# Alignement lexical en corpus comparables : le cas des composés savants et des adjectifs relationnels

Rima Harastani

UNIVERSITÉ DE NANTES

FACULTÉ DES SCIENCES ET DES TECHNIQUES

———

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

Année 2014

# Alignement lexical en corpus comparables : le cas des composés savants et des adjectifs relationnels

———

THÈSE DE DOCTORAT

Discipline : Informatique

Spécialité : Traitement Automatique du Langage Naturel

*Présentée*
*et soutenue publiquement par*

**Rima HARASTANI**

*Le 10 février 2014, devant le jury ci-dessous*

| | |
|---|---|
| Président | Holger SCHWENK, Professeur des universités, Université du Maine |
| Rapporteurs | Ulrich HEID, Professeur des universités, Université de Stuttgart |
| | Hervé BLANCHON, Maître de conférences (HDR), Université Pierre Mendès France |
| Examinateurs | Vincent CLAVEAU, Chargé de recherche, CNRS-IRISA Rennes |
| | Béatrice DAILLE, Professeur des universités, Université de Nantes |
| | Emmanuel MORIN, Professeur des universités, Université de Nantes |

*Directeur de thèse : Emmanuel MORIN, Professeur des universités, Université de Nantes*

*Co-encadrante de thèse : Béatrice DAILLE, Professeur des universités, Université de Nantes*

# Remerciements

Je tiens dans un premier temps à remercier tout particulièrement mon directeur de thèse Emmanuel Morin et ma co-encadrante de thèse Béatrice Daille pour leur aide, leur soutien, leur gentillesse ainsi que pour leurs remarques constructives au cours de ces années passées. Ils m'ont donné l'opportunité d'effectuer cette thèse dans un domaine d'autant plus intéressant qu'il est au coeur de l'actualité. J'ai beaucoup appris à leurs côtés d'un point de vue scientifique et humain.

Je remercie également les membres du jury d'avoir accepté de participer à ma soutenance de thèse et de s'être déplacés : les rapporteurs de thèse, Monsieur Hervé Blanchon, HDR à l'Université Pierre Mendès France, et Monsieur Ulrich Heid, Professeur à l'Université de Stuttgart, ainsi que les examinateurs de thèse, Monsieur Vincent Claveau, Chargé de Recherche au CNRS-IRISA, et Monsieur Holger Schwenk, Professeur à l'Université du Maine. Je les remercie également pour leurs remarques et leurs corréctions précieuses.

Ce travail a été financé et a fait partie du projet européen TTC. Ainsi, j'aimerais remercier tous les membres de ce projet pour leurs collaborations enrichissantes. J'adresse une pensée particulière à Emmanuel Planas pour son suivi de l'avancement de cette thèse. Je remercie aussi Jérôme Rocheteau pour sa patience et son aide.

J'ai eu l'opportunité de travailler dans une ambiance particulièrement agréable et riche au niveau intellectuel et humain au LINA. Je pense particulièrement à Elizaveta, qui est un exemple de gentillesse et est une des personnes les plus agréables avec laquelle on peut travailler. Un grand merci à Audrey, envers qui je suis extrêmement reconnaissante, je la remercie de m'avoir aidé à la relecture avec sourire et gentillesse. Je remercie aussi Nagham, Aurélien et Mathieu V. pour leur compréhension, leur écoute et leur compassion.

J'ai eu aussi la chance d'avoir plusieurs collègues de bureau avec qui j'ai partagé de très bons moments, un grand merci à Mohamed H., à Prajol et à Amir pour m'avoir soutenu, écouté et voyagé avec moi au cours de mes conférences. Je remercie les jeunes docteurs du laboratoire, Thomas C., Thomas V., Olivier, Mohamed M. et Marie [parmi d'autres], d'avoir été une source d'encouragement. Je pense également à ceux qui font toujours leur thèse : Ophélie, Andrea, Matthieu P. et Gabreilla [parmi d'autres]. Merci à tous pour votre bonne humeur.

Un remerciement spécial à Sahab et Malek pour leur soutien et d'avoir été à mes côtés dans les moments difficiles. Merci aussi à mes amis qui sont toujours en Syrie ou qui ont dû la quitter, pour leur courage inspirant et leur encouragement.

Mes plus grands remerciements iront à mes parents qui m'ont toujours soutenu sans fin et qui m'ont transmis leur curiosité et leur passion pour la science depuis tout petite. Merci aussi à mes sœurs et frère si courageux

et si déterminés, Ghada, Rania et Mohammad. Merci à ma famille d'être la source de ma motivation, de mon espoir et de ma réussite.

# Table des matières

# Liste des figures

# Liste des tables

# Notations

| Parties du discours | |
|---|---|
| A (ou Adj) | adjectif |
| AdjR | adjectif relationnel |
| AdjQ | adjectif qualificatif |
| DET | déterminant |
| N (ou Nom) | substantif |
| PREP | préposition |
| **Langues** | |
| DE | allemand |
| EN | anglais |
| ES | espagnol |
| FR | français |
| EN Nom | Nom est un mot en anglais |
| EN-FR | la paire de langues anglais-français |
| **Notations Linguistiques** | |
| [X + Y] | mot ou syntagme contenant X et Y |
| X : : Y | Y est une paraphrase de X |
| X\|Y | X ou Y |
| $X_+$ | X peut apparaître une ou plusieurs fois |
| X? | X peut apparaître zéro ou une fois |
| X- | X est un préfixe |
| -X | X est un suffixe |
| **Abréviations** | |
| ICF | Initial Combining Form |
| FCF | Final Combining Form |
| LCS | similarité de Longest Common Subsequence |
| Lev | distance de Levenshtein |
| TALN | Traitement Automatique du Langage Naturel |
| TAO | Traduction Assistée par Ordinateur |
| TAS | Traduction Automatique Statistique |
| TTC | Terminology Extraction, Translation Tools and Comparable Corpora |

# Introduction

## Contexte et motivations

Le multilinguisme est devenu une réalité omniprésente au niveau social, politique et économique. Cet atout est indispensable pour les échanges oraux comme écrits qui contribuent à la prospérité économique des organismes et des entreprises. Une communication entre deux entités peut être effectuée soit via la langue native de l'une des deux entités concernées, soit via la connaissance d'une langue commune tierce. Dans les deux cas, la barrière linguistique doit être levée de la manière la plus efficace possible en facilitant la communication et en la rendant moins coûteuse. Plusieurs technologies issues du domaine du Traitement Automatique du Langage Naturel (TALN) offrent des solutions à ces fins, adaptées à plusieurs langues, parmi lesquelles on peut citer : la traduction automatique (effectuée entièrement par des machines), la traduction assistée par ordinateur (qui peut se faire en partie manuellement ou de façon interactive avec la machine) ou la recherche d'information interlingue. Ces applications inter- ou multi-lingues doivent être flexibles afin de s'adapter aux particularités de chaque langue. D'un point de vue typologique, les langues peuvent parfois être très éloignées les unes des autres, et chaque langue possède des domaines ayant une forte technicité qui peuvent se révéler difficiles à manipuler. Les applications issues du TALN ne peuvent pas traiter ces défis sans s'appuyer sur des ressources linguistiques telles que les lexiques bilingues (c'est-à-dire des couples de traductions dans deux langues), qui constituent la pierre angulaire de ces applications.

Les lexiques bilingues ne sont pas toujours disponibles ni complets pour certaines paires de langues et pour certains domaines. Ils doivent donc être constamment mis à jour car les langues ne cessent d'évoluer. La construction des lexiques bilingues est particulièrement difficile dans les domaines scientifiques et techniques (dits domaines spécialisés), notamment à cause de l'utilisation de nombreux termes propres à ces domaines (c'est-à-dire des mots ou des expressions linguistiques qui désignent un concept spécifique) et qui ne se trouvent pas toujours dans les dictionnaires bilingues disponibles. Certains termes engendrent plus de problèmes dans un cadre multilingue que d'autres termes, comme les nouveaux termes créés dans une langue qui ne possèdent pas d'équivalents précis dans une autre langue. D'autres termes sont difficiles à traduire à cause des variations linguistiques dans une langue ou à cause de différences de caractéristiques de ces termes entre différentes langues.

L'évolution des langues et les difficultés liées à la traduction des termes rendent la construction ou la mise à jour manuelle des lexiques bilingues

coûteuse et longue, surtout dans les domaines spécialisés. Dans le but d'automatiser la construction de ces lexiques bilingues, des corpus contenant des textes, dans une langue dite source, avec leurs traductions, dans une langue dite cible, ont été largement exploités. La construction de tels ensembles de textes (nommés corpus parallèles) est coûteuse puisque cela nécessite que les textes traduits soient produits ou corrigés par un traducteur humain. Pour cela, les corpus parallèles restent des ressources précieuses mais rares et limitées en taille pour certains domaines et certaines langues (en dehors de quelques grandes entreprises et institutions). D'autres corpus, appelés corpus comparables, sont plus faciles à construire que les corpus parallèles et peuvent être également exploités pour la construction automatique des lexiques bilingues. Un corpus comparable comprend des textes écrits dans deux langues différentes sans aucune relation de traduction entre eux mais dont les textes partagent certaines propriétés (ex. ils appartiennent au même domaine). Lorsqu'un corpus comparable est disponible pour un domaine de spécialité, il pourra être considéré comme un réservoir de termes de ce domaine dans deux langues (langue source - langue cible), où une partie importante des termes d'une langue source peut être alignée avec son équivalent d'une langue cible.

Un corpus comparable dans un domaine de spécialité peut donc être une ressource intéressante pour extraire des alignements de termes dans deux langues différentes. Chacun des textes du corpus comparable est censé employer des termes dans le domaine du corpus et dans l'une des deux langues du corpus. Par des méthodes d'alignement, il va être possible de relier des termes dans la langue source à leurs équivalents dans la langue cible : l'ensemble de ces termes et leurs équivalents est appelé lexique bilingue spécialisé. Cette thèse s'intéresse ainsi aux tâches d'extraction automatique de lexiques bilingues spécialisés à partir de corpus comparables appartenant à un domaine de spécialité, et plus particulièrement à l'intégration des termes possédant des caractéristiques particuliers au sein de ces lexiques. Elle s'inscrit dans le cadre du projet européen TTC « Terminology Extraction, Translation Tools and Comparable Corpora », qui concerne le développement des outils pour plusieurs langues européennes (anglais, français, allemand, espagnol et letton) et deux langues non-européennes (russe et chinois) dans le cadre de la traduction terminologique.

## Problématiques

Dans le but d'extraire des lexiques bilingues à partir des corpus comparables, plusieurs approches sont proposées dans l'état de l'art. Certaines de ces approches font l'hypothèse que les mots ont tendance à apparaître dans des contextes similaires à ceux de leurs traductions. Ces approches, dites **distributionnelles**, dépendent essentiellement des cooccurrences de couples de mots pour aligner des mots avec leurs traductions. La qualité des paires de traductions extraites repose sur plusieurs facteurs : les fréquences des mots dans les corpus, les tailles des corpus, etc. Les approches

distributionnelles donnent généralement de bons résultats avec des corpus de grande taille. D'autres approches ont été proposées pour la traduction de composés syntagmatiques ou de termes complexes. Elles reposent sur une **propriété compositionnelle** : le sens de l'ensemble est une fonction du sens des parties. Ainsi, elles traduisent, par exemple, un terme complexe à partir des traductions de ses composants. Ces approches, dites **compositionnelles**, peuvent extraire des paires de traductions de haute qualité à partir des corpus comparables.

En ce qui concerne l'extraction des lexiques bilingues spécialisés à partir des corpus comparables, malgré les avancées, il y a encore beaucoup de progrès à faire pour améliorer la qualité des lexiques extraits. Cela vient principalement du fait que pour certains domaines de spécialité les tailles de corpus comparables sont limitées et que les termes sont difficiles à traduire. Nous cherchons à améliorer la qualité des lexiques bilingues spécialisés et plus particulièrement en ce qui concerne certains types de termes qui possèdent des caractéristiques les rendant difficiles à traiter : les composés savants et les adjectifs relationnels.

## Contributions

Trois contributions majeures sont proposées dans cette thèse. Les deux premières contributions se basent sur les méthodes compositionnelles et concernent deux types spécifiques de termes. La **première contribution** étudie l'extraction et la traduction des composés savants à partir de corpus comparables. Nous considérons qu'un composé savant est un terme qui contient au moins une racine gréco-latine, ex. radioactivité. La traduction des composés savants est difficile car beaucoup d'entre eux sont des néologismes (nouveaux mots) et ils sont productifs dans les domaines scientifiques (ex. la médecine). Nous menons des expériences avec deux paires de langues (français-anglais) et (français-allemand).

La **deuxième contribution** de cette thèse concerne la traduction des termes complexes qui ont la forme suivante : [Nom + AdjR] (*AdjR* désigne un adjectif relationnel), ex. cancer pulmonaire. La difficulté de la traduction de tels termes provient des variations linguistiques entre les langues. Nous réalisons des expériences avec un corpus français-anglais dans le domaine du cancer du sein et nous obtenons des traductions avec une bonne précision.

Notre **troisième contribution** consiste à explorer la possibilité de réordonner les traductions candidates, déjà fournies à un terme par une approche *distributionnelle*, en exploitant des phrases bilingues dans un corpus comparable. En effet, la traduction des termes ayant une propriété compositionnelle peut donner de hautes précisions. Cependant, quand il s'agit de trouver des traductions des termes simples ne possédant pas une propriété compositionnelle, une approche distributionnelle peut être appliquée mais les résultats ne sont pas toujours satisfaisants. Nous extrayons des phrases pour un terme source et des phrases pour chacune de ces traductions candidates. Ces phrases extraites sont ensuite alignées et aident à ré-ordonner

les paires de traductions. Nous réalisons des expériences avec des corpus dans deux domaines et trois paires de langues (français-anglais, français-espagnol et français-allemand). Nous appliquons cette méthode sur des lexiques bilingues produits par une approche distributionnelle et nous obtenons des améliorations dans les précisions des premières traductions candidates proposées.

## Organisation du travail

Le présent manuscrit est organisé de la manière suivante. Outre la présente introduction, le **chapitre 1** introduit les notions de lexiques bilingues, de termes et de corpus. Nous présentons aussi dans ce chapitre les ressources linguistiques de base utilisées pour mener nos expériences. Le **chapitre 2** présente les méthodes de l'état de l'art qui portent sur l'exploitation de corpus comparables.

Dans les **chapitres 3 et 4**, nous présentons nos travaux sur l'extraction et la traduction de deux types de termes : composés savants et termes complexes contenant des adjectifs relationnels. Nous introduisons également quelques travaux de l'état de l'art et clarifions le positionnement de notre travail par rapport à ces travaux. Le **chapitre 5** cherche à améliorer la qualité d'un lexique bilingue déjà extrait d'un corpus comparable. Il présente une méthode pour ré-ordonner une liste de traductions candidates proposées pour un terme en s'appuyant sur des phrases exemples bilingues dans un corpus comparable. Enfin, une **conclusion** vient conclure ce travail et propose différentes perspectives.

# Concepts de base et ressources linguistiques

1

Sommaire



Dans ce chapitre, nous décrivons les concepts de base de cette thèse ainsi que les principales ressources linguistiques exploitées. Nous abordons des dictionnaires/lexiques bilingues spécialisés et les types d'éléments qu'ils incluent : les termes simples et les termes complexes. Nous introduisons aussi le concept de corpus et les deux types de corpus qui ont été exploités dans la littérature pour la compilation automatique des lexiques bilingues : les corpus parallèles et les corpus comparables. Ensuite, nous présentons les différentes ressources linguistiques exploitées dans nos différentes expériences.

## 1.1 Dictionnaires

Dans le cadre de la construction de dictionnaires, nous distinguons deux notions : la lexicographie et la terminographie. La **lexicographie** est la constitution et l'étude des dictionnaires (Lehmann et Martin-Berthet 2008, p. 16). La **terminographie** est un ensemble d'activités dont l'objectif principal est la normalisation des termes propres à un domaine spécialisé (Cormier et Humbley 1998, p. 79). Ainsi, les dictionnaires généraux et les dictionnaires de spécialité, qui résultent de ces deux processus de construction, se distinguent par la nature de leurs éléments : les éléments de la langue générale versus les termes d'un domaine de spécialité.

Cette thèse, traite principalement de l'extraction automatique de **lexiques bilingues**, et plus précisément de l'extraction automatique de **lexiques bilingues spécialisés**. Nous introduisons d'abord le concept de dictionnaire bilingue, qui est une notion plus large et qui inclut le lexique bilingue. Béjoint et Thoiron (1996) lui donne la définition suivante : « *un dictionnaire bilingue est un dictionnaire dans lequel des expressions dans une langue (dite langue source ou de départ) sont traduits dans une autre (dite langue cible ou langue d'arrivée)* ». Un dictionnaire bilingue est spécialisé quand il contient des termes spécifiques à un domaine précis qui ne sont pas forcément recensés dans des dictionnaires généraux (Béjoint et Thoiron 1996, p. 39). Les dictionnaires fournissent en général des informations telles que les parties de discours des mots, des exemples d'emploi, etc.

Un lexique bilingue est, selon nous, une simple liste d'éléments dans une langue source avec leurs équivalents dans une langue cible. Il se distingue des dictionnaires par le fait qu'il ne contient pas d'informations supplémentaires sur ses éléments comme les catégories grammaticales et les définitions. Un lexique bilingue est dit spécialisé quand il recense des termes dans une langue source alignés avec leurs équivalents dans une langue cible.

Les dictionnaires et les lexiques constituent une ressource importante pour de nombreuses applications issues du TALN (traduction automatique, traduction assistée par ordinateur, recherche d'information, etc.). Cependant, l'incomplétude est un des problèmes majeurs de ces ressources (Bowker et Pearson 2002, p. 15). Les domaines scientifiques et techniques se développent très rapidement alors qu'un dictionnaire ou un lexique nécessitent beaucoup de temps afin d'être compilés et publiés. Par conséquent, les dictionnaires et les lexiques construits manuellement ne reflètent qu'une partie de l'état de la connaissance et doivent être constamment mis à jour.

En revanche, les textes produits dans une langue permettent de suivre automatiquement son évolution et contiennent des exemples concrets d'emplois. En partant de ce principe, des travaux ont exploité des textes afin de compiler ou de compléter des lexiques de langue générale ou spécialisée. Ces textes sont souvent sélectionnés selon un certain nombre de critères et constituent un corpus (cf. la section 1.3).

## 1.2 Termes

De nombreux travaux dans la littérature ont essayé de définir la notion de terme. Selon la définition donnée par l'ISO 1087 Vocabulary of Terminology, un terme est :

« *Une désignation d'un concept défini dans une langue spécialisée par une expression linguistique.* » [1] (1)

Pearson (1998) présente deux types de définition des termes dans la littérature : la définition traditionnelle et la définition pragmatique. La notion de terme selon les définitions traditionnelles est applicable aux unités lexicales avec une référence spéciale dans un domaine spécialisé (Pearson 1998, Sager 1990). Les définitions pragmatiques de termes acceptent les définitions traditionnelles mais elles classifient les termes dans des catégories. Par exemple, elles distinguent les termes spécifiques au domaine de spécialité et les termes non-spécifiques au domaine de spécialité. Trimble (1978) constate que chaque domaine a ses propres termes qui sont très spécifiques au domaine, et qu'il y a des termes qui peuvent être communs entre les différents domaines (Pearson 1998, p. 17).

Nous introduisons dans la suite de cette section les différentes formes de termes et leurs variantes.

### 1.2.1 Termes simples et complexes

Reprenons la définition (1) d'un terme, qu'est-ce qui peut être une « expression linguistique » ? Une réponse valable pour la plupart des langues est que cette expression peut se composer d'un seul mot (simple) ou de plusieurs mots (complexe). Selon (L'Homme 2004, p. 59), les termes en français peuvent être simples ou complexes (du point de vue de la forme graphique) :

- **termes simples** : ils se composent d'une seule entité graphique. Les termes simples peuvent être formés d'une seule base (ex. système), dérivés (ex. anti-âge) ainsi que construits sur des racines gréco-latines (ex. radiologie).
- **termes complexes** : ils se composent de deux ou plusieurs entités graphiques, ces entités forment une expression à sens unique. Les entités graphiques (ou les composants) d'un terme complexe peuvent être séparées par des espaces blancs (ex. système solaire). Elles peuvent être séparées aussi par des diacritiques comme le trait d'union ou l'apostrophe (ex. système-expert).

Les termes simples dérivés sont formés par l'ajout des affixes (préfixes, suffixes) à un radical (ex. postposition / [post + position], alcoolique / [alcool + ique]). Les termes simples sont appelés composés savants quand ils sont construits soit uniquement à partir de plusieurs racines gréco-latines (ex. hydrophile / [hydro + phile]), soit à partir d'une ou plusieurs

1. "Designation of a defined concept in a special language by a linguistic expression."

racines gréco-latines combinées avec un mot (ex. cardiovasculaire / [cardio + vasculaire]).

Les termes complexes possèdent souvent une **propriété compositionnelle** : L'Homme (2004) définit cette propriété par le fait qu'on peut comprendre le sens d'un terme complexe à partir des sens des unités simples qui le composent. Par exemple, le terme *calendrier lunaire* est par définition « *un calendrier réglé sur les phases de la lune* » [2]. Cette propriété a été exploitée par plusieurs travaux dans le but de traduire automatiquement des termes complexes, nous détaillons ces travaux dans le chapitre 2.

### 1.2.2 Variantes des termes

Les différentes usages possibles d'un terme dans une langue font que celui-ci peut être employé sous différentes formes (dites variantes). Depierre (2007) considère que les variantes d'un terme doivent avoir une signification similaire à ce terme. D'autres travaux adoptent une définition plus large des variantes. Ils considèrent que les variantes d'un terme ne sont pas toujours des synonymes de ce terme, mais elles peuvent être en relations d'hyperonymie, d'antonymie, etc. Nous reprenons quelques types de variation citées dans Daille (2003) :

- **variation graphique** : un terme peut prendre plusieurs formes graphiques. Par exemple, en ajoutant un tiret (ex. FR kilowattheure / kilowatt-heure) entre les composants d'un terme, ou en utilisant une écriture valide mais différente de celle du terme d'origine (ex. EN lung color / lung colour, EN antiestrogen / antioestrogen).
- **variation syntaxique** : la structure syntaxique d'un terme change si par exemple un élément est inséré entre les composants du terme (ex. FR mutation des cellules / mutation génétique des cellules), ou si le terme a été construit selon une autre forme linguistique (ex. EN blood group [Nom + Nom] / group of blood [Nom + PREP [3] + Nom]), etc.
- **variation morpho-syntaxique** : les structures syntaxique et morphologique d'un terme peuvent être modifiées (ex. FR groupe du sang [Nom + PREP + Nom] / group sanguin [Nom + Adjectif]).
- **variation sémantique** : un des composants d'un terme complexe peut être remplacé par un synonyme (ex. FR matériel électrique / équipement électrique (Hamon et Nazarenko 2001)).

