ij} avec le titre t_j. Cette relation entre différents mots et titres s'obtient par une pondération tf-idf (équation 6.1) et s'exprime sous forme de la matrice suivante (matrice A) :

$$A = \begin{array}{c} \\ w_1 \\ w_2 \\ \vdots \\ w_i \end{array} \begin{array}{c} \begin{array}{cccc} t_1 & t_2 & \cdots & t_j \end{array} \\ \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1j} \\ a_{21} & a_{22} & \dots & a_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i1} & a_{i2} & \dots & a_{ij} \end{pmatrix} \end{array}$$

$$a_{ij} = tf_{ij} \cdot idf_i \tag{6.1}$$

Cette représentation permet donc de créer un index direct dans lequel on représente chaque titre par les mots qui le décrivent et un index inversé dans lequel on représente chaque mot par son profil d'association avec le titre de Wikipédia.

Une série de modifications destinées à améliorer les performances de ESA ont été utilisées et divulguées ultérieurement [1]. Ces modifications permettent d'écarter les articles d'administration, les listes, les articles trop courts ou avec trop peu de liens. Des poids élevés sont également attribués aux mots constituant le titre de l'article. Dans cette étude, nous avons implémenté notre propre version de ESA. Cette version prend en compte une partie des modifications apportées à la version originale. Pour une tâche d'estimation de similarité sémantique, la valeur de corrélation avec un jugement humain obtenue est de 0.72 par rapport à 0.75 rapportée par (Gabrilovich et Markovitch, 2007). La matrice de l'ESA résultante est généralement creuse, puisque la taille du dictionnaire (nombre de mots dans Wikipédia) est habituellement de l'ordre de centaines de milliers de mots et chaque titre n'est décrit que par un nombre réduit de mots distincts du dictionnaire.

L'ESA a été utilisée dans différentes applications du TAL et en recherche d'information. Dans (Radinsky *et al.*, 2011), une dimension temporelle indiquant le changement de sens d'un mot dans le temps a été ajoutée. Il considèrent que les titres ne sont pas scalaires et les représentent sous forme d'une série temporelle sur un corpus de documents temporellement ordonnées. Leur étude montre que l'ajout de cette dimension améliore les résultats de l'estimation de la similarité sémantique.

(Hassan et Mihalcea, 2011) ont introduit l'analyse sémantique saillante (SSA) qui constitue un développement de l'ESA qui repose sur la détection de concepts pertinents avant la mise en relation des mots et concepts. Cette méthode a permis d'améliorer les résultats de l'ESA en classification de textes. L'avantage principal de l'ESA est que généralement son application ne dépend pas de la langue. C'est pour cette raison qu'elle est déployée dans des contextes multilingues. Une extension de l'ESA pour la langue allemande et son application dans un système de recherche d'information crosslingue a été présentée dans (Sorg et Cimiano, 2012). De même, (Hassan et Mihalcea, 2009) se sont basés sur l'ESA dans différentes langues pour calculer la similarité entre les mots de ces langues. L'objectif principal de ces études est de créer un espace conceptuel indépendant de la langue dans lequel les documents

1. https://github.com/faraday/wikiprep-esa/wiki/roadmap

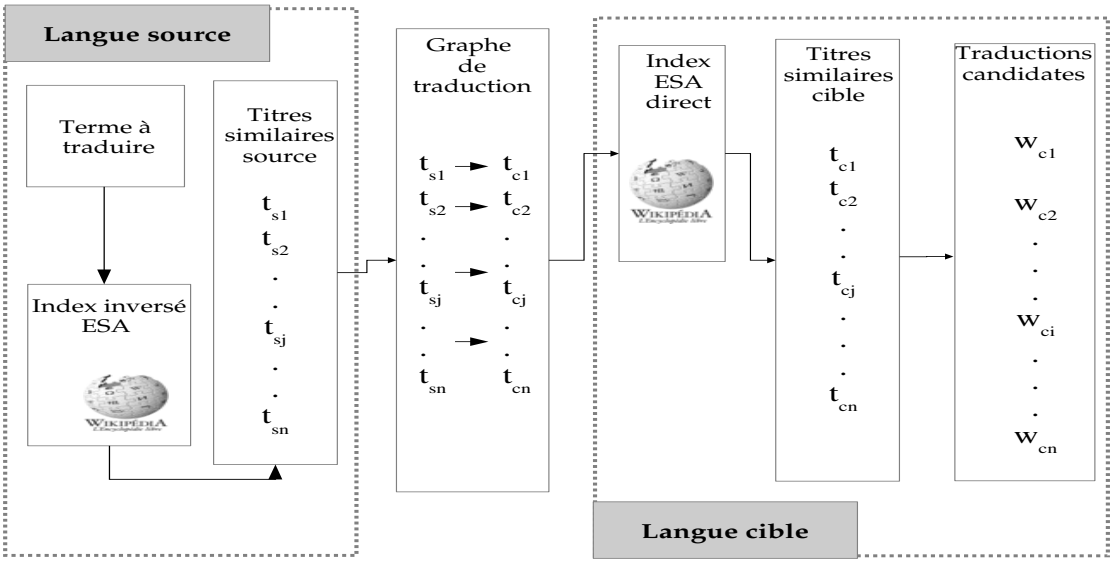


FIGURE 6.1: Aperçu général de l'approche d'extraction de lexiques bilingues avec l'ESA.

seront représentés, puis récupérés. Dans cette étude, nous utilisons également l'ESA dans un cadre multilingue mais avec une finalité différente de celles de (Hassan et Mihalcea, 2009; Sorg et Cimiano, 2012) : l'extraction de lexiques bilingues. Dans l'ESA, certains sujets liés à l'extraction lexicale comprenant la représentation des documents, l'adaptation de la méthode aux domaines de spécialité et la couverture de la ressource dans différentes langues restent ouverts.

6.3 Aperçu général de l'approche

L'objectif principal de notre approche est de mettre au point une méthode d'extraction de lexiques bilingues qui soit facilement applicable à un grand nombre de paires de langues, tout en préservant la qualité globale des résultats. Un objectif subordonné consiste à exploiter des connaissances multilingues à large échelle, telles que le contenu encyclopédique disponible dans Wikipédia. Comme il a été mentionné dans la section 6.1, l'ESA a été exploitée dans de nombreuses applications du TAL. À notre connaissance, c'est une première application de cette technique en extraction de lexiques bilingues à partir de corpus comparables.

Notre approche d'extraction de lexiques bilingues se base sur le contenu encyclopédique de Wikipédia et tente d'extraire pour chaque terme à traduire, une liste de traductions candidates. La figure 6.1 décrit l'architecture globale de cette approche qui se compose des trois étapes suivantes :

1. Tout d'abord, et contrairement à l'approche distributionnelle qui caractérise un terme à traduire par un vecteur de contexte enregistrant son environnement lexical dans un texte, notre approche le représente par les titres des articles de Wikipédia dans lesquels celui ci apparaît. Cette représentation permet de considérer aussi bien les relations syntagmatiques (par contexte) que les relations thématiques qu'entretient le terme à traduire avec les titres de Wikipédia.
2. Les titres représentant le terme à traduire dans la langue cible sont ensuite récupérés en utilisant un graphe de traduction constitué à partir des liens interlingues. Ce graphe pallie les insuffisances des dictionnaires bilingues utilisés dans l'approche distributionnelle, du fait qu'il permet ici de couvrir les langues pour lesquelles des ressources dictionnariques sont disponibles et au moins partiellement des langues peu dotées de ressources.
3. Finalement et en comparaison avec l'approche distributionnelle dans laquelle l'identification d'équivalents de traduction se base sur la comparaison des vecteurs de contexte des langues sources et cibles, dans notre approche les traductions candidates sont trouvées à travers un traitement statistique des descriptions des titres de Wikipédia de l'ESA dans la langue cible. Ces trois étapes seront détaillées dans les sections suivantes.

6.4 Représentation contextuelle

Dans cette section, nous décrivons la première étape de notre approche d'extraction de lexiques bilingues. Cette étape consiste à représenter le terme à traduire dans l'espace des titres de Wikipédia dans la langue source. À cet effet, nous proposons trois représentations qui diffèrent dans la manière dont les titres de Wikipédia sont extraits. La première consiste simplement à extraire les titres des articles de Wikipédia dans lesquels le terme à traduire apparaît. La deuxième représentation se base sur un contexte distributionnel caractérisant le terme à traduire. Nous émettons l'hypothèse qu'en considérant le contexte distributionnel dans lequel le terme apparaît, nous pou-

vons résoudre le problème d'ambiguïté. Dans la troisième représentation nous testons la combinaison des deux premières.

6.4.1 Représentation directe

L'objectif principal de cette étape est d'extraire la liste des titres Wikipédia qui représentent au mieux chaque terme à traduire. Nous utilisons simplement la matrice A introduite dans la section 6.2, pour extraire pour chaque terme à traduire la ligne indiquant son profil d'association avec les titres de Wikipédia. Ces titres sont ordonnés en fonction de leur force d'association avec le terme à traduire et sont définis par l'équation 6.2 :

$$score(t_j)_{directe} = a_{ij} \tag{6.2}$$

où a_{ij} correspond au poids tf-idf qui associe le terme à traduire w_i au titre t_j. Cette représentation peut être considérée comme une nouvelle façon d'obtenir un contexte distributionnel, faisant référence à la distribution du terme à traduire dans la totalité de Wikipédia. Contrairement à l'approche standard, dans laquelle le contexte distributionnel n'enregistre que les relations syntagmatiques qu'entretien le terme à traduire avec les différents mots du contexte, dans notre approche le contexte décrit aussi bien les relations syntagmatiques que les relations thématique qu'entretient le terme à traduire avec les différents titres de Wikipédia. Les relations syntagmatiques enregistrent la fréquence d'apparition du terme dans l'article décrivant le titre de Wikipédia. Quant aux relations thématiques, elles sont prises en compte car les articles de Wikipédia sont classés par catégorie (le système de classement thématique de Wikipédia).

6.4.2 Représentation à partir de contextes

La représentation directe par l'ESA permet d'extraire un contexte global du terme à traduire. Or, comme notre étude se porte sur des domaines de spécialité, cette représentation risque d'être ambiguë surtout lorsque le terme à traduire l'est. Par exemple, le terme *"action"* désigne en finance et en économie un placement boursier qui permet le financement des entreprises et la hauteur des cordes sur le manche d'un

instrument de musique en musique et beaucoup d'autres[2]. Par la prise en compte de tous les titres d'articles où ce terme apparaît, une ambiguïté dans la représentation et la traduction peut avoir lieu puisque les sens d'un d'un même mot sont mélangés dans la matrice de l'ESA.