## 1.3 Corpus

Il existe de nombreuses définitions d'un corpus dans la littérature, prenons une définition souvent citée, celle de (Sinclaire 2003, p. 4) :

« *Un corpus est une collection de morceaux de langues qui sont sélectionnés*

2. http ://fr.wikipedia.org/wiki/Calendrier_lunaire
3. PREP signifie préposition.

*et ordonnés selon des critères linguistiques et extra-linguistiques explicites afin d'être utilisés comme un échantillon de la langue.* »

L'expression « morceaux de langues » est utilisée parce qu'un corpus peut contenir des textes ou des transcriptions des discours complets ou incomplets.

Selon un autre point de vue, plusieurs chercheurs ont considéré le corpus sous l'angle de la méthodologie de constitution (Meyer 2002, McEnery et Gabrielatos 2006, Bowker et Pearson 2002). Par exemple, le corpus est décrit dans (Bowker et Pearson 2002, p. 9) comme « une approche ou une méthodologie pour l'étude de l'usage d'une langue ».

Dans le domaine du TALN, le terme *corpus* désigne des textes numériques construits pour être analysés d'un point de vue linguistique ou statistique. Il est défini par (Bowker et Pearson 2002, p. 9) comme étant « *une large collection de textes authentiques réunis dans une forme électronique selon un ensemble de critères spécifiques* ».

McEnery et Gabrielatos (2006) constatent qu'il y a un consensus croissant pour définir un corpus comme étant (a) des textes numériques (b) et authentiques qui composent (c) un échantillon (d) représentatif d'une langue ou d'une variété d'une langue. La plupart des travaux sur les corpus acceptent les deux premiers points (a) et (b), mais ils ne sont pas d'accord sur les caractéristiques qui font d'un corpus un échantillon représentatif d'une population langagière[4] (Taylor 2008). Par exemple, un corpus n'est représentatif d'une population langagière que si sa taille reflète avec précision cette population (Habert 2000).

Après avoir défini la notion de corpus, nous détaillons maintenant les différents types de corpus, en particulier les corpus parallèles et les corpus comparables.

### 1.3.1 Types de corpus

Les corpus peuvent être classés dans des catégories différentes selon les critères choisis pour les compiler. Nous distinguons les corpus relevant de la langue générale des corpus de la langue de spécialité, les corpus monolingues et les corpus multilingues, ainsi que les corpus parallèles et les corpus comparables.

Un **corpus de langue générale** doit permettre de faire des observations générales sur cette langue. Par exemple, un corpus de langue générale peut contenir des articles de journaux, des émissions de télévision, des débats, etc. C'est-à-dire qu'un corpus de langue générale fait des références à la langue utilisée au quotidien, comprise et admise par tous. Un exemple d'un corpus de langue générale est le Corpus National américain (ANC) contenant de textes de l'anglais américain avec 22 millions de mots de données écrites et orales produites depuis 1990. D'autre part, un **corpus de langue de spécialité** traite un domaine spécifique et doit être représentatif de la langue de spécialité en usage dans le domaine. Dubreil (2006) le

4. Une langue dans son ensemble, une langue familière, une langue de spécialité, etc.

définit comme suit : « *tout regroupement de données langagières créé à des fins spécifiques et représentatif d'une situation de communication ou d'un domaine dans la pratique* ». Un corpus spécialisé est en effet considéré comme « *un vaste réservoir de termes* » (L'Homme 2004, p. 118).

Un corpus est **monolingue** s'il contient des textes dans une seule langue. Alors qu'un corpus est **bilingue** ou **multilingue** s'il s'agit de textes issus de deux ou plusieurs langues (Bowker et Pearson 2002, p. 12).

### 1.3.2 Corpus parallèles

Un corpus parallèle comprend un ensemble de textes dans une langue source accompagnés de leurs traductions dans une ou plusieurs langues cibles (Olohan 2004, Bowker et Pearson 2002, p. 24,92). L'exemple le plus connu de corpus parallèle multilingue est le corpus *Europarl*[5], qui rassemble des textes du Parlement Européen dans 11 langues et contient plus de 20 millions de mots par langue (Koehn 2005). Un autre exemple est le corpus *Hansard*[6] qui est composé de textes anglais et français extraits de débats du Parlement Canadien (il contient une dizaine de millions de mots).

Les corpus parallèles sont utilisés par de nombreuses applications du TALN (ex. traduction automatique, logiciels de concordance, etc.) (Olohan 2004, p. 25) (Munteanu et Marcu 2005). L'alignement des corpus parallèles au niveau des phrases ou des mots a été beaucoup étudié dans la littérature (ex. Brown et al. (1993)), il permet d'extraire des informations bilingues comme des constructions grammaticales ou des instances de lexiques bilingues (Olohan 2004, p. 25). Cependant, les corpus parallèles restent des ressources limitées en taille, disponibles pour certains domaines et certaines langues (Munteanu et Marcu 2005). La construction des corpus parallèles nécessite pour un humain de traduire des documents d'une langue source vers une langue cible, ce qui peut être coûteux en termes de temps et d'argent. Les exemples de grands corpus parallèles cités ci-dessus (Europarl et Hansard) restent des exemples de corpus limités à certains domaines et construits au sein des organismes multilingues. Par ailleurs, la traduction des textes originaux ne peut être qu'influencée par ces derniers. Les textes résultant de la traduction risquent d'avoir des exemples de calques ou d'autres biais de traduction (Delpech et al. 2012). De plus, lors de la traduction des textes scientifiques, un traducteur qui ne connaît pas la traduction d'un terme peut suivre plusieurs stratégies (Abdellah 2003, Baker 1998) qui ne sont pas toujours idéales. Par exemple, il peut paraphraser le terme avec d'autres mots, utiliser des mots moins représentatifs pour expliquer le terme, omettre le terme du texte traduit, etc.

5. http ://www.statmt.org/europarl/
6. http ://www.ldc.upenn.edu/Catalog/catalogEntry.jsp ?catalogId=LDC95T20

### 1.3.3 Corpus comparables

Un corpus comparable contient des textes dans des langues différentes qui ne sont pas des traductions mutuelles. Le terme *comparable* est utilisé pour indiquer que ces textes partagent certaines caractéristiques ou certains traits (Bowker et Pearson 2002, p. 93). Par exemple, on sélectionne souvent des textes qui ont en commun : le sujet, la période ou le degré de technicité, etc. D'autres questions peuvent se poser lors de la construction des corpus comparables. Par exemple, les parties d'un corpus comparables, doivent-ils être comparables en taille (en nombre de textes, de mots ou de phrases) (Maia 2003) ?

Les articles des journaux de différentes sources dans plusieurs langues, qui sont publiés dans une période définie, peuvent constituer un corpus comparable de langue générale (Fung et Cheung 2004). En effet, les corpus comparables de grande taille sont souvent construits à partir des textes issus de journaux et peuvent couvrir plusieurs sujets. Le terme *corpus concurrents* a été utilisé afin de décrire les corpus collectés à partir des textes issus de journaux et qui traitent le même sujet, ou qui sont publiés dans la même période (Maia 2003). Par exemple, des rapports de guerres publiés dans une période définie, de l'agence de presse *Reuters* pour l'anglais et de l'agence de presse *AFP* pour le français, peuvent constituer un corpus comparable concurrent anglais-français.

Les corpus comparables spécialisés peuvent être collectés à partir des textes techniques ou scientifiques avec des degrés variables de spécificité. Cependant, puisque ces textes sont souvent utilisés pour l'extraction terminologique, il est souhaitable qu'ils contiennent une densité importante de termes. Un exemple de corpus comparable spécialisé est le corpus médical utilisé par Déjean et al. (2002) qui est composé de documents médicaux (anglais et allemand) extraits de la base médicale Medline[7] (environ 200 000 mots).

Les corpus comparables peuvent être disponibles pour de nombreux domaines et langues. Ils permettent de pallier le manque de corpus parallèles (Zweigenbaum et Habert 2006). De plus, les corpus comparables répondent aux besoins de textes originaux « naturellement » produits dans une langue. Pour une langue source et une autre cible, il est normalement plus facile de trouver pour chaque langue de textes originaux (sur le Web par exemple) et dans un domaine particulier, que de trouver ou de construire des textes parallèles pour cette paire de langues.

Les corpus comparables ont été exploités pour établir automatiquement des correspondances entre :

– les mots (ex. Fung et Mckeown (1997)) ou les termes (ex. Déjean et al. (2002), Morin et al. (2008)) : l'alignement entre mots ou termes se base principalement sur les similarités de contextes d'utilisation (Zweigenbaum et Habert 2006, p. 23). L'alignement entre termes complexes peut exploiter une propriété compositionnelle de ces termes (ex. Baldwin et Tanaka (2004)).

7. http ://www.ncbi.nlm.nih.gov/PubMed/

– les phrases (ex. Munteanu et Marcu (2005)) ou les segments (ex. Munteanu et Marcu (2006)) : l'alignement entre phrases ou segments se base notamment sur leurs similarités lexicales (ex. le nombre de mots en commun).

Deux caractéristiques des corpus comparables ont été plus particulièrement discutées dans la littérature de l'alignement bilingue : la taille et le degré de comparabilité d'un corpus comparable.

**Taille de corpus** Des corpus comparables de grande taille (plusieurs millions de mots) ont été souvent compilés pour extraire des lexiques bilingues. De tels corpus peuvent être, en quelque sorte, facilement acquis quand il s'agit des corpus de langue générale. Cependant, des corpus de grande taille ne sont pas disponibles pour certains domaines spécialisés. Pour faire face à ce problème, Morin et al. (2008) font l'hypothèse que pour extraire des lexiques bilingues spécialisés (à partir de corpus spécialisés), la qualité de ces corpus serait plus importante que leurs tailles. En théorie, la taille généralement appropriée d'un corpus spécialisé est aux alentours de 500 000 mots (Williams 1999).

**Comparabilité de corpus** La notion de degré de comparabilité d'un corpus permet de quantifier dans quelle mesure les textes d'un corpus sont comparables (Goeuriot 2009). Un corpus a un degré de comparabilité élevé s'il comprend des textes ayant de nombreuses caractéristiques en commun (ex. dates de publication, domaines, thèmes, genres, etc.). Le degré de comparabilité d'un corpus varie selon l'ensemble des caractéristiques ou des critères de comparabilité choisis.

Cependant, la notion de degré de comparabilité reste vague parce qu'elle se base sur des concepts linguistiques ou extra-linguistiques difficilement mesurables. Nous présentons en section 2.5 du chapitre 2 des méthodes opérationnelles de l'état de l'art qui s'intéressent à mesurer le degré de comparabilité des corpus.

## 1.4 Ressources linguistiques

Cette thèse s'inscrit dans l'exploitation de corpus comparables afin d'extraire ou d'enrichir des lexiques bilingues. Nous présentons les corpus comparables ainsi que les dictionnaires bilingues (de base) que nous avons exploités.

### 1.4.1 Corpus comparables

Nous avons à notre disposition deux corpus comparables spécialisés : le premier a été extrait semi-manuellement et le deuxième a été collecté automatiquement.

### Cancer du sein

Nous avons un corpus comparable dans le domaine du cancer du sein en trois langues : le français, l'anglais et l'allemand. Ce corpus a été construit à partir d'articles scientifiques publiés sur le Web. Les articles ont été collectés de manière à ce qu'ils répondent à un certain nombre de critères : l'ensemble d'articles dans une langue doivent (a) contenir le terme clé *cancer du sein* pour le français (ou ses équivalents pour les autres langues : *breast cancer* pour l'anglais et *Brustkrebs* pour l'allemand), (b) être publiés dans la période 2001-2008 et (c) être comparables en taille avec les autres ensembles d'articles dans les autres langues.

La table 1.1 résume la taille de chacun des corpus en nombre de mots et d'articles.

| | **FR** | **EN** | **DE** |
|---|---|---|---|
| **Nb. de mots** | 529 544 | 527 268 | 378 474 |
| **Nb. d'articles** | 130 | 104 | 262 |

TABLE 1.1 – *Caractéristiques des corpus cancer du sein*

### Énergies renouvelables

Dans le cadre du projet européen TTC [8], un corpus comparable dans le domaine des énergies renouvelables a été construit dans plusieurs langues. Nous utilisons ces corpus afin de mener nos expériences sur les langues suivantes : le français, l'anglais, l'allemand et l'espagnol. Les corpus ont été collectés à partir des pages Web en utilisant *Babouk* (Groc 2011), qui est un crawler [9] développé dans le cadre du projet TTC pour la compilation automatique des corpus spécialisés à partir du Web. Cet outil prend en entrée une liste de termes spécifiques à un domaine, appelée liste d'amorces de termes, et trouve des textes sur le Web qui traitent du domaine spécialisé (Loginova et al. 2012). Lors de la première itération (de collection des pages Web), cette liste d'amorces de termes est étendue en s'appuyant sur les nouveaux termes se trouvant dans les pages Web collectées. Pour chercher des textes tirés de sites HTML ainsi que des fichiers PDF et Word, des amorces de termes sont combinés de manière aléatoire [10] pour former une requête. Cette dernière est ensuite soumise à un moteur de recherche qui va retourner les $n$ meilleures pages qui correspondent à cette recherche. Ensuite, des filtres définis choisissent les pages qui sont riches en terminologies spécifiques au domaine et ignorent les pages qui ne sont pas propres après conversion au format texte (ex. à cause de l'encodage) ou celles qui sont de très petite taille ou de très grande taille, etc.

8. www.ttc-project.eu

9. Un crawler est un programme qui, étant donné une ou plusieurs amorces des adresses URL (ou des amorces de mots), télécharge les pages Web associées à ces adresses URL, extrait les hyperliens qu'elles contiennent et continue de manière récursive à télécharger les pages Web identifiées par ces hyperliens. (Olston et Najork 2010)

10. Trois termes par défaut.

Il n'est pas toujours évident pour Babouk de collecter des corpus spécialisés pour certaines langues (ex. letton) peu représentées sur le Web dans le domaine de spécialité. Pour cela, certains corpus monolingues ont été étendus avec des documents recueillis manuellement à partir du Web afin que la taille de chaque corpus monolingue soit au minimum de 300 000 mots.

Les corpus du domaine des énergies renouvelables construits pour plusieurs langues peuvent être téléchargés depuis le site électronique du LINA [11].

La table 1.2 résume la taille des corpus français, anglais, allemand et espagnol en nombre de mots et d'articles.

| | **FR** | **EN** | **DE** | **ES** |
|---|---|---|---|---|
| **Nb. de mots** | 313 943 | 314 549 | 358 602 | 453 953 |
| **Nb. d'articles** | 11 | 28 | 34 | 46 |

TABLE 1.2 – *Caractéristiques des corpus énergies renouvelables*

### 1.4.2 Dictionnaires bilingues

Les dictionnaires bilingues sont indispensables aux approches que nous proposons et mettons en œuvre.

Nous disposons de dictionnaires bilingues de langue générale pour les paires de langues français-anglais, français-allemand et français-espagnol proposés par le catalogue ELRA [12]. Ils contiennent des informations sur les parties du discours des mots. La table 1.3 présente le nombre d'entrées des mots simples (les entrées sont des lemmes) pour chaque dictionnaire.

| | **FR-EN** | **FR-DE** | **FR-ES** |
|---|---|---|---|
| **Nb. d'entrées** | 145 542 | 118 776 | 79 317 |

TABLE 1.3 – *Caractéristiques des dictionnaires bilingues ELRA*

Afin d'évaluer la couverture des dictionnaires par rapport au vocabulaire des corpus, nous calculons l'intersection entre le vocabulaire d'un corpus d'une langue (nous prenons les lemmes des mots obtenus par le logiciel *TreeTagger* [13]) et le vocabulaire d'un dictionnaire de cette langue.

La table 1.4 résume le nombre de mots en commun entre le dictionnaire d'une paire de langues et les corpus correspondant à cette paire. Par exemple, en utilisant le dictionnaire FR-EN, nous trouvons qu'il existe entre le corpus cancer du sein français (énergies renouvelables respectivement) et la partie française du dictionnaire 6 550 mots (3 901 mots respectivement) en commun. L'intersection entre la partie anglaise et le corpus cancer du sein anglais (énergies renouvelables respectivement) est de 7 299 mots (4 347 mots respectivement).

11. http ://www.lina.univ-nantes.fr/ ?Ressources-linguistiques-du-projet.html
12. http ://catalog.elra.info/product_info.php ?products_id=666,
http ://catalog.elra.info/product_info.php ?products_id=667
et http ://catalog.elra.info/product_info.php ?products_id=668
13. http ://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

| Dictionnaire/corpus | Langue source | Langue cible |
|---|---|---|
| FR-EN/cancer du sein | 6 550 | 7 299 |
| FR-EN/énergies renouvelables | 3 901 | 4 347 |
| FR-DE/cancer du sein | 5 743 | 3 750 |
| FR-DE/énergies renouvelables | 2 637 | 3 107 |
| FR-ES/énergies renouvelables | 3 610 | 4 193 |

TABLE 1.4 – *Nombre de mots en commun entre les dictionnaires bilingues et les corpus*

Nous pouvons remarquer que le nombre de mots en commun entre les dictionnaires et les corpus pour la langue allemande est inférieur au nombre de mots en commun entre les dictionnaires et les corpus d'autres langues. En effet, le vocabulaire commun entre les dictionnaires et les corpus (c'est-à-dire la couverture de corpus par les dictionnaires) peut être un indice sur la qualité ou sur le degré de comparabilité de corpus, comme nous verrons dans la section 2.5 du chapitre 2.

## 1.5 Conclusion

Nous avons introduit, dans ce chapitre, les éléments linguistiques et les ressources que nous allons manipuler par la suite.

Nous nous intéressons dans cette thèse au traitement des termes qui se composent de plusieurs éléments : les composés savants et les termes complexes contenant des adjectifs relationnels. Nous notons que les corpus comparables dans les domaines spécialisés sont souvent de taille modeste, ce qui est le cas des corpus à notre disposition. Ceci pose une difficulté supplémentaire à l'extraction de lexiques spécialisés de bonne qualité. L'objectif général de la présente thèse est donc d'enrichir les lexiques bilingues ou d'améliorer leur qualité.

Nos contributions sont inspirées de certaines méthodes de l'état de l'art que nous décrivons dans le chapitre suivant. Ces méthodes sont consacrées à l'exploitation des corpus comparables pour l'extraction des correspondances bilingues.

# État de l'art

# 2

Sommaire



Des correspondances entre des mots, des termes, des segments ou des phrases peuvent être élaborées automatiquement en exploitant des corpus parallèles ou des corpus comparables. L'extraction automatique des correspondances multilingues permet de construire des lexiques bilingues et d'améliorer la performance des systèmes de traduction automatique ou assistée par ordinateur.

Dans ce chapitre, nous mettons l'accent sur des approches de l'état de l'art destinées à l'exploitation des corpus comparables pour l'extraction des lexiques bilingues. Ces derniers peuvent être obtenus à l'aide des approches, nommées distributionnelles, qui se basent sur les distributions de mots dans les textes. Les lexiques bilingues peuvent également être enrichis par des approches, nommées compositionnelles, qui exploitent une propriété compositionnelle de certains types de mots ou de termes.

De même, nous présentons des travaux de l'état de l'art qui se concentrent sur l'extraction des phrases parallèles d'un corpus compa-

rable ainsi que d'autres travaux qui s'intéressent à l'amélioration de la qualité d'un corpus comparable.

## 2.1 Approches distributionnelles

Nous décrivons dans cette section les travaux pionniers de l'état de l'art de Rapp (1995), Fung (1995), Rapp (1999) et Fung et Mckeown (1997), qui exploitent les corpus comparables pour l'acquisition ou l'enrichissement d'un lexique bilingue. Les méthodes proposées, dans ces travaux, supposent qu'un mot et sa traduction partagent des contextes similaires dans un corpus comparable. Elles trouvent des corrélations statistiques entre les mots dans des langues différentes à partir des corpus comparables. Le contexte d'un mot est défini dans un premier temps, puis un mot source est aligné avec un autre mot cible en exploitant la similarité de leurs contextes.

### 2.1.1 Fondements

Le travail présenté dans Rapp (1995) a été l'un des premiers à essayer d'extraire des couples de traductions à partir d'un corpus comparable. Rapp (1995) fait l'hypothèse qu'il existe une corrélation entre les cooccurrences des mots dans un corpus source et les cooccurrences de leurs traductions dans un corpus cible. Par exemple, si les mots *étudiant* et *école* cooccurrent plus souvent que dans le cas du hasard dans un corpus français, leurs traductions respectives en anglais *student* et *school* doivent également cooccurrer plus souvent que dans le cas du hasard dans un corpus anglais. En se basant sur cette hypothèse, la traduction ($m_c$) d'un mot ($m_s$) peut être trouvée s'il y a une corrélation entre les cooccurrences de $m_s$ et de $m_c$ respectivement dans le corpus source et le corpus cible. Les corrélations sont calculées en se basant sur les fréquences de cooccurrences des mots.

Fung (1995) a également essayé d'extraire des couples de traductions à partir des corpus comparables. Elle suppose que les mots qui ont un contexte productif dans une langue sont traduits par des mots qui ont un contexte tout autant productif dans une autre langue. À titre d'exemple, si le mot EN *air* cooccurre avec beaucoup de mots différents dans un corpus anglais, sa traduction en chinois *Ko-ngqì* doit aussi cooccurrer avec relativement autant de mots différents dans un corpus chinois. La différence principale entre cette hypothèse et celle de Rapp (1995), c'est que Fung (1995) ne se base pas sur les fréquences de cooccurrences de mots mais sur le nombre de mots différents qui cooccurrent avec un mot pour trouver sa traduction.