Dans le but d'éviter que l'ambiguïté prenne place, nous proposons une nouvelle représentation par l'ESA, dans laquelle au lieu de représenter directement le terme à traduire, nous représentons des mots d'un contexte distributionnel de ce dernier. Ces contextes sont construits à partir d'un corpus monolingue. Le choix du corpus oriente et détermine le ou les sens majoritaires du terme dans le corpus. Pour ce faire, nous construisons un vecteur de contexte pour chaque terme à traduire et calculons une force d'association entre celui-ci et tous les éléments du vecteur de contexte résultant à l'aide d'une mesure d'association. Cette force, utilisée dans l'approche standard, est un indicateur de l'importance d'une unité lexicale vis à vis du terme à traduire. La nouvelle relation entre les différents mots du contexte et les titres de Wikipédia s'exprime sous forme de la matrice B et les titres servant à la représentation du terme à traduire sont pondérés selon l'équation 6.3.

$$B = \begin{array}{c} \\ w_1 \\ w_2 \\ \vdots \\ w_i \end{array} \begin{array}{c} \begin{array}{cccc} t_1 & t_2 & \cdots & t_j \end{array} \\ \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1j} \\ a_{21} & a_{22} & \dots & a_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i1} & a_{i2} & \dots & a_{ij} \end{pmatrix} \end{array} \begin{pmatrix} c_1 & c_2 & c_3 & \cdots & c_i \end{pmatrix}$$

$$score(t_j)_{contexte} = \sum_{i=1}^{n} a_{ij} \cdot c_i \tag{6.3}$$

où n est le nombre des mots du contexte décrivant le terme à traduire, a_{ij} est le poids tf-idf entre un mot du vecteur contexte w_i et le titre t_j et c_i dénote la force d'association entre le terme à traduire et w_i. L'utilisation de l'information contextuelle dans ce cadre sert à calculer une représentation moyenne des mots du contexte du terme à traduire dans l'espace de Wikipédia et à lever l'ambiguïté et par conséquent à extraire les titres Wikipédia les plus pertinents à la description des termes à traduire dans un contexte particulier.

2. http://fr.wikipedia.org/wiki/Action

Terme	Concepts
action	*évaluation d'action, communisme, actionnaire activiste, socialisme, dévelopement durable ...*
déficit	*crise de la dette dans la zone euro, dette publique, règle d'or budgétaire, déficit, trouble du déficit de l'attention ...*
cisaillement	*taux de cisaillement, zone de cisaillement, cisaillement, contrainte de cisaillement, viscoanalyseur ...*
turbine	*ffc turbine potsdam, turbine à gaz, turbine, turbine hydraulique, cogénération ...*
cryptage	*TEMPEST, chiffrement, liaison 16, Windows Vista, transfert de fichiers ...*
protocole	*Ad-hoc On-demand Distance Vector, protocole de Kyoto, optimized link state routing protocol, liaison 16, IPv6 ...*
biopsie	*biopsie, maladie de Horton, cancer du sein, cancer du poumon, imagerie par résonance magnétique ...*
palpation	*cancer du sein, cellulite, examen clinique, appendicite, ténosynovite ...*

TABLEAU 6.1: Les cinq titres Wikipédia les plus associés aux termes français *action, déficit, cisaillement, turbine, cryptage, biopsie* et *palpation* et leurs vecteurs de contextes.

6.4.3 Combinaison de représentations

Dans cette section, nous présentons une nouvelle représentation du terme à traduire dans laquelle nous considérons la combinaison linéaire des deux représentations précédentes. Le terme à traduire est représenté ici par les titres Wikipédia dans la langue source qui lui sont associés et ceux qui sont associés aux éléments de son vecteur de contexte. Ces titres sont par conséquent pondérés et classés en utilisant l'équation ci-dessous :

$$score(t_j) = \lambda \cdot score(t_j)_{directe} + score(t_j)_{contexte} \tag{6.4}$$

où λ est un facteur permettant de donner plus d'importance au terme à traduire par rapport à ces mots du contexte, $score(t_j)_{directe}$ est le score du titre t_j obtenu par la représentation directe par l'ESA du terme à traduire et $score(t_j)_{contexte}$ est le score du titre t_j obtenu par la représentation par l'ESA à partir de contextes.

Dans le tableau 6.1, nous présentons les cinq titres Wikipédia les plus associés au termes termes français *action, déficit, cisaillement, turbine, cryptage, biopsie* et *palpation* et leurs vecteurs de contextes. Ces termes sont issus des quatre domaines sur lesquels porte notre étude. En observant ces exemple, nous remarquons que malgré les différences entre les domaines de spécialité, notre méthode a l'avantage d'extraire des espaces conceptuels décrivant pertinemment pour chaque terme à traduire, même si celui ci est ambigu comme par exemple les termes *action* et *protocole*.

6.5 Graphe de traduction

Un graphe de traduction de concepts permettant l'extension multilingue de l'ESA a été utilisé pour traduire les concepts de la langue source vers la langue cible. Ce graphe est extrait à partir des liens de traduction explicitement disponibles dans les articles de Wikipédia. Il est exploité afin de connecter l'espace des titres d'un mot dans la langue source avec l'espace des titres lui correspondant dans la langue cible. La taille de l'espace des titres en langue cible est généralement plus petite que celui en langue source. La raison étant que seule une partie des articles peut être traduite. Par exemple, tandis que les versions française et anglaise de Wikipédia contiennent respectivement approximativement 1.4 et 4.25 millions d'article, le graphe de traduction français-anglais extrait est constitué de 940 215 paires de concepts.

6.6 Identification de traductions candidates

La troisième étape de notre approche s'effectue en langue cible, avec pour but d'identifier, pour chaque terme à traduire, une liste ordonnée de candidats à la traduction. La différence principale avec l'approche standard réside dans la façon dont les candidats à la traduction sont identifiés. Si la comparaison des contextes distributionnels sources et cibles est indispensable dans l'approche standard, dans notre approche, nous nous basons sur la représentation directe (par mots) des articles des titres de Wikipédia les plus associés au termes à traduire pour en extraire les candidats à la traduction. Nous émettons l'hypothèse que lorsque l'on représente les titres de Wikipédia les plus associés aux termes à traduire dans la langue cible, la proba-

bilité que les traductions de ces termes apparaissent dans les articles décrivant ces titre est élevée. Le passage de la langue source vers la langue cible est effectuée ici en utilisant le graphe de traduction décrit dans la section 6.5.

En utilisant ce graphe, les contextes des termes à traduire, représentés par les titres de Wikipédia sont traduits dans la langue cible. Nous utilisons ensuite les représentations de ces titres dans la langue cible (colonne de la matrice C) afin de retrouver des traductions potentielles aux termes à traduire. Ces candidats à la traduction w_{ci} sont classés en utilisant cette équation :

$$C = \begin{array}{c} \\ w_{c1} \\ w_{c2} \\ \vdots \\ w_{ci} \end{array} \begin{array}{c} \begin{array}{cccc} t_1 & t_2 & \cdots & t_j \end{array} \\ \begin{pmatrix} d_{11} & d_{12} & \dots & d_{1j} \\ d_{21} & d_{22} & \dots & d_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ d_{i1} & d_{i2} & \dots & d_{ij} \end{pmatrix} \end{array}$$

$$score(w_{ci}) = \sum_{j=1}^{n} \frac{d_{ij}}{avg(d_{kj})} \cdot log|t_j; d_{ij} \neq 0| \tag{6.5}$$

Avec d_{ij} est le poids de la traduction candidate w_{ci} pour le titre t_j de la matrice de l'ESA dans la langue cible, $avg(d_{kj})$ est la moyenne du score tf-idf des mots qui apparaissent dans t_j, $avg(d_{kj})$ est utilisé pour corriger le biais de la mesure du tf-idf envers les articles courts et $log|t_j; d_{ij} \neq 0|$ favorise les mots qui sont associés à un grand nombre de titres. La fonction log a été choisie empiriquement après avoir fait des tests avec un large éventail de fonctions. Dans nos expérimentations, nous avons considérés que les 100 titres de Wikipédia cibles les plus associés aux termes à traduire. Ce seuil a été déterminé empiriquement après avoir fait des expérimentations préliminaires.

6.7 Évaluations

Dans cette section, nous évaluons notre approche qui repose sur l'analyse sémantique explicite de Wikipédia pour extraire des lexiques bilingues à partir de corpus comparables. Comme il a été mentionné dans la section 6.1, la couverture de

Wikipédia en domaine et en langue est très variable. C'est pour cette raison que nous testons notre approche (1) sur des corpus comparables spécialisés dans des domaines variés traitant les thématiques de la *finance des entreprise*, le *cancer du sein*, *l'énergie éolienne* et la *technologie mobile*, et (2) pour les paires de langues *français-anglais*, une paire richement représentée et *roumain-anglais*, une paire qui fait intervenir le roumain, une langue peu dotée de ressources.

Nous présentons dans cette section deux séries d'expérimentations. Dans les premières expérimentations (section 6.7.1), nous évaluons notre approche en utilisant les différentes représentations présentées dans la section 6.4.La deuxième série d'expérimentations est consacrée à présenter la notion de spécificité au domaine et son impact sur nos premiers résultats (section 6.7.2).

6.7.1 Représentations contextuelle

6.7.1.1 Cadre expérimental

Nous évaluons les performances de notre approche d'extraction lexicale qui repose sur l'ESA pour représenter les termes à traduire et identifier leur équivalents de traduction. Ces termes sont représentés de trois manières. Dans la première représentation qu'on note ESA_{Dir}, les termes à traduire sont caractérisés simplement par les titres des articles de Wikipédia dans lesquels ils apparaissent. Dans la deuxième représentation notée ESA_{Cont}, les terme à traduire sont d'abord représentés par un contexte distributionnel construit d'un corpus monolingue. Ce contexte sert ensuite de base pour l'extraction des titres de Wikipédia qui lui sont associés. La troisième représentation qu'on note ESA_{Comb} n'est autre que la combinaison linéaire des deux dernières. Rappelons que dans cette méthode, les titres de Wikipédia sont classés en fonction de leurs association aux termes à traduire et aux contextes distributionnels décrivant ces termes et qu'un facteur λ est utilisé pour préviligier les termes à traduire à ceux constituant son contexte (équation 6.4). Ce facteur est défini comme suit :

$$\lambda = 10 \cdot \max(c_i) \tag{6.6}$$

où $\max(c_i)$ est la valeur d'association maximale entre le terme à traduire w_i et tous les mots de son vecteurs de contexte $w_s i$ et le facteur 10 a été défini empiriquement pour donner plus d'importance au terme à traduire w_i. La force d'association c_i est obtenue en utilisant la mesure du rapport des chances (Odds-Ratio), défini dans l'équation 6.7.

$$OddsRatio_{disc} = \log \frac{(O_{11} + \frac{1}{2})(O_{22} + \frac{1}{2})}{(O_{12} + \frac{1}{2})(O_{21} + \frac{1}{2})} \tag{6.7}$$

Le rapport des chances se base sur une table de contingence regroupant les fréquences d'observations de deux mots dans une fenêtre donnée. O_{11} représente le nombre de fois où i et j cooccurrent, O_{12} est le nombre d'occurrences de j sans i, O_{21} correspond au nombre d'occurrences de i sans j, et O_{22} est le nombre d'occurrences de deux mots autres que i et j.

Nous comparons les performances de notre approche, où les termes à traduire sont représentés de trois manières à l'approche standard, l'approche proposée par (Morin et Prochasson, 2011) et à celle présentée dans le chapitre 5, des approches basées sur l'analyse distributionnelle des termes dans un corpus de mots. Rappelons que ces approches construisent les vecteurs de contexte dans l'espace des mots du corpus, les traduits en utilisant un dictionnaire bilingue amorce et comparent les contextes distributionnels des mots de la langue source et cible pour extraire les candidats à la traduction. Lors du transfert des vecteurs de contexte, l'approche standard considère toute les traductions proposées par le dictionnaire, (Morin et Prochasson, 2011) pondèrent les traductions par leurs fréquences dans la partie cible du corpus comparable, et comme il a été détaillé dans le chapitre précédent, nous avons introduit un processus de désambiguïsation lexicale permettant d'améliorer la représentativité et l'adéquation des vecteurs de contexte dans la langue cible. Ce processus fait appel à cinq mesures de similarité sémantique, puis combine les résultats obtenus par ces dernières par le système de vote Condorcet. La configuration optimale utilisée comme référence ici est celle qui combine les résultats des mesures de similarité sémantique et considère pour chaque mot polysémique les deux traductions les plus similaires aux unités monosémiques des vecteurs de contexte (CONDORCET$_{Merge}$+WN-T$_2$) que l'on nomme WSD$_{Tech}$.