Ces deux travaux, ceux de Rapp (1995) et de Fung (1995), proposent des méthodes qui n'utilisent que des corpus comparables comme ressources afin de trouver les traductions des mots. Cependant, la méthode de Rapp (1995) est coûteuse parce qu'elle calcule, pour représenter le contexte d'un mot, les cooccurrences de ce mot avec tous les mots dans un corpus. La méthode de Fung (1995) nécessite un corpus de très grande taille et pose un réel problème pour mesurer la productivité du contexte d'un mot peu fréquent.

Pour pallier les problèmes de l'approche de Rapp (1995), Fung et Mckeown (1997) se basent sur la même hypothèse que Rapp (1995) propose, mais elles développent une méthode moins coûteuse en s'appuyant sur un dictionnaire bilingue de base (*seed words*). Ainsi, pour représenter le contexte d'un mot, on calcule ses cooccurrences avec les mots qui se trouvent à la fois dans le corpus et dans le dictionnaire de base.

De même, Rapp (1999) améliore sa méthode (Rapp 1995) en utilisant un petit dictionnaire bilingue de base.

Nous détaillons maintenant les représentations des contextes des mots et les méthodes proposées pour calculer des similarités entre les contextes des mots de deux langues.

### 2.1.2 Représentation des contextes et calcul de similarité

Afin d'extraire un lexique bilingue anglais-allemand d'un corpus comparable, Rapp (1995) propose de calculer une matrice de cooccurrences pour chaque langue, à partir des cooccurrences de chaque couple de mots dans le corpus. Les tables 2.1 et 2.2 illustrent deux matrices de cooccurrences, où les lignes et les colonnes d'une matrice sont représentées par des mots : 6 mots pour l'anglais et 6 mots pour l'allemand (exemples tirés de Rapp (1995)). La case qui correspond à un couple de mot, qui cooccurre plus souvent que dans le cas du hasard, est marquée par un point.

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| blue 1 | | • | | | • | |
| green 2 | • | | • | | | |
| plant 3 | | • | | | | |
| school 4 | | | | | | • |
| sky 5 | • | | | | | |
| teacher 6 | | | | • | | |

TABLE 2.1 – *Matrice de cooccurrences de 6 mots anglais (Rapp 1995)*

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| blau 1 | | • | • | | | |
| grün 2 | • | | | | • | |
| Himmel 3 | • | | | | | |
| Lehrer 4 | | | | | | • |
| Pflanze 5 | | • | | | | |
| Schule 6 | | | | • | | |

TABLE 2.2 – *Matrice de cooccurrences de 6 mots allemands (Rapp 1995)*

Rapp (1999) considère que deux mots cooccurrent s'ils se trouvent, dans le corpus, séparés par 11 autres mots au maximum (fenêtre ($w$) de taille 5 : 5 mots avant et 5 mots après un mot). Les valeurs de cooccurrences dans les matrices sont représentées par les nombres de cooccurrences des mots dans le corpus, ces valeurs sont ensuite normalisées pour réduire

| | 1 | 2 | 5 | 6 | 3 | 4 |
|---|---|---|---|---|---|---|
| blue 1 | | • | • | | | |
| green 2 | • | | | | • | |
| sky 5 | • | | | | | |
| teacher 6 | | | | | | • |
| plant 3 | | • | | | | |
| school 4 | | | | • | | |

TABLE 2.3 – *Matrice réordonnée de 6 mots anglais (Rapp 1995)*

l'effet de la fréquence des mots, en utilisant la formule suivante :

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

où $A$ est la matrice de cooccurrences, $f(i\&j)$ est la fréquence des occurrences de deux mots $i$ et $j$ ensemble dans des fenêtres $w$. $f(i)$ et $f(j)$ sont les fréquences de mots $i$ et $j$ respectivement.

La similarité ($s$) entre la matrice de langue source ($E$) et la matrice de langue cible ($G$) est mesurée par les différences absolues des valeurs qui correspondent aux mêmes positions dans les matrices :

$$s = \sum_{i=1}^{N} \sum_{j=1}^{N} |E_{i,j} - G_{i,j}| \tag{2.2}$$

Une des matrices est réordonnée jusqu'à ce que le score de la similarité $s$ entre les deux matrices (source et cible) converge vers une valeur minimale, ce qui indique une similarité maximale entre les matrices. Ensuite, le mot qui correspond à la position $i$ dans la matrice $E$ est considéré comme la traduction du mot correspondant à la position $i$ dans la matrice $G$. Par exemple, la matrice de la table 2.1 peut être réordonnée de manière à ce qu'elle correspond parfaitement à la matrice de la table 2.2, comme dans la table 2.3. Le mot EN *green* dans ligne 2 de la matrice dans la table 2.3 sera aligné avec le mot EN *grün* dans la ligne 2 de la matrice dans la table 2.2.

Une autre approche est présentée dans Fung (1995), elle introduit le trait d'hétérogénéité d'un mot qui est caractérisé par le nombre de mots différents qui le précèdent et qui le suivent immédiatement.

Un mot ($m$) est représenté par un vecteur de l'hétérogénéité de contexte, qui est la paire ($x$, $y$), où $x$ est l'hétérogénéité à gauche et $y$ est l'hétérogénéité à droite. Ces hétérogénéités sont calculées de la manière suivante :

- hétérogénéité à gauche $x = \frac{a}{c}$
- hétérogénéité à droite $y = \frac{b}{c}$

où :

- $a$ est le nombre de mots différents qui précèdent immédiatement $m$ dans le corpus.
- $b$ est le nombre de mots différents qui suivent immédiatement $m$ dans le corpus.

– $c$ est le nombre d'occurrences de $m$ dans le corpus.

L'hétérogénéité du contexte de n'importe quel article, comme EN *the*, peut avoir des valeurs $x$ et $y$ très importantes, dans la mesure où EN *the* peut être précédé et suivi par beaucoup de mots différents. En revanche, la valeur $x$ du mot EN *am* n'est pas importante, parce que ce mot suit très souvent le mot EN *I*. Fung (1995) suppose que l'hétérogénéité du contexte d'un mot dans le corpus source est plus similaire à celle de la traduction de ce mot dans le corpus cible, qu'à celle d'un mot indépendant dans le corpus cible. Par exemple, supposons que l'hétérogénéité du contexte du mot *air* est de (0,676, 0,267) et celle de sa traduction en chinois *Ko-ngqì* est de (0,784, 0,459). L'hétérogénéité du contexte d'un autre mot chinois *Xiu-huì* (litt. ajournement) est de (0,211, 0,091), cela indique que *Ko-ngqì* a un contexte beaucoup plus riche que *Xiu-huì*, car son hétérogénéité de contexte est plus élevée.

Afin de mesurer la similarité entre deux vecteurs d'hétérogénéité du contexte, Fung (1995) propose d'utiliser la distance euclidienne. Reprenons l'exemple précédent, la distance euclidienne entre l'hétérogénéité du contexte de *air* et celle de *Ko-ngqì* est de 0,2205, alors que la distance entre l'hétérogénéité du contexte de *air* et celle de *Xiu-huì* est de 0,497. Cela signifie que *air* a un contexte productif plus similaire à *Ko-ngqì* qu'à celui de *Xiu-huì* et que *Ko-ngqì* est une traduction plus probable de *air* que *Xiu-huì*.

La méthode pionnière de Rapp (1995) a fait l'objet de plusieurs améliorations. Fung et Mckeown (1997) ont proposé une première extension qui porte sur l'exploitation d'un lexique/dictionnaire bilingue de base afin de réduire le calcul coûteux de la méthode de Rapp (1995).

Fung et Mckeown (1997) représentent le contexte d'un mot, comme Rapp (1995), à partir de ses cooccurrences avec d'autres mots. Cependant, elles utilisent des amorces de mots à partir des dictionnaires bilingues afin d'éviter le calcul coûteux de la méthode de Rapp (1995). Pour chaque langue, au lieu de calculer les corrélations d'un mot avec tous les autres mots du corpus, elles calculent les corrélations d'un mot avec les mots du corpus apparaissant dans le dictionnaire utilisé. Ces corrélations sont utilisées pour représenter le contexte d'un mot par un vecteur de relation des mots (Word relation Matrix $WoRm$). L'algorithme suivant est proposé afin de trouver des traductions de mots à partir d'un corpus comparable :

1. un dictionnaire bilingue de base est disponible (c'est-à-dire les amorces de mots).
2. pour chaque mot ($w_x$) dans une langue 1, sa corrélation avec chaque mot dans le dictionnaire bilingue d'une langue 1 est calculée. Cela donne un vecteur de relation ($WoRm_1$).
3. pour chaque mot ($w_c$) dans une langue 2, sa corrélation avec chaque mot dans le dictionnaire bilingue d'une langue 2 est calculée. Cela donne un vecteur de relation ($WoRm_2$).

   Pour trouver les corrélations entre les mots dans le corpus monolingue : la corrélation ($W$) entre deux mots $w_s$ et $w_x$ est calculée à partir des scores de la mesure de vraisemblance (*likelihood*) des cooccurrences des mots dans des segments (des phrases, des paragraphes, etc). Les expériences réalisées dans l'article mènent à choisir

une taille de segment proportionnelle à la fréquence des mots dans le dictionnaire bilingue ($\propto \frac{1}{frequency(W_s)}$), où $W_s$ est l'ensemble des mots dans le dictionnaire bilingue. L'association entre $w_s$ et $w_x$ est calculée à partir de cooccurrences qui sont présentées dans la table de contingence (voir la table 2.4), où $w_1$=$w_s$, $w_2$=$w_x$, occ($w_1$, $w_2$) est le nombre d'occurrences de $w_1$ et $w_2$ ensemble, et $\neg w_1$ signifie tous les mots sauf $w_1$.

| | $w_2$ | $\neg w_2$ |
|---|---|---|
| $w_1$ | a=occ($w_1$, $w_2$) | b=occ($w_1$, $\neg w_2$) |
| $\neg w_1$ | c=occ($\neg w_1$, $w_2$) | d=occ($\neg w_1$, $\neg w_2$) |

TABLE 2.4 – *Table de contingence pour $w_1$ et $w_2$*

$a$, $b$, $c$, et $d$ sont calculés à partir des segments dans le texte monolingue du corpus comparable.

Les probabilités suivantes pour un mot $w_x$ et un autre $w_s$ sont utilisées pour calculer la corrélation finale $W$ entre ces deux mots (de même langue).

$$p(w_s = 1) = \frac{a+b}{a+b+c+d} \tag{2.3}$$

$$p(w_x = 1) = \frac{a+c}{a+b+c+d} \tag{2.4}$$

$$p(w_s = 1, w_x = 1) = \frac{a}{a+b+c+d} \tag{2.5}$$

où $p$ signifie probabilité.

La mesure d'information mutuelle (Fano 1961) est utilisée pour calculer la corrélation $W(w_s, w_x)$, qui est, selon Fung et Mckeown (1997), plus appropriée pour les mots et les termes qui ont des fréquences moyennes :

$$W(w_s, w_x) = p(w_x = 1, w_s = 1) \log_2 \frac{p(w_x = 1, w_s = 1)}{p(w_x = 1)p(w_s = 1)} \tag{2.6}$$

4. la similarité entre le mot $w_x$ de langue 1 et $w_c$ de langue 2 est calculée en utilisant la similarité cosinus entre leurs vecteurs, si elle a une valeur élevée, le mot $w_c$ sera considéré comme traduction de $w_x$.

Afin de comparer les vecteurs de relation de deux mots en deux langues différentes, le vecteur $WoRm_1$ (du mot $w_x$) est d'abord traduit à l'aide du dictionnaire bilingue.

Un exemple, qui représente les vecteurs de relation (appellation utilisée dans Fung et Mckeown (1997)) ou les vecteurs de contexte (appellation souvent utilisée dans la littérature) du mot EN *student* et du mot FR *étudiant*, est illustré dans la figure 2.1. Cet exemple démontre un processus pour calculer un score entre deux mots à partir de la similarité de leurs vecteurs. Un vecteur de contexte est construit pour le mot anglais ainsi que pour le mot français à partir de leurs cooccurrences. Les mots apparaissant dans le vecteur de contexte du mot anglais sont traduits à l'aide d'un dictionnaire bilingue. Au cas où plusieurs traductions (trouvées dans

le dictionnaire bilingue) existent pour un mot qui se trouve dans le vecteur de contexte, une seule traduction possible sera retenue. Un mot qui se trouve dans le vecteur de contexte mais pas dans le dictionnaire bilingue est ignoré. Une mesure (ex. cosinus dans Fung et Mckeown (1997)) est utilisée pour calculer la similarité entre le vecteur du mot anglais et celui du mot français.

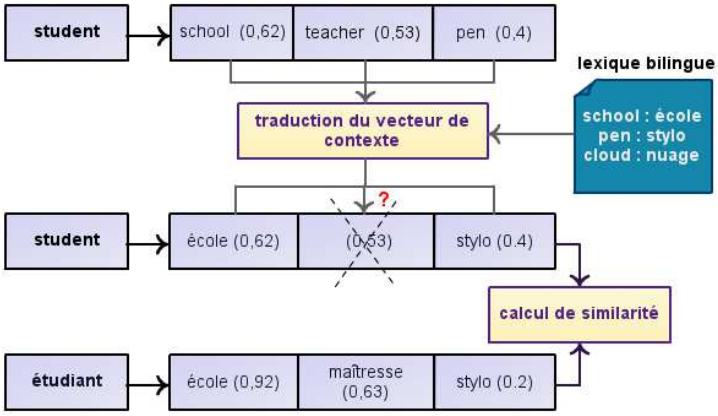


FIGURE 2.1 – *Traduction du vecteur de contexte du mot EN student et calcul de similarité du vecteur traduit avec le vecteur de contexte du mot FR étudiant*

Rapp (1999) améliore aussi sa première méthode (Rapp 1995) en utilisant un dictionnaire bilingue de base. À l'instar de Fung et Mckeown (1997), pour chaque langue, un vecteur de contexte est calculé pour chaque mot à partir de ses cooccurrences avec les mots du corpus apparaissant dans le dictionnaire bilingue de base. Une fenêtre de taille 3 est choisie afin de compter les cooccurrences des couples de mots se trouvant dans une fenêtre de la taille choisie. Mais, au lieu de calculer un seul vecteur pour un mot, plusieurs vecteurs sont calculés, c'est-à-dire un vecteur pour chaque position dans la fenêtre. Par exemple, pour un mot $A$, et une fenêtre d'une taille 2, quatre vecteurs sont calculés :

- un vecteur de cooccurrences dans le cas où $A$ apparaît deux mots avant un mot $B$.
- un vecteur de cooccurrences dans le cas où $A$ apparaît un mot avant un mot $B$.
- un vecteur de cooccurrences dans le cas où $A$ apparaît deux mots après un mot $B$.
- un vecteur de cooccurrences dans le cas où $A$ apparaît un mot après un mot $B$.

Les quatre vecteurs qui sont chacun de taille $n$ sont combinés dans un seul vecteur de taille $4n$ qui représente $A$ (vecteur de contexte). Tous les vecteurs de cooccurrences sont transformés par l'utilisation de la mesure de taux de vraisemblance (*loglikelihood ratio*, (Dunning 1993) [1]) afin de réduire l'influence des fréquences des mots sur les cooccurrences des couples de mots, calculée comme suit :

$$\begin{aligned} -2\log\lambda &= \sum_{i,j\in\{1,2\}} k_{i,j}\log\frac{k_{ij}N}{C_iR_j} \\ &= k_{11}\log\frac{k_{11}N}{C_1R_1} + k_{12}\log\frac{k_{12}N}{C_1R_2} + k_{21}\log\frac{k_{21}N}{C_2R_1} + k_{22}\log\frac{k_{22}N}{C_2R_2} \end{aligned}$$

1. Une autre version de la mesure plus rapide à calculer est utilisée dans Rapp (1999).

où :
- $C_1 = k_{11} + k_{12}$
- $C_2 = k_{21} + k_{22}$
- $R_1 = k_{11} + k_{21}$
- $R_2 = k_{12} + k_{22}$
- $N = k_{11} + k_{12} + k_{21} + k_{22}$

Les paramètres $k_{ij}$ sont exprimés à l'aide de la table de contingence 2.4 où $w_1$=$k_{11}$, $w_2$=$k_{12}$, $\neg w_1$=$k_{21}$ et $\neg w_2$=$k_{22}$.

Pour trouver la traduction anglaise d'un mot allemand ($m_s$), le vecteur de contexte de $m_s$ est calculé et comparé avec tous les vecteurs de contexte de tous les mots anglais à travers une mesure de similarité. Les mots en anglais sont ensuite ordonnés selon les similarités de leurs vecteurs avec le vecteur du mot $m_s$. Les $n$ premiers mots anglais peuvent ensuite être proposés comme traductions pour le mot $m_s$. Un dictionnaire allemand-anglais est utilisé pour traduire les mots du vecteur du mot allemand $m_s$ en anglais. Tous les mots qui n'ont pas de traductions dans le dictionnaire sont supprimés du vecteur. La valeur de la similarité entre deux vecteurs est calculée en utilisant la métrique de *city-block* ($= \sum_{i=1}^{n} |X_i - Y_i|$), qui a donné un meilleur résultat dans les expériences menées dans l'article. Dans la figure 2.2, le processus général de la traduction d'un mot source ($t_s$) par une approche directe basée sur les vecteurs de contexte est démontré.

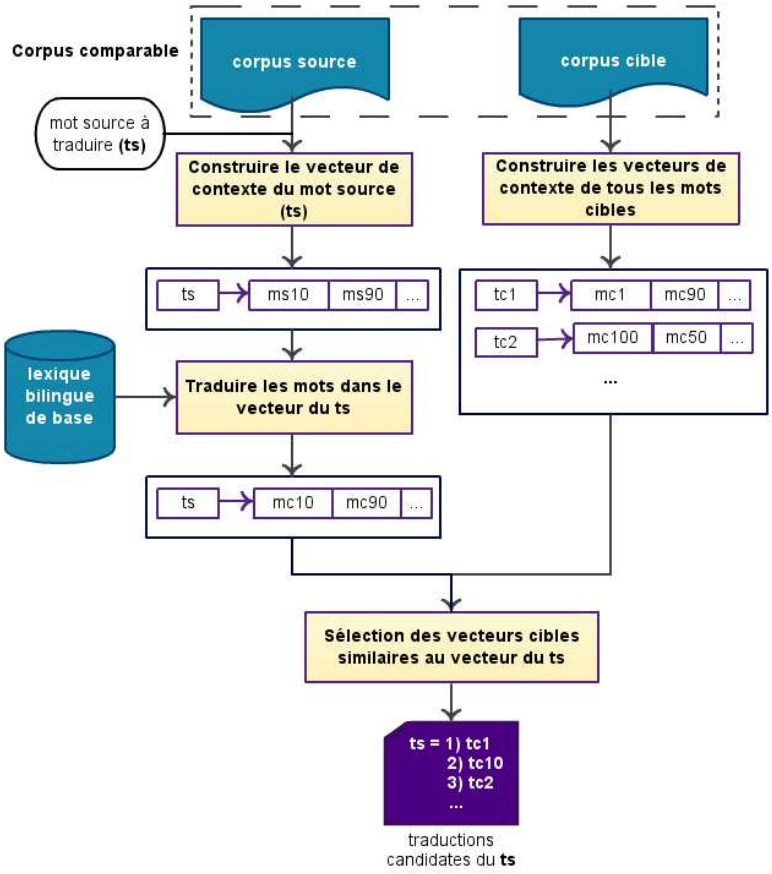


FIGURE 2.2 – *Traduction du mot source $t_s$ par une approche distributionnelle de base en utilisant les vecteurs de contexte*

### 2.1.3 Évaluation

Nous présentons des ressources utilisées pour ces différentes approches et les résultats obtenus.

#### Ressources

Les ressources utilisées par les quatre travaux décrits sont présentées dans la table 2.5. Nous listons pour chaque travail la paire de langues sur laquelle les expériences sont menées, les tailles des corpus comparables utilisés, les tailles des listes de références ainsi que le dictionnaire bilingue de base (au cas où il est exploité par la méthode).

#### Résultats

La qualité de résultats obtenus par les méthodes distributionnelles est souvent estimée par la précision. Cette précision est calculée à différents niveaux après avoir pris les $n$ meilleures traductions candidates pour chaque terme (top 1, top 5, etc.). Elle est définie comme étant le nombre des traductions correctes trouvées divisé par le nombre de termes sources dans la liste de références.

Rapp (1995) confirme par une simulation préliminaire avec 100 mots et leurs équivalents qu'il existe des corrélations entre les cooccurrences d'un mot et sa traduction dans les textes comparables. Les résultats des expériences menées par Fung (1995) montrent que la traduction correcte est trouvée parmi les 12 premiers candidats pour plus de 50 % des mots à traduire. Une précision de 21 % à 52 % est obtenue, avec la liste de 19 mots japonais, par Fung et Mckeown (1997) quand la première traduction candidate est retenue. Dans Rapp (1999), les traductions correctes des 100 mots allemands de référence sont trouvées dans 72 % des cas quand la première traduction est considérée comme acceptable.

Nous pouvons remarquer, à partir de la table 2.5, que Fung et Mckeown (1997) utilisent un corpus et un dictionnaire de plus petite taille par rapport au dictionnaire et aux corpus utilisés dans Rapp (1999). D'ailleurs, Fung et Mckeown (1997) travaillent sur une paire de langues plus éloignée que celle étudiée par Rapp (1999).

En effet, il n'est pas facile de comparer les différentes approches distributionnelles qui sont, en théorie, indépendantes des langues mais pas entièrement à cause de l'utilisation des ressources linguistiques.

Le contexte d'un mot est défini différemment d'une approche à l'autre. Rapp (1995), Fung (1995) et Rapp (1999) définissent le contexte d'un mot par les mots qui cooccurrent dans une fenêtre de taille définie centrée sur le mot, alors que Fung et Mckeown (1997) définissent le contexte d'un mot par les segments qui le contiennent.

De plus, chacune des approches utilise une mesure différente pour normaliser les fréquences de cooccurrences des couples de mots, comme par

| | **Langues** | **Corpus** | **Liste de références** | **Dictionnaire** |
|---|---|---|---|---|
| Rapp (1995) | anglais-allemand | source : 33 millions de mots (des textes de journaux, Brown Corpus) + cible : 46 million de mots (textes de journaux) | 100 mots | - |
| Fung (1995) | anglais-chinois | source : 22 147 de mots uniques + cible : 7 942 mots uniques (des transcriptions des débats du *Hong Kong Legislative Counsil* en anglais et en chinois) | 58 mots | - |
| Fung et Mckeown (1997) | japonais-anglais | source + cible : 16 millions de mots (des textes de journaux, *Wall Street Journal* pour l'anglais et *Nikkei Financal news* pour le japonais) | 19 mots | 1 415 entrées |
| Rapp (1999) | anglais-allemand | source : 163 millions de mots (du journal *The Guardian*, 1990-1994) + cible : 135 millions de mots (du journal *Frankfurter Allgemeine Zeitung*, 1993-1996) | 100 mots | 16 380 entrées |

TABLE 2.5 – *Ressources utilisées par les différentes approches distributionnelles*

exemple la mesure de taux de vraisemblance ou la mesure d'information mutuelle. Ces mesures qui peuvent être utilisées pour normaliser les fréquences de cooccurrences de deux mots sont appelées les mesures d'association. Une étude qui compare ces mesures et d'autres paramètres utilisés par les approches distributionnelles est réalisée par Laroche et Langlais (2010). Cette étude conclut, à partir des expériences menées, que l'utilisation de certaines mesures ou paramètres peut permettre d'améliorer ou de dégrader les résultats obtenus. Par exemple, la mesure d'association *log-odds* (décrite dans Evert (2005)) permet, selon Laroche et Langlais (2010), d'augmenter la précision des approches distributionnelles.

### 2.1.4 Discussion et améliorations

L'hypothèse de Rapp (1995) devrait fonctionner mieux quand elle est appliquée sur des corpus parallèles et moins bien sur les corpus comparables. En effet, les corrélations entre les cooccurrences des mots et les cooccurences de leurs traductions sont plus fortes dans les textes parallèles et moins fortes dans les corpus comparables. De plus, les approches distributionnelles fonctionnent mieux pour les mots assez fréquents dans le corpus. Pour cela, ces approches ont tendance à utiliser des corpus comparables de très grande taille. Dans la tâche d'extraction d'un lexique bilingue spécialisé à partir des corpus comparables, des corpus de très grande taille peuvent ne pas être disponibles pour certains domaines. Morin et al. (2008) estiment que la qualité de ces corpus peut compenser le manque de corpus de grande taille. Li et Gaussier (2010) s'intéressent à améliorer la comparabilité d'un corpus comparable pour l'extraction des lexiques bilingues en préservant le vocabulaire du corpus. Ils proposent une mesure pour estimer la comparabilité d'un corpus bilingue pour pouvoir en extraire une partie qui a un score de comparabilité élevé. Ils enrichissent ensuite la deuxième partie du corpus avec des documents trouvés sur le Web ou des documents parallèles (nous donnons plus de détails sur cette méthode en section 2.5).

Les recherches n'ont cessé d'améliorer la précision des lexiques bilingues obtenus à partir des distributions des mots. Ces recherches concernent soit l'alignement à partir des corpus de langue générale soit l'alignement à partir des corpus issus d'une langue de spécialité.

#### Dictionnaires bilingues de base

La précision obtenue en utilisant les approches de Fung et Mckeown (1997) et de Rapp (1999) dépend des dictionnaires bilingues utilisés. Une précision plus élevée est souvent obtenue si le dictionnaire couvre une grande majorité du vocabulaire du corpus. L'utilisation des thésaurus multilingues qui fournissent des synonymes, comme dans Déjean et al. (2002), permet d'augmenter la précision de l'alignement. Certaines informations supplémentaires sur les mots peuvent être exploitées pour améliorer la qualité du dictionnaire bilingue de base utilisé. Pour les langues

similaires, nous pouvons améliorer la qualité de l'alignement en repérant les cognats (c'est-à-dire les mots similaires orthographiquement ayant un sens similaire) entre deux langues dans les corpus (Koehn et Knight (2002)). Nous pouvons aussi exploiter le phénomène de translittération (populaire dans les domaines scientifiques) afin de trouver l'équivalent d'un mot dans une langue éloignée, comme pour la traduction du chinois vers l'anglais (Shao et Ng 2004).

**Catégories grammaticales**

Sadat et al. (2003) font état d'améliorations en supposant qu'un mot d'une catégorie syntaxique spécifique ne peut être traduit que par des mots qui ont des catégories correspondantes. Par exemple, ils supposent qu'un adjectif en anglais ne peut être traduit que par un adjectif ou un adverbe en japonais.

**Représentation des vecteurs de contexte**

Dans le but d'améliorer la représentation lexicale du vecteur de contexte d'un mot à traduire, des approches essayent de filtrer certains mots ou de donner des poids plus importants à certains mots dans le vecteur de contexte du mot à traduire. Prochasson et Morin (2009) proposent une approche adaptée à l'extraction d'un lexique bilingue des domaines de spécialité. Ils donnent des poids plus importants aux termes scientifiques (ex. composés savants) et aux translittérations dans le vecteur de contexte d'un mot. Ismail et Manandhar (2010) supposent que dans le vecteur de contexte d'un mot ($m$) à traduire : un mot qui a un score élevé doit être préservé seulement s'il a aussi un score élevé dans le vecteur de contexte d'autres mots associés à $m$. Ils appellent les mots qui vérifient cette condition par les termes spécifiques au domaine (*in-domain terms*). Les vecteurs de contextes des mots seront représentés par leurs termes spécifiques au domaine.

Un dictionnaire bilingue de base peut proposer plusieurs traductions lors de la traduction d'un mot qui se trouve dans le vecteur de contexte d'un mot à traduire. Bouamor et al. (2013) et Apidianaki et al. (2013) essayent de désambiguïser les mots polysémiques [2] qui se trouvent dans les vecteurs de contexte en sélectionnant les seules traductions pertinentes.

Dans la plupart des expériences menées sur l'extraction automatique d'un lexique bilingue à partir des corpus comparables, le contexte d'un mot est considéré comme un sac de mots et construit à partir des mots qui le précèdent et qui le suivent dans une fenêtre de taille fixe. Otero (2007) propose de construire le contexte d'un mot à partir des modèles de dépendance linguistiques. Par exemple, du modèle de dépendance linguistique EN [Nom + of + sugar] [3], nous pouvons extraire tous les noms qui précèdent EN [of sugar] afin de représenter le contexte du mot EN *sugar*. Ensuite, des correspondances entre des modèles de dépendance linguistique dans deux langues sont établies à partir d'un corpus parallèle. Otero (2007) fait l'hypothèse que si un mot est traduit par un autre, les deux ont

2. Les mots ayant plus d'une seule traduction dans un dictionnaire.
3. Où Nom est n'importe quel nom.

tendance à apparaître dans des contextes construits à partir des modèles linguistiques parallèles. Hazem et Morin (2013b) proposent d'exploiter la représentation du contexte d'un mot par ses dépendances syntaxiques ainsi que par des fenêtres centrées sur ce mot de manière conjointe. La combinaison des différentes représentations du contexte d'un mot ainsi que la combinaison des scores obtenus par des approches basées sur différentes représentations du contexte d'un mot, permettent d'améliorer la qualité d'un lexique bilingue spécialisé extrait (amélioration de la précision jusqu'à 10 %).

**Termes peu fréquents**

Il existe également des travaux qui se sont concentrés sur l'extraction des traductions pour les mots ou les termes peu fréquents. Prochasson et Fung (2011) utilisent un corpus comparable aligné au niveau de documents (un document peut être aligné avec plusieurs documents, les documents alignés peuvent partager le même thème par exemple) ; pour aligner un mot rare source avec un mot cible, ils calculent la similarité entre leurs vecteurs de contexte et ils exigent que ces mots apparaissent dans des documents alignés.

**Hypothèse distributionnelle**

D'autres approches apportent des modifications sur l'hypothèse de base de Rapp (1995) pour améliorer les lexiques extraits. Chiao (2004) suppose qu'un mot et sa traduction doivent avoir des contextes similaires lors de la traduction de la langue source vers la langue cible, ainsi que lors de la traduction de la langue cible vers la langue source.

## 2.2 Approches compositionnelles

Comme nous l'avons mentionné, les approches distributionnelles dépendent principalement des cooccurrences d'un mot dans un corpus afin de trouver ses traductions. Dans le cadre de la compilation automatique de lexiques bilingues spécialisés, ces approches peuvent aider à aligner les termes simples avec une bonne précision. Cependant, l'application de telles approches aux termes complexes donne des résultats peu satisfaisants ; un terme complexe apparaît généralement peu de fois dans un corpus (Morin et Daille (2010), Baldwin et Tanaka (2004)). Par exemple, Baldwin et Tanaka (2004) trouvent qu'environ 45-60 % des syntagmes de la forme [Nom + Nom] (une des formes possibles des termes complexes anglais) apparaissent une seule fois dans un corpus anglais. Toutefois, il est essentiel de trouver les équivalents des termes complexes car ils constituent une partie importante du vocabulaire terminologique d'un domaine de spécialité. En effet, les nouveaux termes ajoutés à un domaine sont plutôt de type complexe que de type simple (Sag et al. (2001)).

En dehors de leurs fréquences faibles dans un corpus, d'autres aspects des termes complexes rendent leur traduction automatique difficile (Baldwin et Tanaka (2004), Morin et Daille (2010)), nous en citons les suivants :

– **fertilité** : un terme source et son équivalent en langue cible peuvent

comprendre un nombre différent de mots. Par exemple, le terme EN *domestic wind turbine* (composé de trois mots) se traduit par FR *éolienne domestique* (composé de deux mots).

- **non-compositionnalité** : un terme complexe est non-compositionnel quand son sens ne peut pas être induit du sens de ses parties. Dans le cadre de la traduction, un terme complexe se traduit non-compositionnellement quand son équivalent dans une autre langue ne peut pas être obtenu en traduisant le terme complexe mot à mot. Par exemple, le terme FR *produits d'intérêts* se traduit par EN *interest income*, où EN *interest* est la traduction de FR *intérêts*, mais EN *income* n'est pas la traduction de FR *produits*.
- **variation de structure inter-langue** : un terme d'une structure spécifique dans une langue peut être traduit par un terme d'une structure différente dans une autre langue. Par exemple, le terme FR *consommation alcoolique* (de la forme [Nom + Adj]) est traduit par EN *alcohol consumption* (de la forme [Nom + Nom]).
- **variation syntaxique intra-langue** : un terme complexe peut apparaître sous différentes structures dans le même texte. Par exemple, FR *cancer du poumon* et sa variante FR *cancer pulmonaire* sont tous les deux traduits en anglais par *lung cancer*.

Dans cette section, nous décrivons les travaux de Grefenstette (1999), Tanaka et Baldwin (2003), Robitaille et al. (2006) et Vintar (2010) qui sont basés sur une propriété compositionnelle pour traduire les mots ou les termes complexes. Ils trouvent la traduction d'un composé syntagmatique ou d'un terme complexe à partir de leurs composants.

Dans le but d'extraire automatiquement un lexique bilingue spécialisé à partir d'un corpus comparable, une approche compositionnelle peut être considérée comme composée de trois étapes principales (voir figure 2.4) : (a) extraction de termes complexes : deux listes de termes complexes sources et cibles sont extraites des corpus ; (b) génération de traductions candidates : une liste de traductions candidates des termes complexes sources est générée à l'aide d'un lexique bilingue ; (c) sélection de traductions valides : des traductions valides parmi les traductions générées sont choisies en regardant dans la liste des termes complexes cibles.

Nous détaillons d'abord les étapes de génération de traductions candidates et sélection des traductions valides pour un terme complexe comme expliquées dans les travaux de Grefenstette (1999), Tanaka et Baldwin (2003), Baldwin et Tanaka (2004), Robitaille et al. (2006) et Vintar (2010). Ensuite, nous présentons et discutons les résultats obtenus.

### 2.2.1 Méthodes

Afin de traduire des termes complexes d'une langue source, des traductions candidates d'une langue cible sont générées pour ces termes complexes, et ensuite des traductions correctes sont choisies parmi celles qui sont générées.

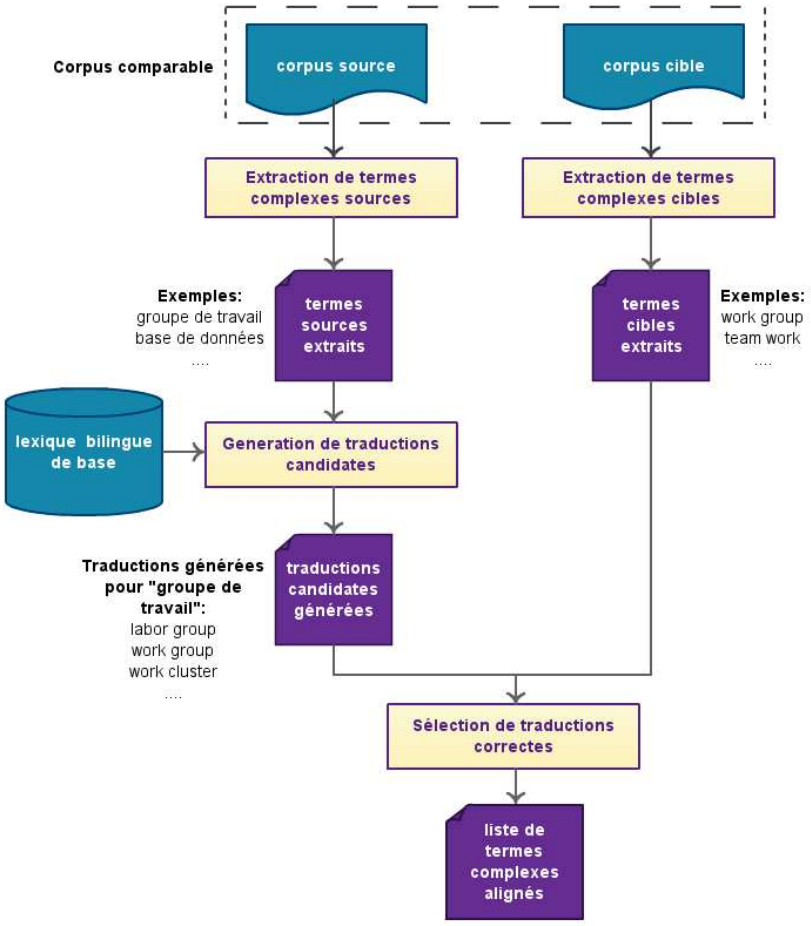


FIGURE 2.3 – *Représentation des trois étapes principales d'une approche compositionnelle de base pour l'extraction d'un lexique bilingue de termes complexes à partir d'un corpus comparable*

**Génération des traductions candidates** Cette étape consiste à générer des traductions pour des composés syntagmatiques ou des termes complexes sources. La génération est faite généralement en utilisant un lexique/dictionnaire bilingue qui peut être probabiliste (c'est-à-dire que chaque paire de traductions est dotée d'une probabilité).

Pour expliquer la génération de traductions candidates, Grefenstette (1999) prend le terme compositionnel *groupe de travail* comme exemple. Dans un dictionnaire français-anglais, le mot français *groupe* peut être traduit en anglais par *cluster*, *group*, *concern*, *grouping* ou *collective*. Le mot français *travail* peut être traduit en anglais par les mots *work*, *labor* ou *labour*. Le mot *groupe* a donc l'ensemble des cinq traductions possibles ($T_1$), et le mot *travail* a l'ensemble des trois traductions possibles ($T_2$). Toutes les combinaisons possibles entre un élément de l'ensemble $T_1$ et un autre de $T_2$ sont ensuite générées : *work group*, *labor group*, *labour group*, *work grouping*, etc.

Tanaka et Baldwin (2003) et Baldwin et Tanaka (2004) traitent les termes complexes qui ont la structure [Nom + Nom] (ex. EN *work group*). les traductions candidates d'un terme complexe source sont générées en traduisant d'abord individuellement les composants du terme. Ensuite, toutes

les combinaisons possibles des traductions individuelles obtenues sont projetées sur des modèles de traduction prédéfinis. Les modèles de traduction définissent les séquences des parties du discours des composants cibles ; ils reformulent le terme source de la forme [Nom + Nom] dans une forme de la langue cible. Par exemple, pour traduire un terme complexe japonais de la forme [$N_{s1}$ + $N_{s2}$][4] : ses composants $N_{s1}$ et $N_{s2}$ sont traduits individuellement, et les traductions obtenues en anglais sont projetées sur des modèles définis tels que [$N_{c1}$ + in + $N_{c2}$] et [$N_{c1}$ + $N_{c2}$]. Le nombre de traductions générées est de $O(mnt)$, où $m$ et $n$ sont les nombres de traductions possibles des mots $N_{s1}$ et $N_{s2}$ respectivement, et $t$ est le nombre de modèles de traductions.

Dans Robitaille et al. (2006), la phase de génération consiste à proposer des traductions candidates pour un terme complexe en le décomposant en éléments plus courts. Les éléments sont ensuite cherchés dans un dictionnaire bilingue. Par exemple, le terme français *base de données relationnelles* peut être décomposé en éléments suivants : ([base de] et [données relationnelles]), ou ([base de données] et [relationnelles]), ou ([base] et [de données relationnelles]), ou ([base] et [données] et [relationnelles]). En effet, plusieurs traductions peuvent être trouvées dans le dictionnaire pour chaque élément, toutes les combinaisons possibles des traductions de ces éléments sont générées. Par exemple, si deux traductions EN *database* et EN *basis* pour l'élément FR *base de données* sont trouvées dans le dictionnaire, et deux traductions EN *relationnal* et EN *related* pour l'élément FR *relationnelles*, les combinaisons possibles de ces traductions peuvent être : *relational database*, *relationnal basis*, *related database*, *related basis*, etc. Si aucune traduction n'est trouvée pour un élément composant dans le dictionnaire, la génération échoue.

L'approche de Vintar (2010) génère toutes les traductions possibles d'un terme complexe comme dans Grefenstette (1999) avec deux différences : (a) chaque traduction d'un composant a une probabilité (car un lexique bilingue probabiliste est utilisé) ; (b) si un des composants d'un terme complexe n'existe pas dans le lexique, il sera ignoré mais la traduction n'échoue pas (une traduction partielle sera proposée). La probabilité finale d'une traduction candidate est calculée dans Vintar (2010) par la somme des probabilités de tous les composants divisée par le nombre de composants.

**Sélection des traductions correctes** Cette étape consiste à sélectionner les traductions candidates générées dans l'étape précédente qui sont susceptibles d'être valides. Des scores peuvent être donnés aux traductions candidates afin de les ordonner.

Grefenstette (1999) et Robitaille et al. (2006) calculent les scores de traductions valides en fonction de leurs fréquences dans le corpus cible. Grefenstette (1999) utilise le Web comme corpus et un moteur de recherche qui permet de trouver les traductions candidates en les mettant entre guillemets (pour trouver la chaine exacte). Le score d'un candidat

4. $N_{s1}$ signifie le premier nom de la langue source dans le syntagme.

est sa fréquence sur le Web. Par exemple, pour le terme français source *groupe de travail*, Grefenstette (1999) démontre que la traduction candidate *work group* est beaucoup plus fréquente sur le Web que toutes les autres traductions candidates générées pour le terme source ; c'est-à-dire qu'elle est plus susceptible d'être la traduction correcte. Robitaille et al. (2006) cherchent les fréquences des traductions cibles dans un corpus cible construit à partir des articles Web trouvés en utilisant des mots-clés. La traduction candidate ayant la fréquence la plus élevée est choisie en tant que traduction valide du terme complexe de la langue source.

Tanaka et Baldwin (2003) s'appuient sur des éléments de preuve dans le corpus et des modèles de traduction pour calculer les scores des traductions candidates générées pour un terme complexe de la forme [Nom + Nom]. Le score d'une traduction candidate comprenant deux composants combinés selon un modèle de traduction $m$ (ex. les composants de la traduction candidate *relation to bandersnatch* sont combinés selon le modèle [Nom + to + Nom]) est calculé comme suit :

$$ctq(w_1^{L_2}, w_2^{L_2}, m) = \alpha \times p(w_1^{L_2}, w_2^{L_2}, m) + \beta \times p(w_1^{L_2}, m)p(w_2^{L_2}, m) + \lambda \times p(w_1^{L_2})p(w_2^{L_2})p(m) \quad (2.7)$$

Où $p$ signifie probabilité, $w_1^{L_2}$ et $w_2^{L_2}$ sont les traductions des composants du terme source et $\alpha + \beta + \lambda = 1$.

Les probabilités $p(w_1^{L_2}, m)$ et $p(w_2^{L_2}, m)$ capturent les caractéristiques de sous-catégorisation de $w_1^{L_2}$ et $w_2^{L_2}$ par rapport au modèle $m$. Par exemple, supposons que $w_1^{L_2}$=bandersnatch et $w_2^{L_2}$=relation, et que $p(w_1^{L_2}, w_2^{L_2}, m)$=0 pour tout $m$, le score *ctq* de *relation to bandersnatch* va être différent de celui de *relation on bandersnatch*, parce que les probabilités *p(relation, [relation to Nom])* et *p(relation, [relation on Nom])* sont différentes et ne sont pas forcément égales à zéro. En d'autres termes, même si *relation to bandersnatch* et *relation on bandersnatch* n'existent pas dans le corpus, on peut donner au premier un score plus élevé qu'au deuxième si *relation to* apparaît plus de fois dans le corpus que *relation on*.

Toutefois, le principal défaut de la méthode compositionnelle de Tanaka et Baldwin (2003) (et d'autres approches compositionnelles) est qu'elle considère toutes les traductions possibles d'un mot composant comme étant équiprobables. En réalité, il existe une variabilité considérable dans leur utilisation. Par exemple, le mot japonais *kiji* peut avoir deux traductions EN *article* et EN *item*, mais une de ces traductions (ex. *article*) est plus générale et doit avoir un score plus élevé. En conséquence, Baldwin et Tanaka (2004) gardent les mêmes traits utilisés pour calculer le score *ctq* (voir équation 2.7), mais utilisent aussi d'autres traits afin de sélectionner les traductions correctes parmi les candidats générés. Ces traits supplémentaires sont dérivés à partir des dictionnaires bilingues. L'un de ces traits est la fréquence d'un terme par rapport à un modèle $m$ dans tous les dictionnaires. Par exemple, à partir des termes complexes de la forme [Nom + Nom] et leurs traductions de la forme [Nom + Nom] dans les dictionnaires bilingues, nous pouvons compter le nombre de fois que *kiji* est traduit par *article*.