La totalité des paramètres nécessaires à l'évaluation de ces différentes méthodes ont été présentés en détail dans le chapitre 4. Ces paramètres incluent les corpus comparables, les listes de référence, les dictionnaires bilingues, les paramètres

d'expérimentation (taille de fenêtre contextuelle, mesure d'association et de similarité) et les paramètres d'évaluation : La F-mesure au $succès_1$ et $succès_{20}$ et la MAP au $succès_{20}$.

6.7.1.2 Résultats et discussion

Le tableau 6.2 décrit les résultats obtenus par notre approche et les différentes approches de l'état de l'art. Une première remarque porte sur la comparaison des méthodes de l'état de l'art. Cette comparaison montre que les performances atteintes par WSD_{Tech} au $succès_{20}$ surpassent ceux obtenus par l'approche standard (SA) et MP11 pour la paire de langues *français-anglais* et présente des performances comparables pour la paire de langues *roumain-anglais*. Pour cette raison, nous utilisons l'approche WSD_{Tech} comme référence pour discuter les résultats obtenus par l'approche introduite dans ce chapitre.

Les résultats présentés dans le tableau 6.2 montrent que la performance de notre approche qui repose sur l'ESA pour représenter les termes à traduire et extraire leurs équivalents à la traduction dépassent ceux obtenus par WSD_{Tech} pour les quatre domaines d'études et pour les deux paires de langues. Pour le domaine de *l'énergie éolienne* par exemple, ESA_{Dir} rapporte des gains de +49% et de +24% de F-mesure au $succès_1$ pour respectivement les paires de langues français-anglais et roumain-anglais. Ce constat montre que notre approche dans laquelle nous représentons les termes à traduire de trois façons dans l'espace des titres de Wikipédia est beaucoup plus efficace que les différentes approches qui représentent ces derniers dans l'espace des mots d'un corpus.

En ce qui concerne les trois manières dont les termes à traduire sont représentés, nous remarquons qu'au $succès_{20}$, ESA_{Dir} rapporte des résultats très similaires à ESA_{Comb} : la combinaison de celui ci et de ESA_{Cont} atteignant les valeurs de F-mesure maximales. Pour le français-anglais et par rapport à WSD_{Tech}, ESA_{Comb} améliore de façon importante la F-mesure au $succès_1$ et $succès_{20}$. Au $succès_{20}$, ces améliorations vont de 0.13 pour la *finance des entreprises* à 0.53 pour l'*énergie éolienne*. Concernant la paire de langues roumain-anglais, les améliorations varient de 0.03 pour la *finance des entreprises* à 0.5 pour l'*énergie éolienne*. En outre et à l'exception du domaine

a) Finance des entreprises

Méthode	français-anglais		roumain-anglais	
	$succès_1$	$succès_{20}$	$succès_1$	$succès_{20}$
AS	0.04	0.17	0.02	0.13
MP11	0.12	0.33	0.02	0.13
WSD_{Tech}	0.10	0.37	0.02	0.14
ESA_{Dir}	0.18	0.49	0.05	0.15
ESA_{Cont}	0.17	0.43	0.05	0.11
ESA_{Comb}	**0.20**	**0.50**	**0.05**	**0.17**

b) Cancer du sein

Méthode	français-anglais		roumain-anglais	
	$succès_1$	$succès_{20}$	$succès_1$	$succès_{20}$
AS	0.28	0.49	0.00	0.21
MP11	0.31	0.55	0.00	0.21
WSD_{Tech}	0.34	0.61	0.00	0.21
ESA_{Dir}	0.32	0.74	0.31	0.76
ESA_{Cont}	0.25	0.46	0.29	0.65
ESA_{Comb}	**0.36**	**0.74**	**0.31**	**0.76**

c) Énergie éolienne

Méthode	français-anglais		roumain-anglais	
	$succès_1$	$succès_{20}$	$succès_1$	$succès_{20}$
AS	0.00	0.08	0.00	0.08
MP11	0.09	0.24	0.00	0.08
WSD_{Tech}	0.03	0.30	0.00	0.08
ESA_{Dir}	0.52	0.83	0.22	0.58
ESA_{Cont}	0.49	0.79	0.14	0.4
ESA_{Comb}	**0.53**	**0.83**	**0.24**	**0.58**

d) Technologie mobile

Méthode	français-anglais		roumain-anglais	
	$succès_1$	$succès_{20}$	$succès_1$	$succès_{20}$
AS	0.01	0.06	0.01	0.16
MP11	0.01	0.05	0.00	0.16
WSD_{Tech}	0.03	0.24	0.01	0.17
ESA_{Dir}	0.39	**0.73**	0.11	0.52
ESA_{Cont}	0.37	0.64	0.09	0.46
ESA_{Comb}	**0.41**	0.72	**0.12**	**0.53**

TABLEAU 6.2: F-mesure au $succès_1$ et $succès_{20}$ des résultats d'extraction de lexiques bilingues pour quatre domaines et deux paires de langues. Trois approches de l'état de l'art sont utilisées à des fins de comparaison : AS est l'approche standard, MP11 est l'amélioration de l'AS introduite dans (Morin et Prochasson, 2011), WSD_{Tech} est l'approche présenté dans le chapitre 5. ESA_{Dir}, ESA_{Cont} et ESA_{Comb} correspondent aux trois façons dont les termes à traduire sont représentés dans notre approche.

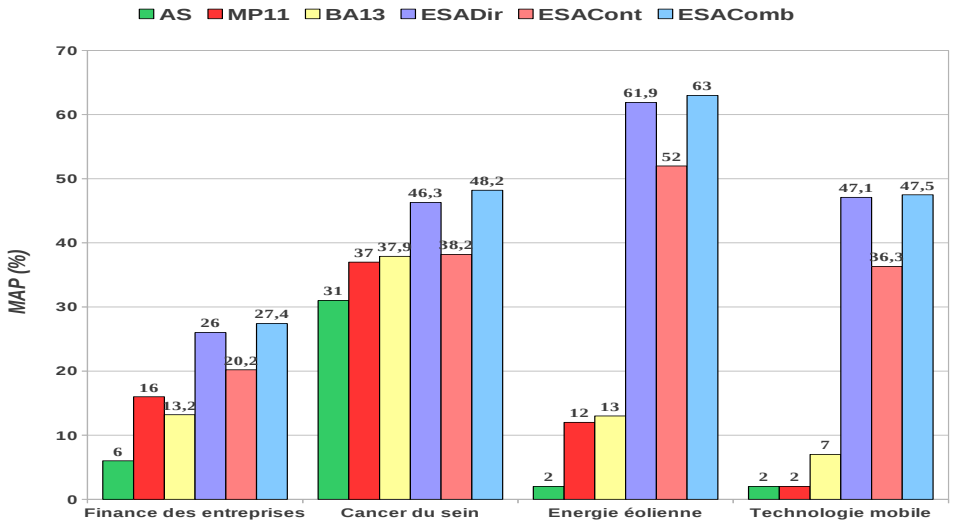
de la *finance des entreprises*, la variation de la performance est beaucoup plus faible pour ESA_{Dir}, ESA_{Cont} et ESA_{Comb} que pour la méthode WSD_{Tech}.

Méthode	Traductions candidates
ESA_{Dir}	*action, party, game, theory, direct, group, political, system, share, dividend*
ESA_{Cont}	*share, stock, action, dividend, earning, game, class, theory, social, decision*
ESA_{Comb}	*share, action, stock, party, game, dividend, use, group, system, company*

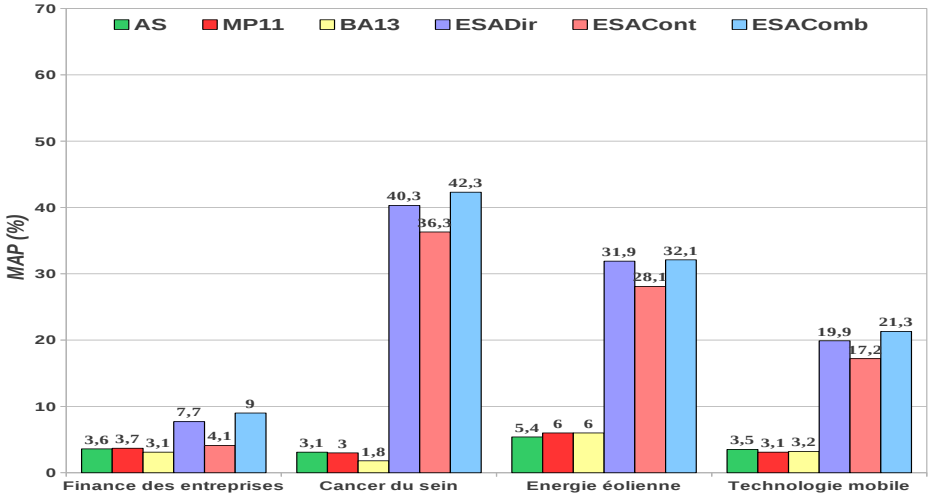
TABLEAU 6.3: Dix premières traductions candidates proposées pour le terme *action* appartenant au domaine de la finance des entreprises, par les trois représentations.

Au *succès*$_1$, c'est ESA_{Comb} qui rapporte les meilleurs résultats. Pour le domaine de la finance des entreprise par exemple, la combinaison linéaire de ESA_{Dir} et ESA_{Cont} obtient une valeur de F-mesure de 20%. Ceci montre que lorsque l'étude se porte sur des domaines de spécialité, l'ajout d'information contextuelle par ESA_{Cont} permet d'extraire les titres de Wikipédia dont la thématique est très similaire au termes à traduire et par conséquent se charge d'assurer implicitement la désambiguïsation de ces derniers. Le tableau 6.3 décrit les dix premières traductions candidates proposées par ESA_{Dir}, ESA_{Cont} et ESA_{Comb} pour terme *action* issu au domaine de la finance des entreprises. Nous constatons que dans ESA_{Dir}, la bonne traduction se trouve au neuvième rang. Par contre, lorsqu'on utilise le contexte distributionnel pour extraire les titres de Wikipédia les plus associés au terme *action*, nous remarquons que la meilleure traduction est classée au premier rang, et une traduction synonyme (*stock*) se trouve au deuxième rang. Dans ce cas, l'ajout de cette information contextuelle a permis à ESA_{Comb} d'identifier correctement la bonne traduction et de la classer au premier rang.