Dans Baldwin et Tanaka (2004), afin de choisir une traduction valide parmi les candidats générés pour un mot, un vecteur est construit pour chaque candidat à partir de ses traits (présentés ci-dessus). L'algorithme d'apprentissage TinySVM[5] est utilisé afin de décider si une traduction est valide ou non. L'ensemble d'entraînement est composé par les vecteurs des traductions candidates générées pour des termes sources (dont les traductions correctes sont connues). Les traits des vecteurs des traductions candidates représentant des traductions correctes pour leurs termes sources sont considérés comme des exemples positifs, et tous les autres traits des autres vecteurs sont considérés comme des exemples négatifs. TinySVM produit une classification binaire pour une traduction candidate, il retourne un score indiquant s'il est plus proche de la classe négative que de la classe positive. Les traductions candidates sont ensuite classées selon leurs scores.

Vintar (2010) calcule le score d'une traduction candidate à partir des probabilités attribuées à ses composants par un dictionnaire probabiliste. Les traductions candidates sont classées selon leurs probabilités et proposées comme traductions au terme complexe. Le dictionnaire probabiliste est obtenu à partir d'un corpus parallèle, un corpus comparable n'est donc pas utilisé par cette approche.

### 2.2.2 Évaluation

Les approches compositionnelles, décrites ci-dessus, utilisent principalement des lexiques/dictionnaires bilingues et un corpus (monolingue ou multilingue) comme ressources. Nous présentons les ressources utilisées par ces approches et les listes de références construites. Nous présentons aussi les résultats obtenus avec ces ressources.

#### Ressources

**Dictionnaires bilingues** Des dictionnaires bilingues sont nécessaires pour générer les traductions candidates.

Grefenstette (1999) utilise des dictionnaires allemand-anglais et espagnol-anglais de 37 600 entrées chacun. Des dictionnaires anglais-japonais composés de 150 000 à 400 000 entrées sont utilisés par Tanaka et Baldwin (2003) et Baldwin et Tanaka (2004). Robitaille et al. (2006) utilisent deux dictionnaires (français-japonais et japonais-français) d'environ 50 000 chacun. Ils étendent les dictionnaires en utilisant des synonymes d'un thésaurus japonais.

Vintar (2010) utilise un lexique bilingue slovène-anglais probabiliste. Il est extrait par l'outil d'alignement de mots *Twente word aligner*[6] à partir des corpus parallèles (de 130 000 à 280 000 mots chacun).

5. chasen.org/ taku/software/TinySVM/
6. http ://archive.is/CcSD

**Corpus** Un corpus source peut être utilisé pour extraire des termes sources à traduire. Un corpus cible peut être utilisé pour valider les traductions générées.

Grefenstette (1999) utilise le Web pour la validation des traductions candidates proposées pour un terme complexe. Il indique que la taille du Web en tant que corpus permet de surmonter le bruit introduit. Des corpus comparables anglais-japonais extrait de *BNC*, *Reuters* et *Mainichi Shimgrams* sont utilisés par Tanaka et Baldwin (2003) et Baldwin et Tanaka (2004). L'approche de Robitaille et al. (2006) se penche sur des couples de mots-clés dans deux langues pour extraire des corpus comparables français-japonais à partir de pages Web. Vintar (2010) utilise un corpus parallèle (d'environ 600 000 mots) aligné au niveau des phrases et constitué de sous-corpus spécialisés dans des domaines différents.

**Liste de références** Des liste de composés syntagmatiques ou de termes complexes sources sont extraits du corpus source ou des dictionnaires afin d'évaluer les approches proposées.

Grefenstette (1999) essaye de trouver les traductions des termes complexes qui se trouvent dans des dictionnaires bilingues. Les traductions de ces termes complexes sont ignorées afin d'essayer de les traduire automatiquement.

Dans Tanaka et Baldwin (2003), Baldwin et Tanaka (2004), Vintar (2010) et Robitaille et al. (2006), un modèle linguistique est défini pour chaque langue (source-cible) afin d'extraire une liste de composés syntagmatiques ou de termes complexes. Ensuite, cette liste est filtrée en utilisant un critère statistique. Par exemple, Tanaka et Baldwin (2003) extraient à partir d'un corpus des syntagmes de la forme [Nom + Nom] en ayant filtré les syntagmes qui ont des fréquences inférieures à 10 dans le corpus.

**Résultats**

La performance d'une approche pour la traduction des termes complexes dépend principalement des facteurs suivants : (a) qualité des dictionnaires bilingues utilisés ; (b) qualité des corpus comparables ; (c) qualité de l'extraction des termes sources et cibles.

Dans Robitaille et al. (2006), une précision de 92 % est obtenue mais avec un rappel de 40 %. L'utilisation d'un dictionnaire bilingue pour la traduction compositionnelle génère peu de candidats. Par conséquent, des thésaurus différents (qui fournissent des synonymes de mots) sont utilisés par Robitaille et al. (2006) afin d'augmenter le rappel de la traduction de termes complexes. L'utilisation des thésaurus montre une augmentation du rappel mais diminue la précision. Les thésaurus ne sont donc utilisés que si le mot à traduire (composant d'un terme complexe) n'existe pas dans le dictionnaire bilingue. De plus, une méthode d'amorçage est utilisée dans Robitaille et al. (2006) pour augmenter la couverture du corpus. Étant donné que la précision obtenue par l'utilisation d'un dictionnaire est

élevée, les *n* meilleures traductions générées par la méthode compositionnelle (en utilisant seulement un dictionnaire) sont insérées dans le système d'extraction de termes : ces traductions sont considérés comme des mots-clés et peuvent aider à étendre les corpus pour extraire plus de termes (les textes des corpus sont trouvés sur le Web en utilisant des mots-clés).

La traduction correcte est identifiée dans plus de 80 % des cas dans Vintar (2010), cependant seulement des corpus parallèles sont utilisés dans les expériences même si Vintar (2010) affirme que la méthode peut donner de bons résultats avec un corpus comparable. La précision rapportée dans Tanaka et Baldwin (2003) est de 60 %, une amélioration de plus de 15 % est atteinte dans Baldwin et Tanaka (2004) par l'utilisation en plus de traits et de la classification de TinySVM. Une précision de 86-87 % est obtenue par Grefenstette (1999) en utilisant le Web comme un corpus pour choisir la traduction correcte parmi des traductions candidates, cependant Grefenstette (1999) teste son approche avec des termes sources extraits d'un dictionnaire bilingue ce qui signifie que ce sont des termes valides (les autres travaux extraient les termes automatiquement du corpus). L'utilisation du Web, comme corpus, est connu pour avoir quelques désavantages : c'est un corpus déséquilibré où il y a beaucoup de redondance (Kilgarriff et Grefenstette (2001)).

Les méthodes de génération des traductions candidates proposée dans Robitaille et al. (2006) et Grefenstette (1999), échouent si aucune traduction de l'un des composants d'un terme complexe n'est trouvée dans le dictionnaire bilingue. Elles échouent aussi si aucune des traductions générées n'est trouvée dans le corpus cible. Les méthodes de calcul des scores proposées par Tanaka et Baldwin (2003), Baldwin et Tanaka (2004) et Vintar (2010) permettent de trouver des traductions même pour un terme contenant des composants qui sont absents du dictionnaire bilingue. Tanaka et Baldwin (2003) et Baldwin et Tanaka (2004) permettent de proposer des traductions candidates pour un terme même si aucune traduction générée pour ce terme n'est trouvée dans le corpus car elles se basent sur d'autres traits existant dans le corpus pour donner un score aux traductions générées. Cependant, l'augmentation du rappel se fait toujours au détriment de la précision.

### 2.2.3 Discussion

Les approches compositionnelles présentées dans cette section ne traitent pas tous les problèmes de traduction des termes complexes (voir les problèmes que nous avons présentés en section 2.2).

**Fertilité** L'approche de Robitaille et al. (2006) peut trouver les traductions des termes fertiles dans certains cas parce qu'elle regroupe les composants des termes complexes lors de la traduction. Par exemple, le terme EN *domestic wind turbine* (trois composants) peut être traduit correctement en français par *éolienne domestique* (deux composants) si (a) il est décomposé en éléments *domestic* et *wind turbine* ; (b) *domestic* est traduit par *domestique* ; (c) *wind turbine* est traduit par *éolienne*. Nous présentons dans

le chapitre 3 les approches de Cartoni (2009) et de Delpech et al. (2012) qui exploitent une propriété compositionnelle des termes morphologiquement construits. Ces approches permettent de traduire des termes simples (contenant une seule entité graphique) en des termes complexes (contenant deux ou plusieurs entités graphiques).

**Non-compositionnalité** Les approches compositionnelles ne peuvent pas identifier les termes complexes qui suivent une traduction non-compositionnelle. De plus, la phase de génération peut générer des traductions candidates dans le corpus cible mais qui ne sont pas des traductions acceptables pour le terme complexe source à traduire. Sag et al. (2001) évoquent ce problème sous le terme de sur-génération.

**Variantes** En outre, les approches compositionnelles qui ont été présentées, dépendent des lexiques/dictionnaires bilingues qui ne sont pas censés capturer toutes variations morphologiques. Par exemple, pour traduire le terme français *consommation alcoolique* par *alcohol consumption* en anglais, il faut que l'adjectif FR *alcoolique* soit traduit par le nom EN *alcohol* dans le lexique/dictionnaire bilingue. Nous avons examiné la traduction de *alcoolique* en anglais dans deux dictionnaires[7] en ligne, et nous avons trouvé que la traduction fournie pour *alcoolique* est l'adjectif EN *alcoholic*.

Pour résumer, un terme complexe peut être mal traduit (par une approche compositionnelle) si ce terme est fertile, non-compositionnel ou quand des traductions candidates sont sur-générées. La traduction peut échouer à proposer des traductions pour un terme complexe dans les cas suivants : (a) un des composants (ou éléments) du terme n'existe pas dans le dictionnaire bilingue utilisé ; (b) la traduction correcte du terme n'existe pas dans le corpus cible.

## 2.3 Approches compositionnelles étendues

Nous présentons dans cette section deux approches qui ont pour but d'améliorer le rappel d'une approche compositionnelle dans le cas où la traduction d'un terme complexe échoue à cause de : (a) la variation morphologique ; ou (b) l'absence de l'un des composants du terme dans le dictionnaire bilingue.

### 2.3.1 Utilisation des connaissances morphologiques

Pour extraire un lexique bilingue, le travail mené par Morin et Daille (2010) propose d'abord de réaliser une approche compositionnelle pour traduire les termes complexes avant de lancer l'approche contextuelle. Ils traitent les termes complexes se composant de deux mots et se trouvant

7. http ://dictionnaire.reverso.net/francais-anglais/,
http ://www.larousse.fr/dictionnaires/francais-anglais

dans un corpus comparable spécialisé. Des documents français et japonais sont extraits du Web (0,7 millions de mots français et 0,8 millions de mots japonais) dans le domaine médical. Des dictionnaires français-japonais disponibles sur le Web sont utilisés.

Des termes complexes sources sont extraits par un programme d'extraction de termes *ACABIT*[8]. Pour traduire ces termes, ils proposent une modification sur la méthode compositionnelle de Robitaille et al. (2006) (présentée en section 2.2). Ils essayent d'associer un composant inconnu à un mot dans le dictionnaire en utilisant des connaissances morphologiques, au lieu de l'ignorer lorsqu'il n'apparaît pas dans le dictionnaire. Il s'agit de relier les adjectifs relationnels et les noms dérivés à leurs noms de base (ex. sanguine avec sang). Ensuite, la traduction d'un adjectif relationnel ou un nom dérivé sera considéré comme identique à la traduction de son nom de base. Par exemple, si le terme complexe FR *groupe sanguin* n'a pas pu être traduit en anglais parce que FR *sanguin* n'existent pas le dictionnaire bilingue, *sanguin* sera relié à FR *sang* et la traduction de ce dernier (ex. EN *blood*) sera attribuée à *sanguine*. La traduction correcte du terme *groupe sanguine*, qui est EN *blood group* peut être ensuite trouvée.

Des règles ont été proposées, nommées *règles de désuffixation recodage*, afin de transformer les adjectifs relationnels et les noms dérivés en forme neutre. Par exemple, l'adjectif FR *forestière* peut être relié au nom FR *forêt* en appliquant la règle suivante : M($Adj$, $N$) = [-estière, -êt].

Les résultats montrent une amélioration de la traduction des termes complexes dont la forme est [Nom + AdjR], où AdjR est un adjectif relationnel. Les adjectifs relationnels en français sont souvent traduits par des noms dans les autres langues. L'évaluation montre que les termes complexes français de la forme [Nom + AdjR] sont traduits en japonais avec une précision de 88 %.

Nous reprenons ces idées au chapitre 4 pour traduire des termes d'une structure [Nom + AdjR]. Nous remplaçons les règles manuellement définies par une méthode qui permet d'extraire et d'aligner automatiquement les adjectifs relationnels en corpus.

### 2.3.2 Utilisation des alignements d'une méthode distributionnelle

Les approches compositionnelles échouent parfois pour la traduction d'un terme complexe même si ce terme possède une propriété compositionnelle. Comme nous l'avons mentionné, il se peut qu'un composant du terme complexe soit absent du dictionnaire bilingue. Dans ce cas, Morin et Daille (2012b) proposent de substituer ce mot par son vecteur de contexte monolingue (la notion de vecteur de contexte est présentée en section 2.1.2). Par exemple, supposons que nous avons le terme complexe FR *antécédent familial* sachant que FR *familial* se traduit par EN *family* et que *antécedent* soit absent du dictionnaire bilingue. Pour traduire ce terme complexe, nous cherchons d'abord le vecteur de contexte de *antécédent*.

8. http ://www.bdaille.com/index.php ?option=com$_c$$ontenttask = blogcategoryid = 5Itemid = 5

Ensuite, nous comparons ce vecteur avec tous les vecteurs de contexte des mots (dans la langue cible) qui composent des termes complexes de la forme [family + Mot][9]. Le composant *antécédent* sera donc relié à tous les mots qui forment un terme avec *family*, avec des scores de similarité. S'il existe deux termes complexes qui comprennent *family* dans le corpus cible : EN *family history* et EN *family story*, le mot FR *antécédent* sera aligné avec EN *history* et EN *story* avec des scores différents. Les alignements sont ensuite classés selon ces scores et utilisés pour la traduction compositionnelle du terme complexe *antécédent familial*.

Morin et Daille (2012b) testent cette méthode sur les paires de langues français-anglais et français-allemand avec des corpus comparables (dans le domaine médical) de 530 000 et de 220 000 mots respectivement. Les bonnes traductions sont trouvées dans 55 % des cas pour le français-anglais et dans 49 % des cas pour le français-allemand, quand les 5 premières traductions candidates d'un terme complexe sont retenues. Ils obtiennent donc des précisions moins élevées que les précisions (79 % et 95 %) obtenues par une méthode compositionnelle qui dépend seulement d'un dictionnaire bilingue. Cependant, l'utilisation des alignements permet d'augmenter le rappel de 15 % à 55 % pour le français-anglais et de 9 % à 53 % pour le français-allemand. De plus, toutes les traductions trouvées par la méthode compositionnelle sont retenues par la méthode compositionnelle étendue.

## 2.4 Extraction de segments parallèles

Nous présentons dans cette section des approches qui s'intéressent à enrichir des corpus parallèles avec des segments ou des phrases parallèles extraits automatiquement des corpus comparables.

Les textes parallèles constituent des ressources importantes pour les systèmes de traduction automatique statistique (TAS), mais ils restent rares pour certaines paires de langues ou dans certains domaines. Les textes parallèles utilisés par les systèmes de TAS appartiennent souvent au même domaine. Un système de TAS donne souvent de moins bons résultats pour un autre domaine que celui sur lequel le système a été entraîné.

Au début des années 2000, l'extraction des correspondances bilingues à partir d'un corpus comparable a pris une nouvelle direction. Il s'agit d'extraire des phrases (ou segments) parallèles acquises de corpus comparables pour faire face au problème de taille et de disponibilité des corpus parallèles. Les phrases extraites d'un corpus comparable peuvent enrichir des corpus parallèles. Afin de les extraire, une approche de base (*brute force*) consiste à calculer une similarité entre toutes les paires de phrases dans un corpus comparable. Cette approche pose un problème de consommation de mémoire et de temps d'exécution. Pour résoudre cela, il faut d'abord réduire l'espace de recherche de paires de phrases parallèles. Par exemple, en alignant les documents d'un corpus comparable, la

9. Mot est n'importe quel mot apparaissant après le mot EN *family*.

recherche de paires de phrases parallèles ne se fait qu'entre les documents alignés. Ensuite, une similarité est calculée entre une phrase source et un sous-ensemble de phrases cibles. Selon les similarités obtenues entre les phrases, ces dernières peuvent être classées comme non-parallèles, comparables, quasi-parallèles ou parallèles.

### 2.4.1 Extraction de phrases parallèles

Nous commençons par expliquer l'approche de Munteanu et Marcu (2005) qui a pour objectif d'extraire des phrases parallèles, elle se compose de trois étapes :

1. alignement des documents : les documents qui sont similaires entre un corpus monolingue source et un corpus monolingue cible sont alignés. Pour un document source, l'approche essaye de trouver un ensemble de document cibles similaires : une requête est formée en prenant les cinq meilleurs traductions pour chaque mot dans le document source. Ensuite, le système de recherche d'information *In-Query* (Callan et al. (1994)) est utilisé pour trouver les 100 documents cibles les plus similaires. Finalement, seulement les documents cibles qui ont été publiés dans les 5 jours avant ou après la date de publication du document source sont retenus.
2. sélection des phrases parallèles candidates : cette étape consiste à d'abord générer toutes les paires de phrases possibles entre un document source et l'ensemble des documents cibles associés. Les phrases sont ensuite filtrées au moyen d'un filtre appelé *word overlap filter*. Deux phrases sont considérées comme candidates à être parallèles si la moitié des mots dans chaque phrase est alignée.
3. chaque paire de phrases candidates est passée par un ME *classifieur d'entropie maximale* qui décide si les phrases sont parallèles ou non.

L'approche de Munteanu et Marcu (2005) est illustrée dans la figure 2.4. Le ME classifieur utilise des fonctions de traits (*feature functions*) qui peuvent être trouvées après que les mots entre les phrases soient alignés à l'aide d'un dictionnaire bilingue probabiliste. Les traits sont les suivants :

- le pourcentage du nombre de mots qui ont des alignements.
- les longueurs des phrases, la différence entre les longueurs des phrases ainsi que le ratio des longueurs des phrases.
- les trois fertilités les plus longues sont utilisés en tant que traits. Un mot est fertile s'il est aligné avec deux mots ou plus. La fertilité des mots est un signe de non-parallélisme entre deux phrases (Brown et al. 1993).
- le score d'alignement entre les deux phrases, qui est la somme de probabilités des mots alignés entre les deux phrases.
- la longueur de la plus longue paire de chaînes où les mots dans une chaîne sont alignés uniquement avec des mots dans l'autre chaîne (*longest contiguous span*).

Pour pouvoir calculer ces traits entre deux phrases, il faut d'abord obtenir un alignement optimal 1 :1 (un mot est aligné avec un seul mot). Le score d'un alignement entre deux phrases est calculé en multipliant les

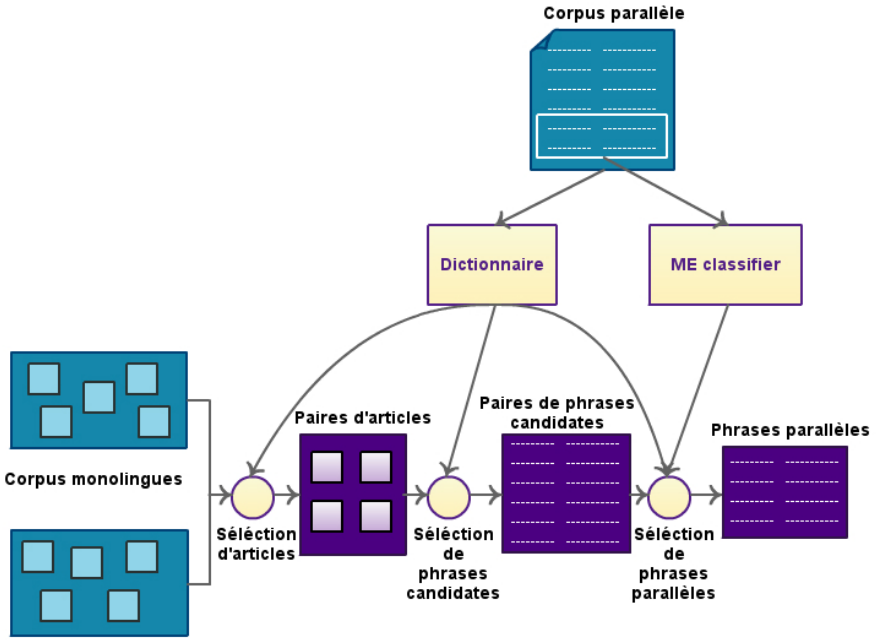


FIGURE 2.4 – *Système d'extraction des phrases parallèles comme illustré dans Munteanu et Marcu (2005)*

probabilités normalisées (modèle IBM 1 Moore (2004)) d'alignements de mots entre les deux phrases. Le meilleur alignement (de mots) possible entre deux phrases est celui qui maximise le score d'alignement pour ces phrases.

Les paramètres du classifieur ME sont estimés en entraînant le modèle sur un corpus parallèle. L'ensemble d'entraînement est construit comme suit : des phrases (sources et cibles) sont extraites d'abord à partir d'un corpus parallèle. Toute les paires de phrases (source-cible) possibles sont générées. Seulement les phrases qui passent le filtre *word overlap filter* de sélection de phrases candidates sont conservées. Ensuite, pour une phrase source, deux phrases cibles au maximum sont préservées : sa traduction correcte et une seule phrase non-parallèle, afin d'assurer d'avoir un ensemble d'entraînement équilibré.

Après avoir estimé les paramètres du classifieur, des phrases parallèles sont extraites du corpus comparable. Munteanu et Marcu (2005) montrent que les phrases parallèles (extraites d'un corpus comparable) peuvent améliorer la performance d'un système de traduction automatique de manière significative.