La MAP au *succès*$_{20}$ a été également utilisée pour évaluer les performances des différentes approches. La particularité de cette mesure est qu'elle cherche si la bonne traduction figure dans les *succès*$_{20}$, et si elle existe, elle se base sur le *rang* de la bonne traduction pour le calcul de la valeur de finale de la MAP. La figure 6.2 décrit les valeurs de MAP obtenues pour les quatre domaine et les paires de langues français-anglais (figure 6.2a) et roumain-anglais (figure 6.2b). À première vue, nous remarquons que les résultats vont dans la même direction que ceux décrit plus haut. Les meilleures valeurs est atteinte par ESA_{Comb} pour les quatre domaines et les deux paires de langues. Pour le domaine de l'énergie éolienne et la paire des langues français-anglais, ESA_{Comb} fait passer la MAP de 13% (WSD_{Tech}) à 63%. Pour la paire de



(a) Français-Anglais



(b) Roumain-Anglais

FIGURE 6.2: Valeurs de la MAP au $succès_{20}$ pour les quatre domaines et les deux paires de langues.

langues roumain-anglais, une valeur maximale de $42,3\%$ de MAP est rapportée par cette technique. Ce constat montre que l'utilisation de l'ESA pour la représentation des termes et l'extraction de candidats à la traduction est robuste au changement de domaine et par conséquent plus générique. Cette constatation vient confirmer l'hypothèse que, une utilisation adéquate d'une ressource multilingue riche comme Wikipédia est appropriée à la tâche d'extraction de lexiques bilingues spécialisés.

Bien qu'il soit plus grand que les les corpus traitant les trois autre domaines, le corpus traitant la thématique *finance des entreprises* obtient les plus faibles améliorations. Ceci est probablement dû au fait que le vocabulaire utilisé en finance des entreprises est beaucoup plus générique que les autres et par conséquent il est plus difficile d'extraire des *contextes spécialisés* dans ce cadre. Nous remarquons aussi qu'en passant du français-anglais au roumain-anglais, la baisse des performances moyennes est plus importante dans WSD_{Tech} que dans les méthodes reposant sur l'ESA.

Il est également intéressant de noter qu'aucune corrélation entre la taille du corpus spécialisé et la qualité des résultats n'a été constatée. En effet, les corpus comparables roumain-anglais sont beaucoup moins riches que ceux de la paire des langues français-anglais. Par exemple, en roumain, le corpus du domaine du cancer du sein est constitué de 22539 mots alors qu'en français et pour le même domaine la taille du corpus est de 396524 mots. Malgré ces différences, les résultats obtenus montrent que notre approche semble être insensible à ce type de paramètres.

En outre, ce qui nous a marqué le plus est que, bien qu'ils soient construits à partir de différentes sources (Wikipédia pour le domaine de la finance des entreprises et le cancer du sein et le projet TTC pour les corpus de l'énergie éolienne et de la technologie mobile, voir section 4.2 du chapitre 4), les résultats obtenus pour ces différents corpus vont dans la même direction. Ces points montrent que la présente approche est robuste à la fois au changement de domaines et de langues.

6.7.2 Spécificité au domaine

6.7.2.1 Spécificité des mots

L'ESA a été généralement utilisée dans des tâches génériques ne nécessitant pas une adaptation au domaine. Dans le cadre de notre étude, comme nous traitons des informations issues de domaines de spécialité, il conviendrait de s'intéresser à la spécificité des mots à ces derniers. La spécificité d'un mot à un domaine donné est calculée en deux étapes. Tout d'abord, nous délimitons chaque domaine et ne gardons que les articles Wikipédia qui en sont issus. Ensuite, une valeur de spécificité de tous les mots apparaissant dans ces articles est calculée.

Pour délimiter un domaine, nous partons d'une liste de titres Wikipédia amorce notée t_{seed} qui décrivent au mieux le domaine en langue cible. Par exemple, pour le domaine de la *finance des entreprises*, la page lui correspondant dans Wikipédia [3] est utilisée comme amorce pour extraire un ensemble de 10 mots (notés SW) ayant le score tf-idf le plus élevé dans cet article. Cette liste de mots est ensuite utilisée pour classer les différents titres de la langue cible en fonction de leur appartenance au domaine en utilisant l'équation 6.8.

$$Score_{dom}(t_j) = (\sum_{i=1}^{n} a_{ij} \cdot a_{it_{seed}}) * count(SW, C_t) \tag{6.8}$$

Où n est le nombre de mots amorces (10 mots) ; a_{ij} constitue le poids associé au mot w_i utilisé pour classer t_j ; $a_{it_{seed}}$ est le poids d'association d'un mot w_i dans le titre amorce du domaine t_{seed} et $count(SW, t_j)$ est le nombre de mots amorces distincts de SW qui apparaissent dans t_j. La première partie de l'équation 6.8 prend en compte les contributions des différents mots de SW qui apparaissent dans l'article du titre t_j. La deuxième partie est destinée a renforcer davantage les articles qui contiennent un plus grand nombre de mots de SW.

La délimitation de domaine est effectué en retenant les articles dont la valeur $Score_{dom}(t_j)$ est au moins égale à 1%. Ce seuil a été fixé durant des expériences préliminaires. Étant donnée la délimitation obtenue avec l'équation 6.8, nous calculons un score de spécificité au domaine $Specif_{dom}(w_i)$ pour chaque mot apparaissant dans celui ci. $Specif_{dom}(w_i)$ estime à quel point l'utilisation d'un mot dans un corpus donné est lié à un domaine cible.

$$Specif_{dom}(w_i) = \frac{DF_{dom}(w_i)}{DF_{gen}(w_i)} \tag{6.9}$$

où DF_{dom} et DF_{gen} représentent les fréquences du domaine et du document générique du mot w_i. La spécificité au domaine est utilisée dans ce cadre pour favoriser les mots ayant une valeur de spécificité élevée par rapport à ceux plus généraux surtout dans le cas où le mot est polysémique. À titre d'exemple, le terme français *action* est ambigu et se traduit en anglais par *action*, *stock*, *share* etc. Dans le domaine général, la traduction la plus fréquente est *action*. Cependant, lorsque l'étude porte sur le domaine de la *finance des entreprises*, *share* ou *stock* sont plus pertinentes. En suivant

3. https://en.wikipedia.org/wiki/Corporate_finance

cette méthode, la spécificité retournée pour les trois traductions par ordre décroissant est *share stock*, *action*. Nous proposons une nouvelle méthode notée ESA_{Spec} qui se base sur cette information pour classer et ordonner ces traductions potentielles.

Dans ESA_{Spec} nous pondérons les traductions qui figurent à la fois dans la liste de traductions proposée par ESA et dans le dictionnaire générique par leur valeur de spécificité. L'intuition qui sous-tend cette méthode est que, en rajoutant l'information sur la spécificité du terme au domaine, un nouveau classement des résultats d'extraction est envisageable.

6.7.2.2 Dictionnaire générique

Le dictionnaire générique est crée en utilisant le graphe de traduction présenté dans la section 6.5. Ce graphe est transformé en un dictionnaire bilingue en supprimant les marques d'homonymie des titres des concepts ambigus, ainsi que les liste, les catégories et d'autres pages d'administration. En outre, comme notre intérêt n'est porté que par des mots simples, nous ne retenons que les paires de traduction composées d'unigrammes dans les deux langues. Les redirections des unigrammes sont également ajoutées lorsqu'ils existent. Les dictionnaires bilingues qui en résultent se composent de $193,543$ mots simple *français-anglais* et de $136,681$ pour le *roumain-anglais*.

Nous présentons une nouvelle méthode qui se base sur ces dictionnaires pour extraire directement la ou les traductions candidates d'un terme à traduire. Cette méthode qu'on note DICO_{Spec} procède comme suit : pour chaque terme à traduire, une liste de traductions candidates sont extraites à partir du dictionnaire générique. La spécificité au domaine introduite dans la section 6.7.2 est ensuite utilisée pour pondérer et classer ces traductions candidates. Par exemple, si l'on considère le terme français *port* dérivée du corpus de la technologie mobile et désignant le système permettant aux ordinateurs de recevoir et d'émettre des informations. Le dictionnaire propose les mots anglais *port* et *seaport* comme traductions à ce terme. En calculant la spécificité de ces mots au domaine de la technologie mobile, les valeurs respectives suivantes sont renvoyées : le terme *port* obtient une valeur de 0.48 contre 0.01 pour *seaport*.

6.7.2.3 Analyse des résultats

Nous évaluons dans cette section les performances des méthodes DICO$_{Spec}$ et ESA$_{Spec}$ et les comparons à ESA$_{Comb}$, l'approche rapportant les meilleurs résultats dans les expérimentations présentées dans la section 6.7.1. Nous utilisons la F-mesure au *succès*$_1$ et *succès*$_{20}$ et la MAP au *succès*$_{20}$ pour évaluer ces méthodes. Les valeurs de F-mesures au *succès*$_1$ et *succès*$_{20}$ de l'approche de base ESA$_{Comb}$ et de ces améliorations DICO$_{spec}$ et ESA$_{spec}$ sont donnés dans le tableau 6.4. Nous constatons que les résultats obtenus par ESA$_{Comb}$ dépassent ceux de DICO$_{spec}$ et, qu'au *succès*$_{20}$ la méthode ESA$_{spec}$ constituant leur combinaison les améliore davantage. Pour les deux paires de langues, les améliorations obtenues par ESA$_{spec}$ vont jusqu'à +7% en F-mesure. Par exemple, pour le domaine du cancer du sein et la paire des langues français-anglais, ESA$_{spec}$ fait passer la F-mesure *succès*$_{20}$ de 74 à 81%. Pour le roumain-anglais, cette méthode rapporte une F-mesure de $0,24$ pour le domaine de la finance.

Pour le *succès*$_1$, les résultats varient en fonction de la paire de langues et du domaine de spécialité. Pour les deux paires de langues, l'ajout de la spécificité au domaine par ESA$_{spec}$ n'obtient des améliorations que pour le domaine de la finance des entreprises. Pour la paire des langues roumain-anglais par exemple, cette méthode réalise un gain de +6%. Ces résultats montrent que la spécifité au domaine fait rapprocher plus de traductions au *succès*$_{20}$ qu'au *succès*$_1$.

Des résultats intéressant sont aussi réalisés par DICO$_{spec}$, qui classe simplement les traductions trouvées dans un dictionnaire générique par leur valeur de spécificité au domaine. Les performances moyennes de cette méthode sont comparables à celles de WSD$_{Tech}$ en français-anglais et plus élevées pour le roumain-anglais. La présence de ce type de dictionnaire constitue un avantage principal. Néanmoins, il est clair que ces dictionnaires sont loin d'être suffisants pour couvrir tous les domaines.