Abdul-Rauf et Schwenk (2009) présentent une méthode pour l'extraction des phrases parallèles de haute qualité. Ils se basent sur l'approche de Munteanu et Marcu (2005), et utilisent des techniques de recherche d'information ainsi que des métriques d'évaluation utilisée pour les systèmes de TAS afin d'évaluer la qualité des phrases extraites. Par exemple, ils utilisent la métriques WER (*Word Error Rate*) qui est basée sur le nombre d'opérations nécessaires (suppression, substitution ou insertion) pour transformer une phrase source traduite en une autre cible, ainsi que la métrique TER (*Translation Edit Rate* : le déplacement d'une séquence

contiguë de mots de distance arbitraire). En analysant les erreurs, Abdul-Rauf et Schwenk (2009) trouvent qu'il se peut qu'une paire de phrases soit parallèle mais qu'une des deux phrases contienne plus d'informations que l'autre à la fin. Pour cela, ils enlèvent ces mots supplémentaires à la fin de la phrase à l'aide de la métrique TER et montrent que ces phrases parallèles filtrées (où les mots supplémentaires sont enlevés) sont plus utiles que les phrases parallèles non-filtrées dans un système de TAS.

### 2.4.2 Extraction de segments parallèles

Le travail de Munteanu et Marcu (2005) ne peut extraire que des phrases complètes susceptibles d'être parallèles. Munteanu et Marcu (2006) étendent cette méthode pour extraire des segments parallèles qui ne sont pas forcément des phrases. L'approche consiste à considérer la phrase cible comme un signal numérique, où les mots alignés (avec des mots dans la phrase source) correspondent à des valeurs positives, et les autres mots correspondent à des valeurs négatives. Le signal est passé sur un filtre de lissage : la valeur à chaque point (mot) est la moyenne de plusieurs valeurs qui l'entourent (5 mots avant et 5 mots après). Ensuite, la partie du signal qui est la plus souvent positive est préservée. Cette partie du signal correspond à un segment dans la phrase cible qui sera aligné avec le segment correspondant dans la phrase source.

### 2.4.3 Extraction de phrases très comparables

Stefanescu et al. (2012) présentent une approche afin d'extraire des phrases parallèles, non-parallèles ou très comparables. Leur approche n'exige pas d'avoir un corpus aligné au niveau des documents. D'abord ils indexent les phrases cibles dans une base de données. Pour trouver les phrases cibles alignées avec une phrase source, il faut d'abord réduire l'espace de recherche. Pour cela, deux étapes sont appliquées :

1. pour chaque phrase source, un moteur de recherche retourne une liste de phrases cibles candidates. Afin de trouver cette liste, chaque mot dans la phrase source est traduit par un dictionnaire bilingue probabiliste (extrait d'un corpus parallèle avec l'outil d'alignement de séquences de mots *Giza++* [10] (Och et Ney 2000)). Les traductions des mots forment une requête utilisée par le moteur de recherche.
2. la liste de phrases candidates retournée par le moteur de recherche est ensuite filtrée selon une fonction définie à partir des longueurs des phrases, le nombre de mots en commun, la moyenne des distances des positions entre les mots alignés, etc.

Après avoir réduit l'espace de recherche, une mesure est proposée pour calculer la similarité entre une phrase source et chaque phrase cible dans la liste des phrases candidates. Cette mesure est la somme pondérée de plusieurs fonctions (où chacune $\in [0,1]$). D'abord, l'alignement optimal (un mot est aligné avec un seul mot) est calculé, c'est celui qui maximise la somme des probabilités de traduction des mots entre les deux phrases.

10. code.google.com/p/giza-pp/

À partir de cet alignement, d'autres fonctions sont définies en exploitant les probabilités de traductions des mots, les alignements des mots vides ou les positions des mots alignés entre les deux phrases.

Les poids des fonctions sont trouvés à l'aide d'un algorithme d'apprentissage *logistic regression classifier*. L'ensemble d'entraînement (*training set*) consiste en 9 500 exemples positifs (phrases parallèles) et en 9 500 exemples négatifs (phrases non-parallèles) extraits du corpus parallèle. L'ensemble de test (*test set*) se compose de 500 phrases parallèles et de 500 phrases non-parallèles.

En suivant cette approche, des phrases alignées sont extraites et insérées dans un système de TAS. Les améliorations obtenues sont comparables avec celles obtenues par l'approche de (Munteanu et Marcu 2005).

Il existe d'autres approches qui permettent d'extraire des phrases/segments parallèles à partir des corpus *très non-parallèles* (ex. (Fung et Cheung 2004)) ou des corpus alignés au niveau des documents (ex. articles extraits de Wikipedia (Smith et al. 2010)).

## 2.5 Comparabilité des corpus

Le degré de comparabilité des corpus comparables joue un rôle important sur la qualité des résultats obtenus, pour n'importe quelle méthode d'extraction de correspondances bilingues à partir de ces corpus. Comme nous l'avons mentionné dans le chapitre 1, la notion de comparabilité d'un corpus bilingue est vague car elle met en jeu des critères linguistiques ou extra-linguistiques difficilement mesurables.

Certains travaux regroupent les corpus bilingues selon trois catégories :
- parallèle : regroupe des textes étant des traductions mutuelles (ex. les corpus du parlement européen *Europal*).
- comparable : regroupe des textes traitant le même sujet. Les textes peuvent être alignés au niveau des documents (ex. des articles des journaux en anglais et en français qui traitent de la guerre en Syrie dans une période définie).
- peu comparable : regroupe des textes traitant le même domaine ou sous-domaine (ex. articles scientifiques en anglais et en français traitant le sujet du cancer du sein).

Cette catégorisation n'est pas suffisante pour un nombre d'applications du TALN qui nécessitent une estimation de la comparabilité des corpus bilingues. Des méthodes pour mesurer le degré de comparabilité (ou de similarité) d'un corpus bilingue ont été proposées. Les mesures proposées se basent notamment sur la similarité lexicale entre les parties du corpus. Elles peuvent être calculées au niveau de documents (ex. (Fung et Cheung 2004)) ou au niveau de corpus (ex. Li et Gaussier (2010)).

Saralegi et al. (2008) estiment que le degré de similarité entre deux corpus dépend de plusieurs critères : les sujets des documents, les dates de publication, la taille de corpus, etc. De plus, ils supposent que ces critères dépendent de la tâche et de la méthode pour lesquelles les corpus sont

utilisés. Ils définissent une mesure qui calcule la similarité globale du corpus à partir des scores de similarité de contenu (ces scores sont calculés en utilisant un lexique bilingue) entre toutes les paires de documents.

Un autre travail, celui de Li et Gaussier (2010), considère deux corpus comme comparables s'ils ont une partie importante de vocabulaire en commun. La comparabilité est définie comme étant « *l'espérance de trouver la traduction d'un mot source/cible donné dans le vocabulaire du corpus cible/source* ». La mesure ($M$) est calculée en utilisant un dictionnaire bilingue (source-cible) et deux corpus (source $S$ / cible $C$), elle est définie comme suit :

$$M(S,C) = \frac{a+b}{a'+b'} \tag{2.8}$$

où :

$a$= le nombre de mots sources dans le vocabulaire commun entre le corpus source et la partie source du dictionnaire, où chaque mot a au moins une traduction (dans le dictionnaire) qui existe dans le corpus cible.

$b$= le nombre de mots cibles dans le vocabulaire commun entre le corpus cible et la partie cible du dictionnaire, où chaque mot a au moins une traduction (dans le dictionnaire) qui existe dans le corpus source.

$a'$ = le nombre de mots sources dans le vocabulaire commun entre le corpus source et la partie source du dictionnaire.

$b'$ = le nombre de mots cibles dans le vocabulaire commun entre le corpus cible et la partie cible du dictionnaire.

À l'aide de cette mesure, Li et Gaussier (2010) extraient une partie ($C_1$) du corpus comparable ($C$) qui a une comparabilité supérieure à un certain seuil. Cela est réalisé en ajoutant à $C_1$ de manière itérative la paire de documents qui a le score de comparabilité le plus élevé. Avant d'ajouter une paire de documents à $C_1$, on vérifie si la paire ajoutée ne dégrade pas le score de comparabilité de $C$ en dessous d'un seuil minimal prédéfini. Après avoir extrait une partie du corpus très comparable, la partie restante ($C_2$) est enrichie avec des documents parallèles à partir du Web. Il est montré que la qualité d'un lexique bilingue obtenu avec une méthode distributionnelle est meilleure avec $C_1$ ou $C_2$ qu'avec $C$.

Su et Babych (2012) relient la comparabilité de textes à leur potentiel d'améliorer la performance de systèmes de traduction automatique statistique. La comparabilité d'un corpus dépend donc du potentiel d'un système de TAS d'extraire des traductions parallèles ou quasi-parallèles (des phrases ou des mots) à partir de ce corpus. Ils proposent trois métriques pour mesurer la comparabilité d'un corpus :

- métrique basée sur l'alignement lexical : des dictionnaires bilingues sont extraits à partir des corpus parallèles en utilisant une méthode statistique d'alignement des mots avec l'outil d'alignement des mots GIZA++. Ces dictionnaires seront ensuite utilisés pour aligner les paires de documents mot-à-mot. Enfin, la similarité cosinus sera utilisée pour calculer la comparabilité entre les paires de documents.

- métrique basée sur les mots clés : les poids des mots dans un document sont calculés en utilisant le TF-IDF [11], les mots qui ont les poids les plus élevés sont considérés comme des mots-clés. Enfin, chaque document sera représenté par sa liste de mots-clés, et la similarité cosinus sera appliquée sur les vecteurs de listes pour calculer la comparabilité entre les paires de documents.
- métrique basée sur la traduction automatique : en utilisant les deux premières métriques, les documents seront considérés comme des listes de mots et seulement les mots apparaissant dans le dictionnaire seront retenus. Par conséquent, les mots non-connus seront supprimés et des informations telles que l'ordre de mots ou la structure syntaxique ne peuvent pas être préservées. De plus, les dictionnaires bilingues ne peuvent pas être obtenus pour certaines paires de langues peu dotées. Pour cela, les auteurs traduisent en anglais les textes d'un corpus non-anglais par un système de traduction statistique. Ensuite, ils utilisent plusieurs traits pour calculer la similarité entre deux documents (traduits en anglais) : trait de mots-clés, trait de mots alignés, trait d'entités nommées et trait de structure. Par exemple, le trait pour mesurer la similarité de la structure entre deux documents est défini en fonction de la comparabilité du nombre de verbes, d'adverbes, et de noms, ainsi que du nombre de phrases entre les documents.

Ensuite, les auteurs classifient les paires de documents selon leurs scores dans des catégories : parallèle, très comparable ou un peu comparable. Les phrases parallèles sont ensuite extraites des documents considérées comme parallèles ou très comparables. Su et Babych (2012) améliorent la qualité d'un corpus comparable en négligeant les documents qui sont peu comparables, ce qui permet d'améliorer les méthodes d'extraction des phrases parallèles.

Liu et Zhang (2012) proposent une méthode adaptée aux corpus spécialisés, ils supposent que la comparabilité de ces corpus reposent sur la distribution et la qualité de la terminologie. Un mot est donné un score de spécificité dans le corpus spécialisé pour mesurer à quel point ce mot est spécifique au domaine du corpus. La spécificité d'un mot par rapport à un domaine est calculée à l'aide de la comparaison des fréquences d'un mot dans deux corpus : le premier issu de la langue spécialisée et le deuxième issu de la langue générale. Chaque corpus est donc représenté par les mots triés selon leurs scores de spécificité. Les mots qui ont des scores élevés auront des poids plus élevés lors du calcul de cosinus entre les vecteurs de corpus.

Toutes ces méthodes, qui proposent des mesures pour calculer le degré de comparabilité d'un corpus, relient l'amélioration de la comparabilité d'un corpus avec la tâche pour laquelle le corpus sera utilisé. Les approches qui calculent une similarité entre les vecteurs de documents de corpus représentés par leurs mots peuvent être couteuses si les corpus sont de grande taille. La mesure de Li et Gaussier (2010) est rapide et fa-

11. Term frequency-inverse document frequency d'un terme est une méthode de pondération souvent utilisée en recherche d'information.

cile à calculer mais elle dépend fortement de la qualité du lexique bilingue utilisé pour calculer la comparabilité entre deux corpus.

## 2.6 Outil d'extraction et d'alignement de termes : TTC TermSuite

Certaines des approches décrites dans ce chapitre sont implémentées dans l'outil TTC TermSuite [12]. Nous utilisons cet outil pour les expériences que nous menons dans les chapitres 3,4 et 5. TTC TermSuite permet de réaliser le pré-traitement du corpus, l'extraction de termes simples ou complexes et la traduction de termes à partir de corpus comparables.

TTC TermSuite est un logiciel librement distribué pour l'extraction et l'alignement terminologique multilingue à partir de corpus comparables spécialisés. Il peut traiter sept langues : anglais, français, allemand, espagnol, letton, russe et chinois. Il est développé en UIMA [13] (*Unstructured Information Management Architecture*), qui est une plateforme libre et extensible conçue pour développer des applications analytiques de traitement de textes.

TTC TermSuite comprend trois composants principaux :

1. Spotter : il effectue le pré-traitement du corpus, c'est-à-dire, découpage des textes en mots, étiquetage en partie du discours et lemmatisation de chaque mot. Cela est réalisé à l'aide de l'outil TreeTagger [14].
2. Indexer : il extrait des termes simples et des termes complexes du corpus. Il est aussi capable de regrouper des variantes de termes selon des modèles linguistiques bilingues (ex. regrouper le terme *cancer du poumon* avec le terme *cancer pulmonaire* à partir du modèle [Nom + préposition + Nom] : : [Nom + Adjectif Relationnel]) ou selon des mesures de distance/similarité (ex. regrouper le terme *tumeur* avec *tumeurs* par la mesure de Levenshtein). Levenshtein est une mesure de distance d'édition (*edit distance*) qui permet de calculer le nombre minimum de modifications (opérations) nécessaires pour transformer une chaîne en une autre. Les opérations autorisées sont l'insertion, la suppression et la substitution d'un seul caractère. Le coût de chaque opération est égal à 1. Des mesures qui permettent de calculer la similarité ou la distance entre deux chaînes sont présentées dans le paragraphe *Travaux connexes* en section 4.4 du chapitre 4.
3. Aligner : ce composant est capable de proposer des traductions à partir d'un corpus comparable pour des termes sources que nous voulons traduire.

   Les termes simples peuvent être traduits par la méthode distributionnelle de Rapp (1999) (présentée en section 2.1). Les paramètres par défaut pour la méthode distributionnelle dans TermSuite sont

12. https ://code.google.com/p/ttc-project/
13. http ://uima.apache.org/
14. http ://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

ceux qui donnent les meilleurs résultats avec le corpus cancer du sein français (ces paramètres sont trouvés à l'aide des expériences). Une fenêtre de taille 3 et une mesure de vraisemblance sont choisies pour calculer les valeurs de cooccurrences de mots dans le vecteur de contexte d'un mot. La mesure de Jaccard (Grefenstette 1994) est utilisée pour calculer la similarité entre deux vecteurs de contexte.

Les termes complexes seront alignés en appliquant une méthode compositionnelle de base. Cette méthode est celle de (Grefenstette 1999) (présentée en section 2.2) après avoir replacé le corpus du Web par un corpus comparable. Les méthodes compositionnelles étendues de Morin et Daille (2010; 2012b) (présentées en section 2.3) sont aussi implémentées dans ce composant et peuvent être utilisées pour aligner les termes complexes.

## 2.7 Conclusion

Dans ce chapitre, nous avons présenté des méthodes de l'état de l'art pour : l'extraction des traductions de mots/termes (approches distributionnelles et compositionnelles), l'extraction des segments parallèles et l'estimation de la comparabilité d'un corpus comparable.

Les contributions de cette thèse portent sur la traduction des termes à partir d'un corpus comparable. Nous exploitons le principe d'une approche compositionnelle de base ainsi que le principe des approches compositionnelles étendues présentées dans ce chapitre pour traduire les composés savants et les termes complexes d'une structure [Nom + Adjectif]. Nous cherchons à améliorer la qualité des résultats obtenus par l'approche distributionnelle implémentée dans l'outil TTC TermSuite (présenté en section 2.6) en s'inspirant des travaux concernant l'extraction des segments parallèles.

Dans le chapitre suivant, nous nous basons sur le principe d'une approche compositionnelle pour traduire les composés savants ayant des éléments qui n'existent pas dans les dictionnaires bilingues. Nous testons également une méthode inspirée du travail de Morin et Daille (2012b) (présentée en section 2.3.2) pour la traduction de composés savants.

# Traduction des composés savants

3

Sommaire



Dans ce chapitre, nous nous intéressons à la traduction automatique des composés savants à partir des corpus comparables. Nous considérons qu'un terme simple est un composé savant s'il contient au moins une racine gréco-latine combinée avec d'autres racines gréco-latines ou un mot. Par exemple, *aérogénérateur* est un composé savant parce qu'il comprend la racine gréco-latine *aéro* et le mot *générateur*. Nous faisons l'hypothèse qu'un composé savant dans une langue source peut être traduit de manière compositionnelle ou semi-compositionnelle par un composé savant dans une langue cible. Par exemple, *aérogénérateur* se traduit par *aerogenerator* en anglais et par *aerogenerador* en espagnol si nous traduisons les composants *aéro* et *générateur* individuellement. Nous menons des expériences avec deux paires de langues et nous obtenons une haute précision

pour les traductions des composés savants par une approche compositionnelle.

## 3.1 Introduction

Un nouveau terme (appelé néologisme) peut être formé par des processus de formation de mots (ex. dérivation, abréviation, composition, etc.) qui sont propres à chaque langue. Dans les langues romanes et germaniques (entre autres), un de ces processus de formation des mots est la formation savante (également dite *néoclassique*). La formation savante consiste à combiner des éléments empruntés au grec ou au latin afin de créer des termes nommés *composés savants*. Nous appelons les éléments qui constituent les composés savants les racines gréco-latines (également dits *éléments néoclassiques*). Par exemple, la combinaison des racines gréco-latines FR *hydro* et FR *logie* conduit au composé savant FR *hydrologie*. Une langue peut emprunter de nouvelles racines au grec ou au latin lorsque cela est nécessaire pour former de nouveaux composés savants.

Dans les corpus spécialisés, les composés savants constituent une partie non négligeable du vocabulaire de ces corpus, notamment dans les corpus issus du domaine médical. Par conséquent, la traduction automatique des composés savants d'une langue source vers une langue cible peut aider à enrichir les lexiques bilingues spécialisés.

Nous proposons, dans ce chapitre, d'identifier les composés savants dans un corpus comparable et de trouver leurs équivalents qui sont aussi des composés savants.

## 3.2 Racines gréco-latines

Les racines gréco-latines sont des éléments empruntés aux langues grecque et latine (ex. patho-, bio-, -logie, etc). Ces éléments ne sont pas considérés comme des unités lexicales car ils ne peuvent pas jouer le rôle de mots autonomes dans la syntaxe d'une langue, c'est-à-dire qu'ils sont toujours vus comme des formes combinées à d'autres éléments (ex. *bio-* dans le mot *biologie*) (Amiot et Dal 2008). Chaque langue peut assimiler ses emprunts du grec ou du latin phonologiquement. En d'autres termes, un mot grec ou latin subit un minimum d'adaptation avant d'être adopté par une langue hôte. Par exemple, les racines FR *pathie* et EN *pathy* ont été empruntées au mot grec *pathos*.

En outre, chaque racine gréco-latine peut avoir des allomorphes différents, ce qui signifie qu'un élément emprunté au grec ou au latin peut être assimilé à des formes différentes dans une seule langue. Par exemple, la racine *neuro* en anglais peut prendre deux formes en français : *neuro* comme dans FR *neurologie* et *névro* comme dans FR *névrodermite*. Il peut exister aussi des relations sémantiques entre les racines (ex. synonymie entre les racines EN *opt* et EN *ophtalm* car les deux signifient *vision*, hyponymie entre les racines EN *blast* et EN *cyt* car la deuxième signifie *cellule* et la première est un type du mot *cellule*, etc.), nous n'exploitons pas ces types de relations dans le cadre de la traduction de composés savants élaborée dans ce travail.

Les racines gréco-latines peuvent apparaître à des positions différentes dans les composés savants. Bauer (1983) distingue les formes de combinaison initiales et finales :

- racines en position initiale (ICF) [1] : les ICFs comprennent des formes de racines gréco-latines qui apparaissent en position initiale (ex. bio-, cardio-, patho-, etc.). Les ICFs se terminent souvent par les lettres de liaison *o* ou *i*.
- racines en position finale (FCF) [2] : les FCFs comprennent des racines gréco-latines qui apparaissent en position finale (ex. -logie, -cide, -pathie, etc.).

Plusieurs ICFs peuvent apparaître successivement dans un composé savant (ex. histo- et patho- dans histopathologie). Un mot emprunté au grec ou au latin peut être adapté à la fois dans les deux formes : ICF et FCF. Par exemple, *patho-* (ICF) dans *pathologie* et *-pathie* (FCF) dans *cardiopathie*, les deux éléments étant adaptés de *pathos*.

Nous distinguons les racines gréco-latines des affixes (ex. préfixes et suffixes) dans une langue. Les affixes peuvent apparaître soit en position initiale (préfixes, ex. FR pré-) soit en position finale (suffixes, ex. FR -ique). Cependant, certaines racines prennent toujours une position initiale comme EN *micro-*, d'autres prennent toujours une position finale comme EN *-ectomy*. Il n'est, en effet, pas toujours facile de distinguer les racines gréco-latines des préfixes ou des suffixes. Nous supposons dans cette thèse qu'une racine gréco-latine est de longueur supérieure à 2 (en nombre de lettres) et qu'elle a des équivalents similaires orthographiquement dans plusieurs langues européennes.

## 3.3 Formes des composés savants

Nous définissons un composé savant comme étant un terme simple qui comprend au moins une racine gréco-latine combinée avec d'autres racines gréco-latines ou un mot. Selon cette définition, les mots *aérogénérateur* et *cardiologie* sont des composés savants, alors que *hépatique* n'est pas un composé savant parce qu'il est de la forme [racine + suffixe].

La formation néoclassique des mots dans différentes langues suit généralement le modèle des langues grecque et latine pour former des termes (Amiot et Dal 2008). Le modèle gréco-latin pour un terme $XY$ qui se compose de deux éléments $X$ et $Y$ est le suivant : [déterminé + déterminant]. Selon ce modèle, *cardiologie* se compose de *logie* (étude) étant le déterminé (identifie la classe dont le composé savant est une sorte) et *cardio* (cœur) étant le déterminant (donne le trait distinctif).