Pour prendre en considération les rangs des bonnes traductions trouvés par ces différentes méthode, nous évaluons également ces deux méthodes et ESA$_{Comb}$ en utilisant la MAP. Les figures 6.3a et 6.3b illustrent les valeurs de MAP au *succès*$_{20}$ obtenus pour les quatre domaines et les deux paires de langues. Nous remarquons encore une fois que les performances de ESA$_{Comb}$ surpassent ceux de DICO$_{spec}$. Néanmoins pour la paire des langues français-anglais, la méthode ESA$_{spec}$ atteignant les meilleures

a) FE

Méthode	français-anglais		roumain-anglais	
	$succès_1$	$succès_{20}$	$succès_1$	$succès_{20}$
ESA_{Comb}	0.20	0.50	0.05	0.17
DICO_{spec}	0.06	0.20	0.05	0.11
ESA_{spec}	**0.23**	**0.56**	**0.11**	**0.24**

b) CS

Méthode	français-anglais		roumain-anglais	
	$succès_1$	$succès_{20}$	$succès_1$	$succès_{20}$
ESA_{Comb}	**0.36**	0.74	**0.31**	0.76
DICO_{spec}	0.13	0.50	0.07	0.44
ESA_{spec}	0.28	**0.81**	0.31	**0.78**

c) ÉÉ

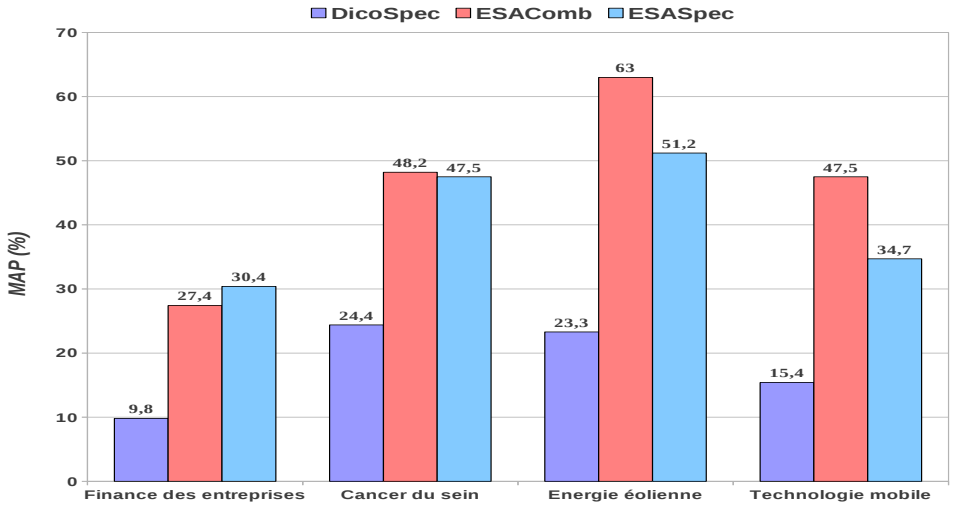
Méthode	français-anglais		roumain-anglais	
	$succès_1$	$succès_{20}$	$succès_1$	$succès_{20}$
ESA_{Comb}	**0.53**	0.83	**0.24**	**0.58**
DICO_{spec}	0.16	0.36	0.05	0.21
ESA_{spec}	0.31	**0.86**	0.18	0.58

d) TM

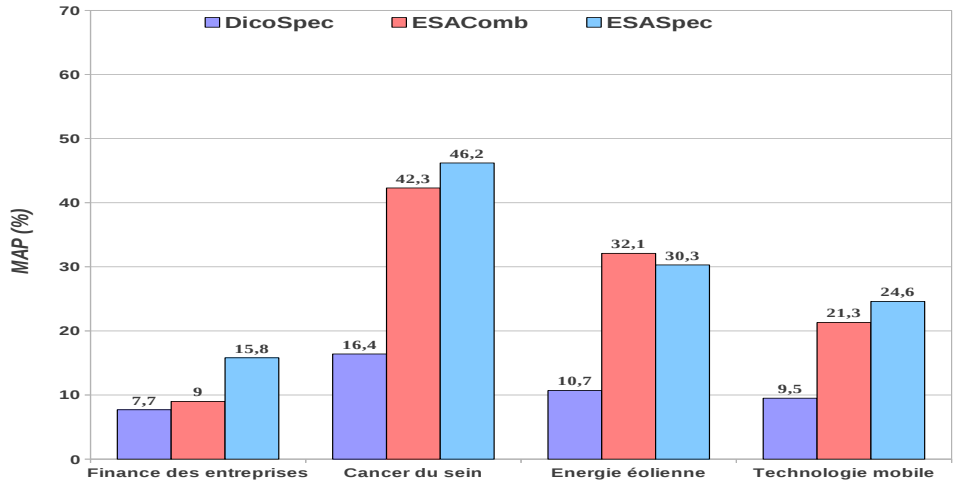
Méthode	français-anglais		roumain-anglais	
	$succès_1$	$succès_{20}$	$succès_1$	$succès_{20}$
ESA_{Comb}	**0.41**	0.72	**0.12**	0.53
DICO_{spec}	0.11	0.25	0.05	0.16
ESA_{spec}	0.19	**0.75**	0.12	**0.55**

TABLEAU 6.4: F-mesure au $succès_1$ et $succès_{20}$ des résultats d'extraction de lexiques bilingues pour quatre domaines et deux paires de langues. DICO_{spec} exploite un dictionnaire générique, combiné avec la spécificité au domaine (section 6.7.2). ESA_{Comb} est la représentation obtenant les meilleurs résultats et ESA_{spec} combine les résultats de DICO_{spec} et ESA_{Comb}. FE, CS, ÉÉ et TM désignent respectivement les domaines de la finance des entreprises, du cancer du sein, de l'énergie éolienne et de la technologie mobile.

valeurs de F-mesure au $succès_{20}$ pour tous les domaines (tableau 6.4) n'a montré son efficacité que pour le domaine de la finance des entreprise, un domaine dont le vocabulaire se rapproche le plus du domaine général. Ceci montre que la nature des corpus utilisés influence la qualité des résultats de ESA_{spec}. Dans ce domaine, cette méthode réalise un gain de +3% par rapport à ESA_{Comb}. Pour les autres domaine, la diminution du niveau de la MAP s'explique par le fait que le reclassement des traductions par l'application de la spécificité des traductions candidates au domaine, ramène dans certains cas des traductions très liées au domaine mais ne constituant pas la meilleure traduction. Une autre raison peut être liée au fait que la MAP privilégie les bonnes traductions trouvées au premier rang et que la F-mesure compte



(a) Français-Anglais



(b) Roumain-Anglais

FIGURE 6.3: Valeurs de la MAP au $succès_{20}$ pour les quatre domaines et les deux paires de langues.

simplement le nombre de bonnes traductions trouvées dans les $succès_{20}$ premiers candidats.

Pour la paire des langues roumain-anglais, ESA_{spec} améliore davantage la MAP pour les domaines de la finance des entreprises, du cancer du sein et de la technologie mobile avec des améliorations allant de $+3,3\%$ pour le domaine de la technologie mobile à $+6,8\%$ pour le domaine de la finance des entreprises. Ce constat montre

que l'ajout de la spécificité au domaine des candidats à la traduction améliore leur classement.

6.8 Conclusion

Nous avons présenté une nouvelle approche pour l'extraction de lexiques bilingues à partir de corpus comparables. L'approche proposée aborde directement les deux principaux enjeux recensés dans la section 6.1. La rareté des ressources est traitée par une exploitation adéquate de Wikipédia, une ressource disponible dans des centaines de langues. La qualité des lexiques bilingues extraits a été améliorée par une délimitation appropriée de domaine, une liaison entre les langues ainsi que par un traitement adéquat des titres les plus associés à un terme, apparaissant dans un contexte donné.

La contribution principale présentée dans ce chapitre est de proposer une méthode automatique qui exploite le contenu multilingue de Wikipédia d'une manière innovante pour améliorer les résultats d'extraction lexicale à partir de corpus comparables. Par rapport aux approches de l'état de l'art, l'avantage principal de notre approche est sa capacité à être appliquée dans de différentes langues.

Une validation expérimentale est obtenue grâce à une évaluation de cette approche sur quatre domaines différents et pour deux paires de langues. Pour le français-anglais, deux langues dont la représentation dans Wikipédia est très riche, les résultats dépassent de loin celles des approches état de l'art et surtout de l'approche standard. Au $succès_{20}$ la F-mesure obtenue est autour de 0.8 pour trois domaines sur quatre.

Pour la paire de langues roumain-anglais, une paire faisant intervenir une langue avec une représentation moins dense dans Wikipédia, les performances de notre méthode sont moindres que celles pour le français-anglais. Cependant, elles ne baissent pas avec les mêmes proportions que celles des approches de référence. Cette constatation montre que notre approche est plus générique et compte tenu de sa faible dépendance à la langue, elle peut être étendue à un grand nombre de paires de langues.

La rareté des ressources est traitée par une exploitation adéquate de Wikipédia, une ressource disponible dans des centaines de langues. La qualité des lexiques bi-

lingues extraits a été améliorée par une délimitation appropriée de domaine, une liaison entre les langues ainsi que par un traitement statistique adéquat des concepts similaire à un terme dans un contexte donné.

Conclusion et perspectives

Nous nous sommes intéressés dans cette thèse à la tâche d'extraction de lexiques bilingues à partir de corpus de textes multilingues *parallèles* et *comparables*. Nos premières contributions ont porté sur l'extraction de lexiques bilingues à partir de corpus parallèles en se concentrant sur le cas des expressions polylexicales (EPL).

Nous avons présenté une approche qui identifie les EPL et leurs équivalents à la traduction à partir d'un corpus parallèle fançais-anglais. La méthode d'identification d'EPL repose sur une approche symbolique et permet d'extraire des mots composés, des collocations, des expressions figées prépositionnelles et des entités nommées. Nous avons également proposé un algorithme d'alignement qui se base sur la comparaison des distributions des EPL sources et cibles. Les expérimentations menées dans ce cadre ont montré que notre méthode rapporte un faible niveau de rappel. La raison en est que nous avons posé l'hypothèse qu'une EPL se traduit forcément par une EPL, alors que nous avons pu identifier des cas où une EPL se traduit par un mot simple par exemple. En outre, nous considérons que les patrons morphosyntaxiques définis dans ce travail ne couvrent pas la totalité des expressions présentes dans notre corpus.

Comme il n'existe à ce jour aucun protocole commun pour l'évaluation et la comparaison des différents travaux en extraction et alignement d'EPL, nous avons mené une évaluation extrinsèque de la qualité du lexique bilingue d'EPL, où nous avons étudié l'apport de ce lexique pour un système de traduction statistique. Nous avons exploré différentes stratégies d'intégration de ces unités dans le système de traduction Moses. Deux types de stratégies ont été présentés :

– Trois stratégies d'intégration dynamiques (TRAIN, TABLE et TRAIT), où nous avons modifié le modèle de traduction de différentes façons pour une prise en considération des EPL bilingues.
– Une stratégie d'intégration statique (FORCÉ) dans laquelle nous avons incorporé ces unités sans changer le modèle de traduction.

Notre première série d'expériences a montré que la stratégie dynamique TRAIT, où un trait additionnel indique pour chaque entrée de la table de traduction s'il s'agit d'une EPL ou pas, améliore significativement les résultats obtenus par Moses avec un gain allant jusqu'à $+0,23$ points BLEU. Une évaluation lexicale fine des résultats de traduction a aussi été mise en place. Cette évaluation a montré que la plupart des stratégies d'intégration améliorent d'une manière significative les résultats de traduction et que la qualité du lexique construit automatiquement est suffisamment bonne pour améliorer la qualité de traduction statistique.

Les recherches en extraction lexicale à partir de corpus parallèles est une activité qui est passée du domaine de la recherche au domaine commercial. Cependant, et en dépit des bons résultats obtenus avec ces corpus, leur rareté en particulier pour des domaines spécialisés et pour des paires de langues peu dotées en ressources a conduit en outre à orienter les recherches en extraction de lexiques bilingues vers l'utilisation des corpus comparables. Dans la troisième partie de ce manuscrit, nous nous somme concentrés sur l'extraction de lexiques bilingues à partir de ce type de corpus. Cette tâche a été largement étudiée et les travaux menés dans ce cadre se sont appuyés sur l'approche distributionnelle. L'étude de ces méthodes distributionnelles a soulevé l'importance de la caractérisation des mots et nous a ainsi conduit à lui porter un intérêt particulier. Nos deux dernières contributions tournent autour de cette notion de représentation contextuelle.

Nous avons présenté une approche qui étend l'approche standard et vise à améliorer la représentativité des vecteurs de contexte. Cette approche aborde et vise à résoudre le problème de la *polysémie* des mots dans les vecteurs de contexte en introduisant un processus de désambiguïsation lexicale. Ce processus se base sur des mesures de similarité sémantique dérivées de WordNet pour résoudre le problème de la polysémie des mots dans les vecteurs de contexte. Les évaluations ont été menées sur des corpus comparables spécialisés dans les domaines de la *finance des entreprises, cancer du sein, énergie éolienne* et de la *technologie mobile* pour les paires

de langues français-anglais et roumain-anglais. Dans nos expérimentations, nous avons tout d'abord évalué et comparé les performances des cinq mesures de similarité sémantique. Ensuite, nous avons utilisé la méthode Condorcet, une méthode de fusion de données pour fusionner et reclasser les résultats obtenus par toutes les mesures. Nous avons constaté que la fusion de données surpasse considérablement les résultats obtenus par deux approches de l'état de l'art. Plus intéressant encore, lorsque les corpus sont fortement polysémiques, la désambiguïsation des vecteurs de contexte affecte positivement les résultats globaux. Cela montre que l'ambiguïté présente dans les corpus comparables spécialisés entrave l'extraction de lexiques bilingues.

La principale faiblesse de cette approche est qu'elle s'appuie sur WordNet. Son application dépend donc de l'existence de cette ressource dans la langue cible. C'est la raison pour laquelle nous avons proposé une nouvelle approche qui résout implicitement le problème d'ambiguïté des mots. Cette approche considère Wikipédia comme un corpus comparable et se base sur l'analyse explicite sémantique pour la représentation des termes à traduire et l'extraction de leurs équivalents de traduction. Au lieu de représenter les termes à traduire par analyse distributionnelle, qui consiste à les représenter dans l'espace des *mots du corpus*, dans notre approche, ces termes sont représentés dans l'espace des *titres de Wikipédia*. Cette représentation permet de considérer aussi bien les relations syntagmatiques (par contexte) et les relations thématique qu'entretien le terme à traduire avec les titres de Wikipédia. En outre, pour passer de la représentation d'un terme de la langue source vers sa représentation dans la langue cible, nous remplaçons l'usage du dictionnaire bilingue par les liens interlingues de Wikipédia. Ceci permet de traiter aussi bien des langues pour lesquelles des ressources dictionnariques sont disponibles et des langues peu dotées de ressources. Finalement et en comparaison avec l'approche distributionnelle dans laquelle l'identification d'équivalents de traduction se base sur la comparaison des vecteurs de contexte des langues sources et cibles, dans notre approche les traductions candidates sont trouvées à travers un traitement statistique des descriptions des titres de Wikipédia de l'ESA dans la langue cible.

Une validation expérimentale est obtenue grâce à une évaluation de cette approche sur quatre domaines différents et pour deux paires de langues. Pour le français-anglais, deux langues dont la représentation dans Wikipédia est très riche, les résultats dépassent de loin celles des approches état de l'art et surtout de l'approche standard. Au $succès_{20}$ la F-mesure obtenue est autour de 0.8 pour trois domaines sur quatre.

Pour la paire de langues roumain-anglais, une paire faisant intervenir une langue avec une représentation moins dense dans Wikipédia, les performances de notre méthode sont moindres que celles pour le français-anglais. Cependant, elles ne baissent pas avec les mêmes proportions que celles des approches de référence. Cette constatation montre que cette approche est plus générique et compte tenu de sa faible dépendance à la langue, elle peut être étendue à un grand nombre de paires de langues.

Perspectives

Les résultats obtenus tout au long de cette thèse sont encourageants et peuvent être améliorés de différentes façons. Nous envisageons tout d'abord d'améliorer notre approche d'extraction de lexique bilingue d'EPL, qui se base sur une liste prédéfinie de patrons morphosyntaxiques pour l'identification de ces unités. Au lieu de les définir manuellement pour chaque langue, nous souhaitons apprendre l'ensemble des patrons morphosyntaxiques à partir de larges corpus monolingues. Ceci permettra de couvrir un large éventail d'EPL dans différentes langues.

Puis, nous comptons également porter quelques améliorations à notre application du lexiques bilingue d'EPL au système de traduction statistique moses. Dans notre approche, le modèle de traduction est estimé sur des lemmes plutôt que sur des formes de surface. Nous avons d'abord l'intention d'utiliser un modèle de génération pour produire les formes de surfaces adéquates à partir des résultats de traduction, présentés ici en lemmes. Nous comptons ensuite entraîner notre système de traduction sur un corpus de taille plus importante afin d'évaluer l'impact du volume des données sur les résultats obtenus. Rappelons également que dans la stratégie Table, les valeurs de probabilité de traduction sont représentées par l'indice de Jaccard. Il semble intéressant de transformer les valeurs de l'indice de Jaccard en une probabilité de traduction. Cet ajustement pourrait assurer l'uniformité et la cohérence des probabilités dans la table de traduction et par conséquent améliorer non seulement les résultats obtenus par la stratégie Table, mais aussi la stratégie Trait. En plus de leur application dans un système de TAS, nous tenterons d'étudier l'impact de ces EPLs sur la pertinence des résultats d'un système de recherche d'informations ou un système de traduction à base de règles.

Nos deux approches d'extraction de lexiques bilingues à partir de corpus comparables peuvent également être améliorées. Ces deux approches n'ont été évaluées que

pour deux paires de langues (français-anglais et roumain-anglais). Nous envisageons tester ces deux approches pour d'autres paires de langues. Si pour la première approche, qui propose d'ajouter un processus de désambiguïsation lexicale à l'approche standard, cette tache s'avère être difficile du fait qu'elle soit très dépendante de plusieurs ressources dont les dictionnaires bilingues, qui ne sont pas forcément disponibles pour toutes les langues, son application est possible dans notre deuxième approche reposant sur l'analyse sémantique explicite et Wikipédia. Elle est possible dans la mesure ou la paire de langues que l'on souhaite traitée soit couverte par Wikipédia.

Finalement, nos approches d'extraction lexicale à partir de corpus comparables ne tiennent pas compte des termes composés et se limitent à l'extraction bilingue de termes simples. Nous envisageons appliquer nos approches aux termes complexes, surtout qu'en domaine de spécialité, la plupart des entrées d'un lexiques bilingue se composent de termes composés. Une version de l'ESA qui divise les texte dans des mots composés plutôt que des mots simples est actuellement développée et sera testée pour l'extraction de lexiques bilingues.

Publications

Les idées et les résultats présentés dans cette thèse ont déjà été publiés dans les articles suivants :

- Dhouha Bouamor, Adrian Popescu, Nasredine Semmar, Pierre Zweigenbaum, "Using Large-Scale Background Knowledge for Bilingual Lexicon Extraction from Comparable Corpora" *EMNLP 2013. Conference on Empirical Methods in Natural Language Processing*, Seattle, USA : 2013.
- Dhouha Bouamor, Nasredine Semmar, Pierre Zweigenbaum, "A Generic Approach for Bilingual Lexicon Extraction from Comparable Corpora" *MT SUMMIT 2013. Machine Translation Summit*, Nice, France : 2013.
- Dhouha Bouamor, Nasredine Semmar, Pierre Zweigenbaum,"Context Vector Disambiguation for Bilingual Lexicon Extraction from Comparable Corpora" *ACL-HLT 2013. The 51th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, Sofia, Bulgaria : 2013.
- Dhouha Bouamor, Nasredine Semmar, Pierre Zweigenbaum, "Using WordNet and Semantic Similarity for Bilingual Terminology Mining from Comparable Corpora" *BUCC 2013. The ACL 2013 Workshop on Building and Using Comparable Corpora*, Sofia, Bulgaria : 2013.
- Dhouha Bouamor, Nasredine Semmar, Pierre Zweigenbaum, "Utilisation de la similarité sémantique pour l'extraction de lexiques bilingues à partir de corpus comparables" *TALN 2013. 20ème Conférence sur le Traitement Automatique des Langues Naturelles*, Les Sables d'Olonne, France : 2013.
- Dhouha Bouamor, Nasredine Semmar, Pierre Zweigenbaum, "Acquisition de lexique bilingue d'expressions polylexicales : Une application à la traduction statistique" *RECITAL 2013. 15ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, Les Sables d'Olonne, France : 2013.

– Dhouha Bouamor, Nasredine Semmar, Pierre Zweigenbaum, "A Study on Using Arabic-English Multiword Expressions for Statistical Machine Translation", *In International journal of Computational and General Linguistics : LINGUISTICA COMMUNICATIO, Vol. 5, 2013.*.
– Dhouha Bouamor, Nasredine Semmar, Pierre Zweigenbaum. "Automatic Construction of a Bilingual Lexicon of Multiword Expressions : A Statistical Machine Translation Evaluation Perspective".*In Proceedings of the 2012 joint COLING Workshop on Gognitive Aspects of the Lexicon*, Mumbai, India : 2012.
– Dhouha Bouamor, Nasredine Semmar, Pierre Zweigenbaum, "Identifying Bilingual Multi-word Expressions for Statistical Machine Translation" *LREC 2012. The 8th international conference on Language Resources and Evaluation*, Istanbul, Turkey : 2012.
– Dhouha Bouamor, Nasredine Semmar, Pierre Zweigenbaum. "A Study on Using Arabic-English Multiword Expressions for Statistical Machine Translation". *CITALA 2012. In Proceedings of the IEEE techniqually sponsored 4th International Conference on Arabic Language Processing*, Rabat, Maroc : 2012.
– Dhouha Bouamor, Nasredine Semmar, Pierre Zweigenbaum."Using Bilingual Multiword Expressions to Improve Statistical Machine Translation". *LIHMT-2011. In Proceedings of the International Workshop on Using Linguistic Information for Hybrid Machine Translation* , Barcelona, Spain : 2011.

Bibliographie

Abdul-Rauf, S. et Schwenk, H. (2009). On the use of comparable corpora to improve smt performance. *In EACL*, pages 16–23.

Allauzen, A. et Wisniewski, G. (2009). Modèles discriminants pour l'alignement mot à mot. *Traitement Automatique des Langues*, 50(3):173–203.

Andrade, D., Nasukawa, T. et Tsujii, J. (2010). Robust measurement and comparison of context similarity for finding translation pairs. *In Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 19–27. Association for Computational Linguistics.

Apidianaki, M., Ljubešić, N. et Fišer, D. (2013). Cross-lingual wsd for translation extraction from comparable corpora. *In Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 1–10, Sofia, Bulgaria. Association for Computational Linguistics.

Babych, B. et Hartley, A. (2004). Modelling legitimate translation variation for automatic evaluation of MT quality. *In LREC 2004*.

Baldwin, T., Bannard, C., Tanaka, T. et Widdows, D. (2003). An empirical model of multiword expression decomposability. *In proceedings of the ACL-SIGLEX workshop on multiword expressions : analysis, acquisition and treatment*, pages 89–96.

Banerjee, S. et Pedersen, T. (2002). An adapted lesk algorithm for word sense disambiguation using wordnet. *In Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '02, pages 136–145, London, UK, UK. Springer-Verlag.

Besançon, R., De Chalendar, G., Ferret, O., Gara, F., Laib, M., Mesnard, O. et Semmar, N. (2010). Lima : A multilingual framework for linguistic analysis and linguistic resources development and evaluation. *In Proceedings of LREC*, Malta.

Bowker, L. et Pearson, J. (2002). *Working with Specialized Language*. Routledge (Taylor and Francis), New York.