Un composant d'un composé savant peut être une racine, un mot, un préfixe ou un suffixe. (Namer 2009, p. 333) cite quelques formes possibles d'un composé savant en considérant qu'il peut être décomposé en deux éléments. Quelques formes possibles selon (Namer 2009, p. 333) sont les suivantes :

1. Initial Combining Form.
2. Final Combining Form.

1. ICF + FCF (ex. anthropophage, ICF=anthropo & FCF=phage).
2. Mot + FCF (ex. liberticide, Mot=liberté & FCF=cide). Le mot *liberté* a été tronqué et combiné avec la racine *cide* à l'aide de la lettre de liaison *i*.
3. ICF + Mot (ex. hydralcool, ICF=hydr & FCF=alcool).
4. Composé savant + FCF. (ex. rhinopharyngite, ICF=rhinopharynx & FCF=ite).
5. ICF + Composé savant (ex. biotechnologie, ICF=bio & Composé savant=technologie).

D'autres formes de composés savants existent. Par exemple, un composé savant peut être composé d'un préfixe et de racines gréco-latines (ex. *colohyperplasie*, où *hyper* est un préfixe). Un composé savant peut aussi être préfixé (ex. *antiandrogène*, où *anti* est un préfixe).

## 3.4 Difficulté de traduction

Les composés savants peuvent être des termes très spécifiques aux domaines. La productivité de ces termes dans les domaines spécialisés rend leur traduction difficile, car beaucoup d'entre eux ne sont pas susceptibles d'être trouvés dans les dictionnaires bilingues et ont des fréquences faibles. De plus, les racines gréco-latines qui forment les composés savants ne se trouvent généralement pas dans les dictionnaires bilingues.

Les composés savants partagent certains problèmes de traduction avec les termes complexes que nous avons présentés dans le chapitre 2 (section 2.2). Nous présentons certains problèmes qui concernent la traduction des composés savants ci-dessous :

- **fertilité** : un composé savant source peut être traduit par un terme complexe. Par exemple, le composé savant FR *télécommande* peut être traduit par EN *remote control*.
- **non-compositionnalité** : un composé savant se traduit non-compositionnellement quand on ne peut pas obtenir son équivalent dans une autre langue en traduisant ses composants individuellement. Par exemple, le composé savant FR *cytoponction* se traduit par EN *cytology aspiration*, où FR *ponction* n'est pas la traduction directe de EN *aspiration*, et *cytology* n'est pas la traduction de la racine *cyto*.
- **variation de structure inter-langue** : un composé savant d'une structure spécifique dans une langue peut être traduit par un terme d'une structure différente dans une autre langue. Par exemple, le composé savant FR *mastectomie* (de la forme [ICF + FCF]) peut être traduit par EN *breast removal* (de la forme [Mot + Mot]).
- **variation syntaxique intra-langue** : un composé savant peut apparaître sous différentes structures dans le même texte. Par exemple, FR *lumpectomie* et sa variante FR *tumorectomie* peuvent être traduits en anglais par *lumpectomy*.

## 3.5 Travaux connexes

Nous avons présenté dans le chapitre 2 des approches compositionnelles, telles que les approches de Robitaille et al. (2006), Baldwin et Tanaka (2004) et Grefenstette (1999), proposées pour traduire des termes complexes en s'appuyant sur une propriété compositionnelle de ces termes. Ces approches consistent à traduire un terme complexe en cherchant les traductions de chacun des mots qui le composent individuellement à l'aide d'un dictionnaire bilingue. Ensuite, les traductions individuelles sont combinées de manière à générer toutes les combinaisons possibles ou selon des modèles appropriés à chaque paire de langues, afin de produire des traductions candidates du terme complexe. Comme les termes complexes, beaucoup de composés savants possèdent une propriété compositionnelle. Contrairement aux mots qui composent les termes complexes, les équivalents de racines gréco-latines qui forment les composés savants ne sont pas susceptibles d'être fournis par les dictionnaires bilingues.

Nous présentons dans cette section des travaux qui dépendent d'autres ressources que les dictionnaires bilingues pour traiter les composés savants. Ces travaux concernent principalement l'analyse morphologique ou la traduction des composés savants.

### 3.5.1 Utilisation de langue pivot

L'approche présentée par Claveau et Kijak (2010) essaye d'acquérir automatiquement la sémantique des racines gréco-latines. Pour cela, l'analyse morphologique des composés savants est faite en s'appuyant sur une langue pivot : le japonais. Plus précisément, l'approche se penche sur les termes écrits avec l'alphabet kanjis. Claveau et Kijak (2010) font l'hypothèse que les composés savants sont traduits en japonais par des termes qui font référence à des mots simples. Un exemple est donné dans la figure 3.1.

photochimiothérapie = 光化学療法
photo (光=lumière)
chimio (化学=chimie)
thérapie (療法=thérapie)

Figure 3.1 – *Exemple d'un composé savant où chacun de ses composants est aligné avec un mot simple en japonais (Claveau et Kijak 2010)*

L'avantage de l'utilisation de termes écrits en kanjis alignés avec des composés savants, selon les auteurs, est que la morphologie de tels termes est une simple concaténation de mots faciles à traduire en utilisant un dictionnaire général. Leur approche consiste donc à aligner automatiquement chaque racine gréco-latine avec son équivalent qui sera un mot simple en japonais. Les composants d'un composé savant sont appelés *morphes*.

Étant donnée une liste de termes français alignés avec leurs équivalents japonais, Claveau et Kijak (2010) développent un algorithme qui fournit

les morphèmes des termes français et leurs sens en kanjis. Cet algorithme d'alignement est basé sur un algorithme Expectation-Maximization (EM) de type Baum-Welch étendu (alignement many-to-many) (Jiampojamarn et al. 2007). L'algorithme est capable d'aligner un symbole (lettre ou caractère vide) de la langue source avec un symbole de la langue cible. La phase *Expectation* calcule les comptes partiels de toutes les correspondances possibles entre les sous-séquences de kanjis et de lettres. La phase *Maximization* estime les probabilités d'alignement entre ces sous-séquences de kanjis et de lettres.

Le même kanji peut être aligné avec plusieurs morphes d'un même morphème mais avec des probabilités différentes. Par exemple, pour le kanji (translittéré) *Saikin*, il peut être associé à trois morphes : *bactérie*, *bactério* (comme dans bactériolyse) et *bactéri* (comme dans mycobactériose), chacun avec une certaine probabilité. Cette dispersion des probabilités est due à la variation morphologique : bactério, bactérie et bactéri étant trois morphes d'un même morphème. Afin de les regrouper et de renforcer leurs probabilités dans l'étape *Maximization*, les alignements kanjis-morphèmes sont normalisés en s'appuyant sur le calcul de l'analogie. Une analogie est une relation entre quatre éléments qui peut être notée par « a : b : : c : d » et qui veut dire « a est à b ce que c est à d ». Une telle analogie sera par exemple dermato : dermo : : hémato : hémo. Sachant que *dermato* et *dermo* appartiennent à un même morphème, on peut en déduire que *hémato* et *hémo* appartiennent aussi au même morphème. La mise en œuvre de cette analogie consiste à apprendre une règle de réécriture de préfixe et de suffixe permettant de passer de *dermato* à *dermo* et de vérifier que cette règle s'applique bien à *hémato* et *hémo* (ex. ato $\rightarrow$ o).

Afin de pouvoir utiliser la normalisation par analogie, une liste de base de morphes alignés (comme *dermato* avec *dermo*) doit être disponible. Une telle liste est donc extraite automatiquement à l'aide d'une similarité de lettres (le préfixe le plus long entre deux chaînes). Les morphes qui partagent une similarité supérieure à un certain seuil seront considérés comme liés. Cette liste sera ensuite utilisée par la normalisation par analogie afin de regrouper d'autres morphes liés.

Les expériences sont faites sur 8 000 paires de termes français à aligner et leurs traductions en kanjis extraits du méthathesaurus de l'UMLS (Tuttle et al. 1990). Dans le but d'évaluer les résultats, 1 600 paires sont alignées manuellement (morphème-kanji) et une précision supérieure à 70 % est obtenue avec cette liste. Cependant, l'hypothèse que cet article fait n'est pas toujours vérifiée parce que certains composés savants ne se traduisent pas de manière compositionnelle en kanjis. Les analyses des résultats montrent aussi que certains termes ont été décomposés alors qu'ils ne sont pas des composés savants.

Dans le but de traduire un terme français inconnu, les traductions des morphes en kanji avec leurs probabilités sont exploitées dans un algorithme de type *Viterbi* (Ryan et Nudd 1993) afin de décomposer un terme en morphes, les traductions de ces morphes seront utilisées ensuite pour traduire le terme. Un test sur 128 termes montre que 58 termes sont bien traduits, 36 termes ne sont pas traduits et 34 termes sont mal traduits.

### 3.5.2 Utilisation de règles de réécriture

Une approche proposée dans Claveau (2009) a pour objectif de traduire des termes biomédicaux d'une langue vers une autre. L'approche dépend de règles de réécritures (au niveau des lettres) trouvées à l'aide d'un algorithme d'apprentissage entraîné sur une liste de termes alignés entre deux langues. Le processus de traduction à l'aide de règles de réécriture ressemble à celui de la traduction automatique statistique. L'article suppose qu'il existe une large classe de termes liés morphologiquement d'une langue à une autre et que les différences entre ces termes sont assez régulières pour être apprises automatiquement. Ces hypothèses sont basées sur le fait que les termes biomédicaux dans différentes langues partagent souvent des racines gréco-latines (ex. FR ophtalmorragie traduit par EN ophthalmorrhagia).

Les paires de traductions utilisées pour l'apprentissage sont alignées au niveau des lettres en utilisant l'outil DPalign [3] qui minimise la distance d'édition [4] entre deux chaînes. Des règles sont d'abord inférées à partir de l'alignement produit par l'outil DPalign. Par exemple, s'il suffit de remplacer *e* par *a* pour transformer FR *ophtalmorragie* en EN *ophthalmorrhagia*, la première règle trouvée sera e→a (une règle pour chaque différence). Ensuite, des règles plus spécifiques sont générées en ajoutant, à cette règle, les lettres qui précèdent et qui suivent les lettres substituées : ie→a, ie→ia, gie→gia [5], etc. Finalement, des scores sont attribués à chaque règle, le score d'une règle est le nombre de fois que la règle peut être appliquée, divisé par le nombre de fois où la prémisse de la règle correspond à une suite de lettres dans un terme source.

Cet ensemble de règles est utilisé pour traduire un terme (plusieurs traductions possibles peuvent être générées pour un terme). Afin de choisir la traduction la plus probable, un modèle de langage de n-grammes (n=7) [6] est utilisé. Pour un mot qui se compose de $k$ lettres : $l_1$, $l_2$,.., $l_k$, sa probabilité ($P$) est calculée comme suit :

$$P(m) = \prod_{i=1}^{k} P(l_i | l_{i-n+1}, .., l_{i-1}) \tag{3.1}$$

où $l_i$ est la lettre à la position $i$ de l'ensemble des lettres du mot $m$. L'utilisation d'un modèle de langage permet de trouver la traduction la plus probable car le modèle du langage favorise les traductions qui « ressemblent » à des mots corrects dans la langue cible. Ce modèle est entraîné sur les mots cibles dans la liste des paires de termes alignés (déjà disponible).

3. http ://www.cnts.ua.ac.be/ decadt/ ?section=dpalign

4. Les coûts choisis par DPalign sont appris sur l'ensemble des mots, l'alignement avec un caractère vide est possible

5. Toutes les règles de réécritures possibles qui sont de la forme *i* → *e* sont générées dans un treillis, où * signifie 0 lettres ou plus.

6. La probabilité d'observer la *i*ème lettre dans l'historique de contexte des précédentes n-1 lettres.

Les expériences ont été réalisées sur plusieurs paires de langues. Pour la paire FR-EN, la précision de traduction sur 1 000 termes[7] est de 85,8 %. Les expériences montrent que pour inférer des règles de réécriture, un nombre limité de termes alignés pour l'apprentissage peut donner une bonne précision mais un nombre limité de règles inférées. Pour d'autres paires de langues, qui sont un peu plus éloignées, la précision était inférieure à celles obtenues sur des langues plus proches. Par exemple, pour la paire de langues espagnol-anglais, la précision était de 71,7 %. Toutefois, pour la paire de langue anglais-allemand (langues germaniques), la précision de traduction était de 68,8 %.

### 3.5.3 Utilisation de règles de préfixation

Cartoni (2009) travaille sur les néologismes italiens et français construits par préfixation. Il s'appuie sur des ressources lexicales (des lexiques monolingues, des lexiques bilingues et des corpus monolingues) ainsi que sur un ensemble de règles de formation de lexèmes bilingues afin de détecter des néologismes construits et de générer leurs traductions. Ces traductions sont trouvées à l'aide des règles de génération appelées « règles de transfert ».

Cartoni (2009) traite les néologismes qui sont de la forme [préfixe + base]. Par exemple, pour traduire le néologisme italien *retrobottega* : ce néologisme est d'abord identifié (à l'aide des listes de préfixes et de suffixes) comme composé du préfixe *retro* et du nom non-déverbal[8] *bottega*. Ensuite, les règles du transfert associées au préfixe *retro* seront appliquées afin de générer la traduction du néologisme. Ces règles attribuent deux traductions possibles à *retro* en français (*arrière* ou *rétro*) en fonction de la nature du mot de base préfixé par *retro*. Par exemple, puisque la base *bottega* a été identifiée comme un nom non-deverbal[9] : *retro* sera traduit par *arrière* et *bottega* sera traduit par un nom (ex. *boutique*). La traduction construite sera donc *arrière-boutique* selon la règle du transfert du préfixe *retro* avec une base nominale non-déverbale. Si la base préfixée par *retro* était déverbale, *retro* aurait été traduit par *rétro* en français.

Pour les expériences, Cartoni (2009) utilise un corpus italien (composé d'articles de journaux) et essaye d'analyser chaque mot existant dans le corpus non présent dans le lexique monolingue. Il utilise le Web comme corpus pour la langue française afin de chercher les traductions générées. Une traduction générée est considérée comme fiable si elle apparaît plus de 5 fois sur le Web. L'évaluation montre que certains préfixes d'une longueur égale à 1 ou 2 (ex. a-, di-, s-) introduisent beaucoup de bruit dans le système, et que l'identification correcte des formes des néologismes est l'étape la plus importante pour avoir une bonne qualité de traduction des néologismes. En outre, la qualité de traduction varie selon le préfixe en

7. Les termes alignés utilisés pour l'entraînement et le test ont été extraits du méthathesaurus de l'UMLS (Tuttle et al. 1990).

8. Nom qui n'est pas dérivé d'un verbe.

9. En vérifiant si le suffixe du mot n'est pas l'un des suffixes déverbaux.

question. Par exemple, pour le préfixe italien *super* la précision est de 42 % alors qu'elle est de 93,9 % pour le préfixe *de*.

### 3.5.4 Utilisation de listes de racines gréco-latines

Certains travaux ont exploité des listes de racines gréco-latines monolingues ou bilingues prédéfinies afin d'analyser ou de traduire des composés savants dans plusieurs langues.

#### Analyse morpho-sémantique des composés savants

Namer et Baud (2007) présentent un système capable d'associer les composés savants à des mots reliés morphologiquement et de leur donner des définitions. Le travail exploite le fait que les racines gréco-latines sont communes dans la plupart des langues occidentales et donc une méthode indépendante de la langue est développée. Cette méthode trouve des relations entre des composés savants au niveau monolingue. Afin d'appliquer une analyse morphosyntaxique (analyse morphologique et interprétation sémantique) en donnant des définitions aux composés savants à partir de leurs composants, Namer et Baud (2007) construisent des listes de racines gréco-latines avec leurs significations et les relations entre elles. Une telle liste est définie pour plusieurs langues. Ensuite, une dérivation de relations lexicales entre les composés savants est menée en utilisant des règles de calcul et en exploitant les listes de racines.

Le système développé suppose que les racines gréco-latines peuvent être reliées par quatre liens :

1. Synonymie : représenté par $=$. Par exemple, EN OPT=OPHTALM, vision. C'est-à-dire que la racine EN *opt* est un synonyme de la racine EN *ophtalm* et que les deux signifient *vision*.
2. Hyponymie : représenté par $<$. Par exemple, EN BLAST, embryonal cell < CYT, cell.
3. Méronymie : représenté par $\leftarrow$. Par exemple, EN CORO, pupil $\leftarrow$ OCUI, eye.
4. Proximité sémantique : representé par $\approx$. Par exemple : EN DISC, intervertebral disk $\approx$ SPONDYL, verterbra.

Après avoir défini les relations entre les racines, un ensemble de règles est défini pour relier deux composés savants $A$ et $B$, où $A$ est de la forme $[Y_1 + X_1]$ [10] et $B$ est de la forme $[Y_2 + X_2]$. Un composé savant peut être décomposé en deux éléments : une racine + un mot, racine + racine ou un mot [11] + racine. Les liens déjà établis entre les racines peuvent établir des liens entre des composés savants. Par exemple, supposons que $A$ et $B$ partagent le même composant $X$ (ex. $A$=abdominoscopy et $B$=laparoscopy, où X=scopy). Si les racines *abdo* et *laparo* sont liées par le lien de synonymie, $A$ et $B$ seront considérés également comme des synonymes.

10. $Y$ et $X$ sont des composants.
11. Le mot peut être tronqué.

Les expériences sont menées sur la langue française afin de valider la méthode. Les expériences sont donc effectuées au niveau monolingue. Il est montré que le lien de synonymie est le plus fort et que les définitions correctes sont trouvées pour plus de 77,3 % d'une liste de 100 composés savants inconnus.

**Traductions des composés savants à partir d'un corpus comparable**

Un travail récent qui concerne la traduction des composés savants – réalisé en parallèle avec le présent travail – est celui de Delpech et al. (2012). Ce travail propose une méthode permettant d'extraire des traductions de termes morphologiquement construits (préfixés, composés savants ou natifs [12], suffixés ou des combinaisons de ces processus) à partir de corpus comparables.

La méthode se base sur la traduction compositionnelle en exploitant les traductions des composants des termes. Ces composants peuvent donc être des préfixes, des suffixes, des racines ou des mots. Les auteurs travaillent aussi sur la traduction de termes simples (morphologiquement composés) en termes complexes. Pour traduire un terme, la méthode se compose principalement en trois étapes : (a) **décomposition** : le terme est décomposé en morphèmes (ex. EN post-menopause est décomposé en *post* et *menopause*) en se référant à des listes de préfixes, suffixes, racines et mots ; (b) **traduction** : les morphèmes composants sont traduits en morphèmes ou en mots (ex. EN *post* peut être traduit en français par le morphème *post* ou par le mot *après*) et les mots sont traduits par des mots (ex. EN *menopause* peut être traduit par FR *ménopause*) ; (c) **recomposition** : les composants traduits individuellement sont combinés afin de produire les traductions candidates (ex. postménopause, après la ménopause, etc.) [13]. Toutes les permutations possibles des traductions des composants sont générées. Les composants de chaque permutation possible sont combinés selon plusieurs formes (avec des espaces, sans espaces, etc.) afin de produire des traductions candidates. Les traductions candidates sont ensuite filtrées à partir du corpus cible (celles qui ne se trouvent pas dans le corpus sont ignorées).

Plusieurs traductions possibles peuvent donc être proposées pour un terme. Afin de ré-ordonner ces traductions, plusieurs traits sont exploités : (a) la fréquence de la traduction dans le corpus cible ; (b) la probabilité que le terme source d'une partie du discours $x$ soit traduit par un terme cible d'une partie du discours $y$ ; (c) la similarité entre les vecteurs de contexte [14] du terme source et du terme cible ; et (d) la fiabilité des traductions individuelles trouvées. Le deuxième et le dernier traits sont calculés préalablement à partir de données annotées. Tous les traits sont combinés par une interpolation linéaire ou par un algorithme d'apprentissage du type *Learning to Rank* (Liu 2011).

12. Termes simples qui se composent de deux ou plusieurs mots.
13. Si un des composants n'a pas été identifié, la traduction échoue.
14. La notion du vecteur de contexte est introduite dans le chapitre 2 en section 2.1.2.

Delpech et al. (2012) utilisent plusieurs types de données pour trouver les équivalents des composants d'un terme : (a) listes alignées de racines gréco-latines, préfixes et suffixes ; (b) dictionnaires bilingues ; (d) liste de cognats extraits automatiquement du corpus comparable.

Les expériences sont faites avec des corpus comparables pour les paires de langue EN-FR et EN-DE dans le domaine du cancer du sein (chaque corpus est d'une taille $\approx$ 0.4 million de mots). Les résultats sur une liste de références de 126 mots, montrent que la précision au top 1 pour le couple EN-FR peut atteindre 93 %. La combinaison des traits par une interpolation linéaire ou par un algorithme d'apprentissage peut donner les mêmes résultats. Cela est peut-être dû au fait que les algorithmes *Learning to Rank* ont été proposés à la base dans le domaine de la recherche d'information pour résoudre la difficulté de la modélisation manuelle de la combinaison de nombreux traits, mais le travail de Delpech et al. (2012) combine peu de traits pour ré-ordonner les traductions.

### 3.5.5 Discussion

Nous avons présenté dans cette section cinq méthodes de l'état de l'art qui peuvent traiter des composés savants. L'approche que nous proposons dans ce chapitre exploite des listes de racines gréco-latines alignées entre les langues comme le travail de Namer et Baud (2007). D'ailleurs, notre approche emploie, comme Cartoni (2009) et Delpech et al. (2012), une méthode compositionnelle afin de traduire les composés savants.

Cartoni (2009) traite des néologismes qui peuvent être des composés savants ou non, car il ne fait pas de distinction entre les préfixes natifs et les racines gréco-latines. Dans notre approche, nous n'exploitons pas les catégories grammaticales des composants d'un terme comme dans Cartoni (2009).

Contrairement au travail de Delpech et al. (2012) qui traite plusieurs types de termes morphologiquement construits, notre travail concerne seulement l'extraction et l'alignement automatique des composés savants. Nous nous intéressons aux propriétés des composés savants et nous supposons qu'une partie importante des composés savants dans une langue source peut être traduite de manière compositionnelle par des composés savants dans une langue cible. Delpech et al. (2012) génèrent les traductions d'un composé savant en s'appuyant sur toutes les permutations possibles de traductions de ses composants. La méthode compositionnelle que nous employons se base sur le modèle gréco-latin suivi généralement lors de la construction des composés savants par différentes langues pour générer des traductions candidates.