Brown, P., Della Pietra, S., Della Pietra, V. et Mercer, R. (1993). The mathematics of statistical machine translation : Parameter estimation. *In Computational linguistics*.

Brown, P. F., Lai, J. C. et Mercer, R. L. (1991). Aligning sentences in parallel corpora. *In Proceedings of the 29th annual meeting on Association for Computational Linguistics*, ACL'91, pages 169–176, Stroudsburg, PA, USA. Association for Computational Linguistics.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C. et Schroeder, J. (2008). Further meta-evaluation of machine translation. *In Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08, pages 70–106. Association for Computational Linguistics.

Calzolari, N., Fillmore, C. J., Grishman, R., Ide, N., Lenci, A., Macleod, C. et Zampolli, A. (2002). Towards best practice for multiword expressions in computational lexicons. *In LREC*. European Language Resources Association.

Carpuat, M. et Diab, M. (2010). Task-based evaluation of multiword expressions :a pilot study in statistical machine translation. *In Proceedings of HLT-NAACL*.

Chen, S. F. (1993). Aligning sentences in bilingual corpora using lexical information.

Chiao, Y.-C. et Zweigenbaum, P. (2002). Looking for candidate translational equivalents in specialized, comparable corpora. *In Proceedings of the 19th international conference on Computational linguistics - Volume 2*, COLING '02, pages 1–5. Association for Computational Linguistics.

Chiao, Y.-C. et Zweigenbaum, P. (2003). The effect of a general lexicon in corpus-based identification of french-english medical word translations. *In Proceedings Medical Informatics Europe, volume 95 of Studies in Health Technology and Informatics*, pages 397–402, Amsterdam.

Cho, M., Choi, C., Kim, H., Shin, J. et Kim, P. (2007). Efficient image retrieval using conceptualization of annotated images. *In MCAM 2007*, numéro 4577 de LNCS, pages 426–433.

Choi, D., Kim, J., Kim, H., Hwang, M. et Kim, P. (2012). A method for enhancing image retrieval based on annotation using modified wup similarity in wordnet. *In Proceedings of the 11th WSEAS international conference on Artificial Intelligence, Knowledge Engineering and Data Bases*, AIKED'12, pages 83–87, Stevens Point, Wisconsin, USA. World Scientific and Engineering Academy and Society (WSEAS).

Constant, M., Tellier, I., Duchier, D., Dupont, Y., Sigogne, A., Billot, S. *et al.* (2011). Intégrer des connaissances linguistiques dans un crf : application à l'apprentissage d'un segmenteur-étiqueteur du français. *In Actes de TALN*, Montpellier, France.

Cook, P., Fazly, A. et Stevenson, S. (2007). Pulling their weight : exploiting syntactic forms for the automatic identification of idiomatic expressions in context. *In Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, MWE '07, pages 41–48, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dagan, I. et Church, K. (1994). Termight : Identifying and translating technical terminology. *In Proceedings of the 4th Conference on ANLP*, pages 34–40, Stuttgart, Germany.

Daille, B. (2001). Extraction de collocation à partir de textes. *In* Maurel, D., éditeur : *Actes de TALN 2001 (Traitement automatique des langues naturelles)*, Tours. ATALA, Université de Tours.

Daille, B. (2002). Terminology mining. *In SCIE*, pages 29–44.

de Groc, C. (2011). Babouk : Focused web crawling for corpus compilation and automatic terminology extraction. *In Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '11, pages 497–498, Washington, DC, USA. IEEE Computer Society.

Déjean, H. et Gaussier, E. (2002). Une nouvelle approche à l'extraction de lexique bilingues à partir de corpus comparables. *Lexicometrica*, 19(4):1–22.

Déjean, H., Gaussier, E. 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 - Volume 1*, COLING '02, pages 1–7. Association for Computational Linguistics.

Deleger, L., Namer, F. et Zweigenbaum, P. (2009). Morphosemantic parsing of medical compound words : Transferring a French analyzer to English. *International Journal of Medical Informatics*, 78(1):48–55.

DeNero, J. et Klein, D. (2010). Discriminative modeling of extraction sets for machine translation. *In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1453–1463, Stroudsburg, PA, USA. Association for Computational Linguistics.

Diab, M. et Finch, S. (2000). A statistical word-level translation model for comparable corpora. *In Proceedings of the Conference on Content-based multimedia information access (RIAO)*.

Dorow, B., Laws, F., Michelbacher, L., Scheible, C. et Utt, J. (2009). A graph-theoretic algorithm for automatic extension of translation lexicons. *In Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, GEMS '09, pages 91–95, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fellbaum, C. (1998). *WordNet : An Electronic Lexical Database*. Bradford Books.

Frantzi, C., Ananiadou, S. et Mima, H. (2000). Automatic recognition of multi-word terms : the c-value/nc-value method. *In Int. J. on Digital Libraries 3(2)*, pages 115–130.

Fung, P. (1995). Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. *In Proceedings of the Third Workshop on Very Large Corpora*, pages 173–183.

Fung, P. (1998). A statistical view on bilingual lexicon extraction : From parallel corpora to non-parallel corpora. *In Parallel Text Processing*, pages 1–17. Springer.

Fung, P. et Yee, L. Y. (1998). An ir approach for translating new words from nonparallel, comparable texts. *In Proceedings of the 36th Annual Meeting of the*

Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1, ACL '98, pages 414–420. Association for Computational Linguistics.

Gabrilovich, E. et Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. *In Proceedings of the 20th international joint conference on Artifical intelligence*, IJCAI'07, pages 1606–1611, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Gale, W. A. et Church, K. W. (1991). A program for aligning sentences in bilingual corpora. *In Proceedings of the 29th annual meeting on Association for Computational Linguistics*, ACL '91, pages 177–184, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gamallo, O. (2007). Learning bilingual lexicons from comparable english and spanish corpora. *In Proceedings of MT SUMMIT*, pages 191–198.

Gamallo, O. (2008). Evaluating two different methods for the task of extracting bilingual lexicons from comparable corpora. *In Proceedings of LREC 2008 Workshop on Comparable Corpora*, pages 19–26.

Gamallo, P. et Garcia, M. (2012). Extraction of bilingual cognates from wikipedia. *In Proceedings of the 10th international conference on Computational Processing of the Portuguese Language*, PROPOR'12, pages 63–72, Berlin, Heidelberg. Springer-Verlag.

Garera, N., Callison-Burch, C. et Yarowsky, D. (2009). Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences. *In Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, pages 129–137, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gaussier, É., Renders, J.-M., Matveeva, I., Goutte, C. et Déjean, H. (2004). A geometric view on bilingual lexicon extraction from comparable corpora. *In ACL*, pages 526–533.

Gaussier, É. et Yvon, F. (2011). *Modèles statistiques pour l'accès à l'information textuelle*. HERMÈS / LAVOISIER. Collection : Recherche d'information et web.

Germann, U. (2008). Yawat : Yet another word alignment tool. *In ACL (Demo Papers)*, pages 20–23. The Association for Computer Linguistics.

Haghighi, A., Blitzer, J., DeNero, J. et Klein, D. (2009). Better word alignments with supervised itg models. *In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 923–931, Suntec, Singapore. Association for Computational Linguistics.

Haghighi, A., Liang, P., Berg-kirkpatrick, T. et Klein, D. (2008). Learning bilingual lexicons from monolingual corpora.

Halavais, A. et Lackaffb, D. (2008). An Analysis of Topical Coverage of Wikipedia. *Journal of Computer-Mediated Communication*, 13(2):429–440.

Harris, Z. (1954). Distributional structure. *Word*.

Hassan, S. et Mihalcea, R. (2009). Cross-lingual semantic relatedness using encyclopedic knowledge. *In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 3 - Volume 3*, EMNLP '09, pages 1192–1201, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hassan, S. et Mihalcea, R. (2011). Semantic relatedness using salient semantic analysis. *In AAAI*.

Hazem, A. et Morin, E. (2012a). Adaptive dictionary for bilingual lexicon extraction from comparable corpora. *In Proceedings, 8th international conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey.

Hazem, A. et Morin, E. (2012b). Qalign :a new method for bilingual lexicon extraction from comparable corpora. *In Proceedings of CICLING*, India.

Hazem, A. et Morin, E. (2013). Extraction de lexiques bilingues à partir de corpus comparables par combinaison de représentations contextuelles. *In Actes de TALN 2013 (Traitement automatique des langues naturelles)*, Les Sables d'Olonne. ATALA, LINA.

Hogan, D., Cafferkey, C., Cahill, A. et van Genabith, J. (2007). Exploiting multi-word units in history-based probabilistic generation. *In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 267–276, Prague, Czech Republic. Association for Computational Linguistics.

Hwang, M., Choi, C. et Kim, P. (2011). Automatic enrichment of semantic relation network and its application to word sense disambiguation. *IEEE Transactions on Knowledge and Data Engineering*, 23:845–858.

Jackendoff, R. (1997). The architecture of the language faculty. *MIT Press*.

Ji, H. (2009). Mining name translations from comparable corpora by creating bilingual information networks. *In Proceedings of the 2nd Workshop on Building and Using Comparable Corpora : from Parallel to Non-parallel Corpora*, BUCC '09, pages 34–37, Stroudsburg, PA, USA. Association for Computational Linguistics.

Katz, G. et Giesbrecht, E. (2006). Automatic identification of non-compositional multi-word expressions using latent semantic analysis. *In Proceedings of the Workshop on Multiword Expressions : Identifying and Exploiting Underlying Properties*, MWE '06, pages 12–19, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kay, M. et Röscheisen, M. (1993). Text-translation alignment. *Computational Linguistics*, 19(1):121–142.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. *In Proceedings of EMNLP*.

Koehn, P. (2005). Europarl : A parallel corpus for statistical machine translation. *In Proceedings of MT-SUMMIT*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. et Herbst, E. (2007). Moses : Open source toolkit for statistical machine translation. *In Proceedings of ACL, demo session*, Prague, Czech Republic.

Koehn, P. et Knight, K. (2000). Estimating word translation probabilities from unrelated monolingual corpora using the em algorithm. *In Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 711–715. AAAI Press.

Koehn, P. et Knight, K. (2001). Knowledge sources for word-level translation models. *In Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 27–35.

Koehn, P., Och, F. et Marcu, D. (2003). Statistical phrase-based translation. *In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 115–124, Edmonton, Canada.

Kun, Y. et Tsujii, J. (2009). Bilingual dictionary extraction from wikipédia. *In Proceedings of MT SUMMIT*.

Kupiec, J. (1993). An algorithm for finding noun phrases correspondences in bilingual corpora. *In Proceedings of the 31st annual Meeting of the Association for Computational Linguistics*, pages 17–22, Columbus, Ohio, USA.

Lambert, P. et Banchs, R. (2005). Data inferred multi-word expressions for statistical machine translation. *In Proceedings of MT SUMMIT*.

Lambert, P. et Banchs, R. (2006). Grouping multi-word expressions according to part-of-speech in statistical machine translation. *In Proceedings of the Workshop on Multi-word Expressions in a multilingual context*.

Lardilleux, A. (2010). *Contribution des basses fréquences à l'alignement sous-phrastique multilingue : une approche différentielle*. These, Université de Caen.

Laroche, A. et Langlais, P. (2010). Revisiting context-based projection methods for term-translation spotting in comparable corpora. *In 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 617–625, Beijing, China.

Leacock, C. et Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. *In WordNet : An Electronic Lexical Database*, pages 305–332. MIT Press.

Leturia, I., San Vicente, I. et Saralegi, X. (2009). Search engine based approaches for collecting domain-specific basque-english comparable corpora from the internet. *In 5th International Web as Corpus Workshop (WAC5)*, Donostia-San Sebastian, Spain.

L'Homme, M.-C. (2004). *La terminologie : Principes et techniques*. La presse de l'université de Montréal.

Li, B. et Gaussier, É. (2010). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. *In 23rd International Conference on Computational Linguistics (Coling 2010)*, Beijing, China.

Lin, D. (1998). An information-theoretic definition of similarity. *In Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Lopez, A. et Resnik, P. (2006). Word-based alignment, phrase based translation : what's the link ? *In Proceedings of the association for machine translation in the Americas : visions for the future of machine translation*, pages 90–99.

Manning, C. D., Raghavan, P. et Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Melamed, I. D. (1996). A geometric approach to mapping bitext correspondence. *In Conference on Empirical Methods in Natural Language Processing*, pages 1–12.

Melamed, I. D. (1999). Bitext maps and alignment via pattern recognition. *Computational Linguistics*, 25(1):107–130.

Miller, G. A. et Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Moirón, B. V. et Tiedemann, J. (2006). Identifying idiomatic expressions using automatic word alignment. *In Proceedings of the EACL 2006 Workshop on Multiword Expressions in.*

Montague, M. et Aslam, J. A. (2002). Condorcet fusion for improved retrieval. *In Proceedings of the eleventh international conference on Information and knowledge management*, CIKM '02, pages 538–548, New York, NY, USA. ACM.

Moore, R. C. (2002). Fast and accurate sentence alignment of bilingual corpora. *In Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation : From Research to Real Users*, AMTA '02, pages 135–144, London, UK, UK. Springer-Verlag.

Morin, E. et Daille, B. (2006). Comparabilité de corpus et fouille terminologique multilingue. *In Traitement Automatique des Langues (TAL).*

Morin, E., Daille, B., Takeuchi, K. et Kageura, K. (2008). Brains, not brawn : The use of smart comparable corpora in bilingual terminology mining. *ACM Trans. Speech Lang. Process.*, 7(1):1 :1–1 :23.

Morin, E. et Prochasson, E. (2011). Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. *In Proceedings, 4th Workshop on Building and Using Comparable Corpora (BUCC)*, page 27–34, Portland, Oregon, USA.

Nivre, J. et Nilsson, J. (2004). Multiword units in syntactic parsing. *In Workshop on Methodologies and Evaluation of Multiword Units in Real-World Applications.*

Nuray, R. et Can, F. (2006). Automatic ranking of information retrieval systems using data fusion. *Inf. Process. Manage.*, 42(3):595–614.

Och, F.-J. (2003). Minimum error rate training in statistical machine translation. *In Proceedings of ACL*.

Och, F. J. et Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Okita, T., Guerra, M., Alfredo Graham, Y. et Way, A. (2010). Multi-word expression sensitive word alignment. *In Proceedings of the 4th International Workshop on Cross Lingual Information Access at COLING 2010*, pages 26–34, Beijing.

Ozdowska, S. (2006). *ALIBI, un système d'ALIgnement BIlingue à base de règles de propapagation syntaxique*. These, Université de Toulouse-Le Mirail.

Papineni, k., Roukos, S., Ward, T. et Zhu, W. J. (2002). Bleu : a method for automatic evaluation of machine translation. *In 40th Annual meeting of the Association for Computational Linguistics*.

Patwardhan, S. (2003). Incorporating Dictionary and Corpus Information into a Context Vector Measure of Semantic Relatedness. Mémoire de D.E.A., University of Minnesota, Duluth.

Pecina, P. (2008). *Lexical Association Measures : Collocation Extraction*. Thèse de doctorat, Faculty of Mathematics and Physics, Charles University in Prague, Prague, Czech Republic.

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.

Piao, S. S., Rayson, P., Archer, D. et McEnery, T. (2005). Comparing and combining a semantic tagger and a statistical tool for {MWE} extraction. *Computer Speech and Language*, 19(4):378 – 397. Special issue on Multiword Expression.

Prochasson, E. (2009). *Alignement multilingue en corpus comparables spécialisés*. These, Université de Nantes.

Prochasson, E. et Fung, P. (2011). Rare word translation extraction from aligned comparable documents. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies - Volume 1*, HLT '11, pages 1327–1335, Stroudsburg, PA, USA. Association for Computational Linguistics.

Prochasson, E., Morin, E. et Kageura, K. (2009). Anchor points for bilingual lexicon extraction from small comparable corpora. *In Proceedings, 12th Conference on Machine Translation Summit (MT Summit XII)*, page 284–291, Ottawa, Ontario, Canada.

Radinsky, K., Agichtein, E., Gabrilovich, E. et Markovitch, S. (2011). A word at a time : computing word relatedness using temporal semantic analysis. *In Proceedings of the 20th international conference on World wide web*, WWW '11, pages 337–346, New York, NY, USA. ACM.

Ramisch, C., Villavicencio, A. et Kordoni, V. (2013). Introduction to the special issue on multiword expressions : From theory to practice and use. *ACM Trans. Speech Lang. Process.*, 10(2):3 :1–3 :10.

Rapp, R. (1995). Identifying word translations in non-parallel texts. *In Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, ACL '95, pages 320–322. Association for Computational Linguistics.

Rapp, R. (1999a). Automatic identification of word translations from unrelated english and german corpora. *In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 519–526. Association for Computational Linguistics.

Rapp, R. (1999b). Automatic identification of word translations from unrelated english and german corpora. *In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 519–526. Association for Computational Linguistics.

Rapp, R., Sharoff, S. et Babych, B. (2012). Identifying word translations from comparable documents without a seed lexicon. *In* Chair), N. C. C., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J. et Piperidis, S., éditeurs : *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Ren, Z., Lu, Y., Liu, Q. et Huang, Y. (2009). Improving statistical machine translation using domain bilingual multiword expressions. *In Proceedings of the Workshop on Multiword Expressions : Identification, Interpretation, Disambiguation and Applications*, pages 47–57.

Robitaille, X., Sasaki, Y., Tonoike, M., Sato, S. et Utsuro, T. (2006). Compiling french-japanese terminologies from the web. *In EACL'06*.

Rubino, R. et Linarès, G. (2011). A multi-view approach for term translation spotting. *In Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, pages 29–40.

Sadat, F. et Terrasa, A. (2010). Exploitation de wikipédia pour l'enrichissement et la construction des ressources linguistiques. *In Proceedings of TALN*, Montréal, Canada.

Sag, I., Baldwin, T., Francis Bond, F., Copestake, A. et Flickinger, D. (2002). Multiword expressions :a pain in the neck for nlp. *In CICLing 2002*, Mexico City, Mexico.

Sager, J. C. (1990). *A Practical Course in Terminology Processing*. John Benjamins, Amsterdam/Philadelphia.

Sagot, B., Clément, L., De La Clergerie, É., Boullier, P. *et al.* (2005). Vers un méta-lexique pour le français : architecture, acquisition, utilisation. *In Actes de TALN*.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. *In Proceedings of the ACL SIGDAT-Workshop*, pages 47–50.

Seco, N., Veale, T. et Hayes, J. (2004). An Intrinsic Information Content Metric for Semantic Similarity in WordNet. *In ECAI'2004, the 16th European Conference on Artificial Intelligence*.

Semmar, N., Servan, C., de Chalendar, G., le Ny, B. et Bouzaglou, J.-J. (2010). A hybrid word alignment approach to improve translation lexicons with compound words and idiomatic expressions. *In ASLIB Translating and the Computer Conference*, London (UK).

Seretan, V. et Wehrli, E. (2007). Collocation translation based on sentence alignment and parsing. *In* Benarmara, F., Hatout, N., Muller, P. et Ozdowska, S., éditeurs : *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

Shezaf, D. et Rappoport, A. (2010). Bilingual lexicon generation using non-aligned signatures. *In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 98–107. Association for Computational Linguistics.

Simionescu, R. (2011). Hybrid pos tagger. *In Proceedings of "Language Resources and Tools with Industrial Applications" Workshop (Eurolan 2011 summerschool)*.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.

Smadja, F., McKeown, K. et Hatzivassiloglou, V. (1996). Translating collocations for bilingual lexicons : A statistical approach. *In Computational Linguistics*, pages 1–38.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. et Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. *In Proceedings of Association for Machine Translation in the Americas*.

Somers, H. (2001). Bilingual parallel corpora and language engineering. *In Proceedings of workshop on language engineering for south-asian languages*.

Sorg, P. et Cimiano, P. (2012). Exploiting wikipedia for cross-lingual and multilingual information retrieval. *Data Knowl. Eng.*, 74:26–45.

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. Association for Computational Linguistics.

Tomlinson, S. (2012). Measuring robustness with frst relevant score in the trec 2012 microblog track. *In The Twenty-First Text REtrieval Conference Proceedings (TREC)*.

Tufis, I. et Ion, R. (2007). Parallel corpora, alignment technologies and further prospects in multilingual resources and technology infrastructure. *In Proceedings of the 4th International Conference on Speech and Dialogue Systems*, pages 183–195.

Tutin, A. et Grossmann, F. (2002). Collocations régulières et irrégulières : esquisse de typologie du phénomène collocatif. *Revue Française de Linguistique Appliquée, Lexique : recherches actuelles*, vol VII:p 7–25.

Vechtomova, O. (2005). The role of multi-word units in interactive information retrieval. *In ECIR2005*, pages 403–420, Berlin.

Vintar, S. et Fisier, D. (2008). Harvesting multi-word expressions from parallel corpora. *In Proceedings of LREC*, Marrakech, Morocco.

Vogel, S., Ney, H. et Tillmann, C. (1996). Hmm-based word alignment in statistical translation. *In Proceedings of the 16th conference on Computational linguistics - Volume 2*, COLING '96, pages 836–841, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vulić, I. et Moens, M.-F. (2012). Detecting highly confident word translations from comparable corpora without any prior knowledge. *In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 449–459, Avignon, France. Association for Computational Linguistics.

Wagner, C. (2005). Breaking the knowledge acquisition bottleneck through conversational knowledge management. *Information Resources Management Journal*, 19(1):70–83.

White, J. S. (1995). Approaches to black box mt evaluation.

Wu, D. (1997). Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403.

Wu, H., Wang, H. et Zong, C. (2008). Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. *In Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 993–1000, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wu, Z. et Palmer, M. (1994). Verbs semantics and lexical selection. *In Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, ACL '94, pages 133–138. Association for Computational Linguistics.

Xu, J. et Chen, J. (2011). How much can we gain from supervised word alignment ? *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies : short papers - Volume 2*, HLT '11, pages 165–169, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yu, K. et Tsujii, J. (2009). Extracting bilingual dictionary from comparable corpora with dependency heterogeneity. *In Proceedings of Human Language Technologies : The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume : Short Papers*, pages 121–124, Boulder, Colorado. Association for Computational Linguistics.

Zribi, A. (1995). *Contribution à l'étude de l'appariement de textes bilingues et monolingues*. These, Université Parix XI.

Zweigenbaum, P. et Habert, B. (2006). Les corpus naissent tous comparables en droit : apports méthodologiques de l'acquisition lexicale en contexte multilingue. *Glottopol*, pages 22–44. Version disponible sur Internet : [http ://www.univ-rouen.fr/dyalang/glottopol/numero_8.html].