En apprenant des règles de réécriture entre les termes biomédicaux dans plusieurs langues, Claveau (2009) suppose que les composés savants entre deux langues sont morphologiquement liés. C'est-à-dire qu'il suppose implicitement que l'ordre des composants entre un composé savant source et un autre cible est le même. Le modèle compositionnel sur lequel

nous nous appuyons exige également que l'ordre des composants entre deux langues soit respecté lors de la traduction.

## 3.6 Contribution à la traduction des composés savants

Dans cette section, nous décrivons les hypothèses que nous faisons pour traduire les composés savants. Nous présentons également les formes que nous traitons. Enfin, nous expliquons l'approche d'alignement des composés savants dans une langue avec leurs équivalents dans une autre langue.

### 3.6.1 Hypothèses

Notre approche de traduction des composés savants entre deux langues est basée sur les hypothèses suivantes :

**Propriété compositionnelle ou semi-compositionnelle**

Un composé savant d'une langue source peut être traduit de manière compositionnelle en un composé savant d'une langue cible. Chaque composant du composé savant de la langue source est traduit individuellement, puis la traduction finale du composé savant est la combinaison de ses composants traduits.

Le sens d'un composé savant peut être parfois obtenu compositionnellement (Estopa et al. 2000). Par exemple, la traduction du composé savant EN *hydrology* (ayant une propriété compositionnelle) en français est *hydrologie*. Cette traduction peut être obtenue par la combinaison des traductions des composants du mot *hydrology* : EN *hydro* (eau) = FR *hydro* et EN *logy* (étude) = FR *logie*.

En effet, la traduction compositionnelle d'un composé savant vers un autre composé savant peut donner des résultats précis, même dans le cas où le sens d'un composé savant ne peut pas être restauré compositionnellement. McCray et al. (1988) affirment que le sens de la plupart des composés savants ne peut pas être restauré à partir de leurs composants. Nous croyons que malgré cela, la traduction compositionnelle des composés savants peut aboutir même quand le sens d'un composé savant ne peut être obtenu à partir de ses composants. Prenons à titre d'exemple le composé savant EN *leukopathy* ('une maladie entraînant la perte de pigmentation de la mélanine de la peau'). Ce composé savant n'est pas déterminé par ses éléments puisque sa définition ne contient pas une référence explicite à *blanc* qui est le sens de son premier composant EN *leuko*. Toutefois, les équivalents de *leukopathy* dans d'autres langues sont les suivants : FR *leucopathie*, DE *leukopathie*, ES *leucopatia*. Cet exemple montre que les composés savants peuvent être traduits de manière compositionnelle, même lorsque leur sens ne peut pas être restauré de manière compositionnelle.

Comme les termes complexes, certains composés savants ne peuvent pas être traduits compositionnellement. Cependant, nous suppo-

sons que certains de ces composés savants peuvent être traduits de manière semi-compositionnelle d'une langue à une autre. Par exemple, la traduction du composé savant FR *multivarié* peut être EN *multivariate* [15]. Le premier composant FR *multi* dans FR *multivarié* peut être traduit par EN *multi* qui est aussi le premier composant du mot EN *multivariate*. Cependant, FR *varié* ne peut pas être traduit par EN *variate* en utilisant un dictionnaire bilingue [16].

**Conservation de l'ordre des composants**

Nous supposons que l'ordre des éléments d'un composé savant source est conservé dans son équivalent de composé savant cible. Prenons le composé savant FR *hydrologie* à titre d'exemple, l'équivalent de *hydro* doit apparaître avant l'équivalent de *logie* lorsque l'on combine les équivalents des composants pour former la traduction finale. Cette hypothèse est fondée sur le fait que la formation néoclassique des mots dans différentes langues suit le modèle des langues grecque et latine pour former des termes.

Selon cette hypothèse, l'ordre des éléments doit être respecté lors de la traduction d'un composé savant. Par conséquent, chaque racine gréco-latine composante est traduite par une racine gréco-latine du même type (par exemple un ICF par un ICF, un FCF par un FCF).

Ci-après, nous nous intéressons à la traduction des composés savants qui sont des adjectifs ou des substantifs, malgré le fait que certains verbes ou adverbes peuvent contenir des racines gréco-latines (ex. le verbe EN *hydrogenate* ou l'adverbe FR *histologiquement*).

Nous n'ignorons pas le fait que les hypothèses que nous faisons ne sont pas valides pour la traduction de tous les composés savants d'une langue à une autre. Cependant, nous supposons qu'une partie importante des composés savants peut être traduite par des composés savants dans une autre langue à l'aide d'une méthode compositionnelle ou semi-compositionnelle, notamment dans les domaines médicaux comme les travaux de l'état de l'art le montrent. De plus, même si un composé savant peut être traduit par un terme qui n'est pas un composé savant dans une autre langue, une traduction en un composé savant peut exister : par exemple, le composé savant FR *mastectomie* peut être traduit par *mastectomy* et *breast removal* en anglais. La première traduction (un composé savant) a plus tendance à être utilisée dans les textes scientifiques.

### 3.6.2 Formes traitées

Les composés savants peuvent prendre des formes différentes. Nous avons présenté dans la section 3.3 des formes possibles de composés savants. Les composés savants contiennent au moins une racine gréco-latine mais peuvent aussi contenir des mots natifs et des préfixes combinés dans

15. *multivarié* est aligné avec *multivariate* sur http ://www.linguee.fr/francais-anglais/.

16. Nous avons vérifié la traduction de *varié* du français vers l'anglais sur http ://www.larousse.fr/dictionnaires/francais-anglais/
et http ://dictionnaire.reverso.net/francais-anglais/.

| Langue | Exemples |
|---|---|
| EN | radiology (radio$_{ICF}$ & logy$_{FCF}$), histogram (histo$_{ICF}$ & gram$_{FCF}$) |
| FR | histopathologie (histo$_{ICF}$ & patho$_{ICF}$ & logie$_{FCF}$), monomorphe (mono$_{ICF}$ & morphe$_{FCF}$) |
| DE | radiometrie (radio$_{ICF}$ & metrie$_{FCF}$), biotechnologie (bio$_{ICF}$ & techno$_{ICF}$ & logie$_{FCF}$) |
| ES | geomorfologia (geo$_{ICF}$ & morfo$_{ICF}$ & logia$_{FCF}$), histopatología (histo$_{ICF}$ & pato$_{ICF}$ & logía$_{FCF}$) |

TABLE 3.1 – *Exemples des composés savants de la forme [ICF$_+$ + FCF]*

| Langue | Exemples |
|---|---|
| EN | photobioreactor (photo$_{ICF}$ & bio$_{ICF}$ & reactor$_{Mot}$), biomedical (bio$_{ICF}$ & medical$_{Mot}$) |
| FR | cardiovasculaire (cardio$_{ICF}$ & vasculaire$_{Mot}$), photosensibilisateur (photo$_{ICF}$ & sensibilisateur$_{Mot}$) |
| DE | ferroelektrisch (ferro$_{ICF}$ & elektrisch$_{Mot}$), kardiovaskulär (kardio$_{ICF}$ & vaskulär$_{Mot}$) |
| ES | multidisciplinario (multi$_{ICF}$ & disciplinario$_{Mot}$), fotosensibilizador (foto$_{ICF}$ & sensibilizador$_{Mot}$) |

TABLE 3.2 – *Exemples des composés savants de la forme [ICF$_+$ + Mot]*

des ordres différents. Nous choisissons de représenter ci-dessous de manière simplifiée certaines formes de composés savants. Notre méthode ne peut aligner que les composés savants qui appartiennent à l'une des formes suivantes :

- **[ICF$_+$ + FCF]** [17]
  Cette forme comprend des composés savants qui se composent uniquement des racines gréco-latines. Un ou plusieurs ICFs peuvent apparaître successivement avec un seul FCF. Des exemples de cette forme des composés savants sont donnés dans la table 3.1.
- **[ICF$_+$ + Mot]**
  Cette forme comprend un ou plusieurs ICFs combinés avec un mot. Des exemples de cette forme de composés savants sont fournis dans la table 3.2.

### 3.6.3 Traduction compositionnelle des composés savants

La méthode que nous proposons identifie tout d'abord des composés savants candidats des langues source et cible à partir de corpus comparables en utilisant deux listes de racines gréco-latines, la première pour la langue source ($l_{NE_s}$) et la deuxième pour la langue cible ($l_{NE_c}$). Cela produit des listes de candidats cibles et sources de composés savants : $l_{NC_s}$ et $l_{NC_c}$. Ensuite, chaque composé savant ($NC_s$) dans $l_{NC_s}$ est aligné avec son (ses) équivalent(s) dans $l_{NC_c}$, à l'aide d'un dictionnaire bilingue ($Dico_{Bi}$) et d'une liste de racines gréco-latines alignées ($NE_A$). Notre approche suit les principales étapes des méthodes compositionnelles pour les termes

17. ICF$_+$ signifie un ou plusieurs ICFs.

complexes (présentées en section 2.2 du chapitre 2) : (a) l'identification de composés savants candidats ; (b) la génération de traductions candidates ; et (c) la sélection de traductions correctes. Les étapes principales sont illustrées en figure 3.2.

Cela produit donc une liste de composés savants alignés entre deux langues. Ensuite, une méthode semi-compositionnelle essaye de traduire les composés savants candidats (de la langue source) qui n'ont pas été traduits par l'approche compositionnelle et qui n'existent pas dans le dictionnaire bilingue. La méthode semi-compositionnelle aligne les composés savants avec leurs équivalents à l'aide d'une mesure de similarité de lettres ainsi qu'à l'aide d'une approche distributionnelle.

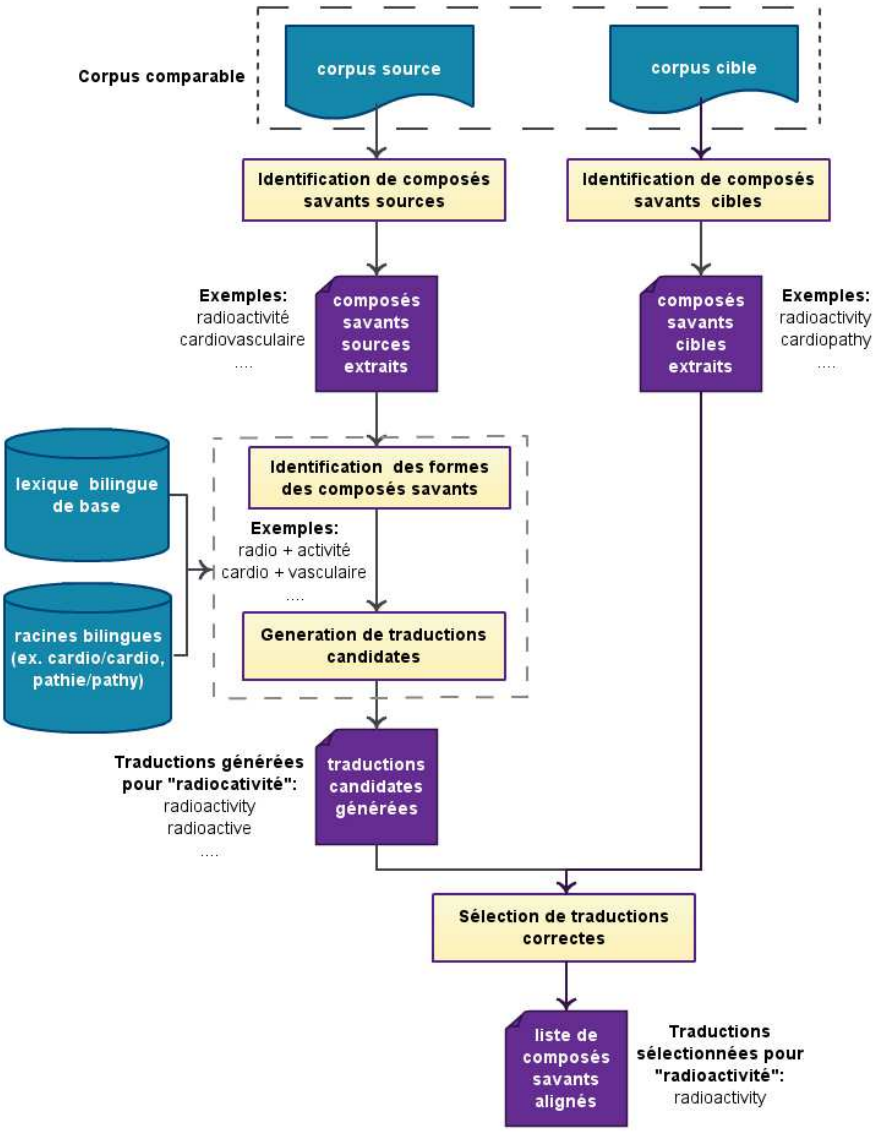


FIGURE 3.2 – *Représentation des trois étapes principales de notre méthode compositionnelle pour l'extraction d'un lexique bilingue de composés savants*

**Identification des composés savants candidats dans le corpus**

Les listes candidates de composés savants sources et cibles ($l_{NC_s}$ et $l_{NC_c}$) sont obtenues en projetant la liste de racine gréco-latines sources $l_{NE_s}$ sur le corpus de langue source et la liste de racine gréco-latines cibles $l_{NE_c}$ sur le corpus de langue cible. Ces listes de composés savants candidats contiennent des mots comportant des composants qui appartiennent aux listes de racines gréco-latines (ex. si la liste de racines contient la racine *techno*, cette racine sera détectée dans un mot comme *biotechologie*).

En effet, les adjectifs ou les substantifs qui ont au moins une racine gréco-latine potentielle (ICF ou FCF) seront considérés comme des composés savants candidats. Un ICF peut apparaître au début ou n'importe où au milieu d'un composé savant, par exemple, les ICFs *bio-*, *géo-* et *morpho-* apparaissent dans *biogeomorphological*. Un FCF se trouve à la fin d'un composé savant, tel que *-pathie* dans *neuropathie* et *-logie* dans *biotechnologie*.

**Identification de la forme d'un composé savant**

La projection faite dans l'étape d'identification des composés savants décompose chaque composé savant candidat en deux ou plusieurs composants, où au moins un de ces composants est une racine gréco-latine potentielle.

La forme d'un composé savant candidat doit être identifiée (voir les formes traitées en section 3.6.2) avant de procéder à la génération des traductions possibles de ce composé. Tous les composants d'un composé savant candidat doivent être identifiés comme une racine ou un mot (dans la liste de racines ou dans le dictionnaire). La concaténation de ces composants identifiés doit recomposer le composé savant, sinon l'identification de la forme du composé savant échoue. Plusieurs formes d'un composé savant candidat peuvent être identifiées, par exemple, un composé savant comme *histopathologie* peut être identifié comme composé de *histo* et *pathologie* ou comme composé de *histo*, *patho* et *logie*.

**Génération des traductions candidates**

Des traductions candidates sont générées pour chaque composé savant d'une forme identifiée tout en respectant les hypothèses expliquées en section 3.6.1. Les équivalents des ICFs et FCFs identifiés sont trouvés en utilisant la liste de racines alignées $NE_A$, tandis que les traductions des mots composants identifiés sont obtenues à partir du dictionnaire $Dico_{Bi}$.

Les traductions candidates d'un composé savant sont obtenues en formant toutes les combinaisons possibles (suivant le modèle gréco-latin) des traductions de chaque composant du composé savant ($NC_s$). Un tiret peut être ajouté entre les composants traduits lors de la génération des traductions candidates.

Par exemple, supposons que nous identifions les deux éléments (*neuro-* et *-logy*) comme des racines gréco-latines dans le composé savant EN *neurology*. Pour générer les traductions candidates françaises de ce composé savant, nous recherchons les équivalents (de type ICF) de *neuro-* dans $NE_A$, ce qui serait par exemple FR *névro* et *neuro*, ainsi que les équivalents (de type FCF) de *logy*, ce qui serait FR *-logie*. En conséquence, quatre traductions candidates seront générées en combinant les traductions des composants, cela donne : *neurologie*, *neuro-logie*, *névro-logie* et *névrologie*.

Prenons un autre exemple : nous voulons générer des traductions candidates anglaises pour FR *bioscience*, qui correspond à la deuxième forme des composés savants que nous traitons. Nous pouvons identifier FR *bio* comme racine gréco-latine et FR *science* comme un mot dans le dictionnaire. Supposons que l'ICF équivalent de FR *bio* est EN *bio*, obtenu à partir de $NE_A$, tandis que les traductions de FR *science* dans $Dico_{Bi}$ sont *art*, *science*, *information*, et *knowledge*. En conséquence, huit traductions candidates seront générées : *bioart*, *bio-art*, *bioscience*, *bio-science*, *bioinfomation*, *bio-infomation*, *bioknowledge* et *bio-knowledge*.

#### Sélection des traductions correctes

Chaque traduction candidate (obtenue lors de l'étape de génération) est cherchée dans la liste cible des composés savants $l_{NC_c}$. Dans le cas où une traduction candidate est trouvée dans $l_{NC_c}$, elle sera considérée comme une traduction correcte pour son composé savant source respectif $NC_s$. Par exemple, si deux traductions candidates en français ont été générées pour FR *neurologie* : *neurologie* et *névrologie*, elles seront recherchées dans la liste $l_{NC_c}$. Le candidat *névrologie* ne peut pas être trouvé car il n'existe pas comme mot, mais il y a une probabilité que *neurologie* soit trouvé dans $l_{NC_c}$, et sera donc considéré comme étant une traduction correcte.

### 3.6.4 Traductions semi-compositionnelles des composés savants candidats non-traduits compositionnellement

Cette méthode est appliquée sur les mots [18] qui : (a) n'existent pas dans le dictionnaire bilingue ; (b) sont identifiés comme contenant des racines gréco-latines mais aussi un ou plusieurs composants ne se trouvant pas dans le dictionnaire bilingue ou dans la liste de racines alignées. Par exemple, si seulement le composant FR *ectomie* a été identifié comme racine dans le composé savant FR *tumorectomie* et que le composant FR *tumor* n'a pas pu être identifié comme une racine ou un mot, la traduction semi-compositionnelle sera appliquée.

La traduction semi-compositionnelle suppose qu'au moins une racine gréco-latine dans un composé savant source $NC_s$ se transpose en une racine dans l'équivalent de $NC_s$.

18. De longueur supérieure ou égale à 5 et d'une fréquence supérieure ou égale à 2 ou 5.

**Variantes graphiques** Nous essayons d'abord d'aligner un composé savant candidat (source) avec un des composés savants (sources) déjà traduits dans la phase de traduction compositionnelle, en utilisant une mesure de similarité des lettres (ex. similarité basée sur la distance de Levenshtein). Si un composé savant source est aligné avec un autre composé savant source dont les traductions sont connues, les traductions du dernier seront proposées au premier. Les deux composés savants sources alignés doivent au moins partager une racine gréco-latine. La similarité entre les deux mots alignés doit être supérieure à 0,85 s'ils sont de longueurs < 10, sinon le seuil de similarité sera fixé à 0,75 [19]. Cela aide à trouver des traductions pour les variantes graphiques des composés savants (ex. FR *histologiques* sera aligné avec FR *histologique*, FR *carcinogénique* sera aligné avec FR *carcinogène*). Cette méthode est appelée la méthode **semi-compositionnelle 1**.

**Adaptation de la méthode de Morin et Daille (2012b)** Dans le cas où un composé savant candidat source ($NC_s$) n'a pas pu être traduit par la méthode compositionnelle ou semi-compositionnelle 1, nous appliquons l'approche distributionnelle de Rapp (1999) (présentée en section 2.1 du chapitre 2) afin de trouver une liste de traductions candidates ($L_{Tcs}$) pour $NC_s$. Ensuite, nous traduisons individuellement les composants identifiés dans $NC_s$. Puis, nous cherchons des traductions candidates qui sont dans les premiers $n$ traductions proposées dans $L_{Tcs}$ et qui comprennent au moins un des équivalents des composants identifiés dans $NC_s$. Les composants non-identifiés dans une traduction candidate trouvée doivent être chacun au minimum de longueur 3. Par exemple, si FR *tumorectomie* n'a pas été traduit par la méthode compositionnelle ou semi-compositionnelle 1 mais que son composant *ectomie* a été identifié comme racine, l'équivalent de cette racine en anglais sera cherché (ex. *ectomy*). Ensuite, tous les mots qui ont la forme [X + ectomy] dans la liste $L_{Tcs}$ seront cherchés, où X est une suite de caractères de longueur supérieure ou égale à 3. Ces traductions seront proposées comme meilleures traductions candidates pour $NC_s$. Par exemple, si la liste de traductions candidates $L_{Tcs}$ est EN {surgery, operation, lumpectomy, etc.} pour le terme FR *tumorctomie*, la traduction EN *lumpectomy* sera proposée car elle se termine par *ectomy* (l'équivalent de FR *ectomie*). Les traductions trouvées seront ordonnées par les scores qui leur sont donnés par l'approche distributionnelle. Cette méthode repose donc sur une propriété semi-compositionnelle et sur une approche distributionnelle pour traduire les composés savants, ce qui ressemble au travail présenté dans Morin et Daille (2012b;a) (voir section 2.3 dans le chapitre 2). Nous appelons cette méthode **semi-compositionnelle 2**.

## 3.7 Évaluation

Nous menons différentes expériences pour évaluer notre approche de traduction des composés savants. D'abord, nous essayons de traduire tous

19. Ces seuils ont été définis par des expériences menées sur un corpus comparable français-anglais dans le domaine du cancer du sein.

| Langue | Nb. d'adjectifs et de noms/Cancer du sein | Nb. d'adectifs et de noms/Énergies renouvelables |
|---|---|---|
| **EN** | 6 227 | 5 879 |
| **FR** | 4 153 | 5 102 |
| **DE** | 4 904 | 11 567 |

TABLE 3.3 – *Nombre d'adjectifs et de noms en corpus*

les composés savants candidats extraits d'un corpus à l'aide de notre approche compositionnelle (présentée en section 3.6.3). Ensuite, nous appliquons notre méthode semi-compositionnelle (présentée en section 3.6.4) sur une liste de mots inconnus dans le corpus et qui contiennent au moins une racine gréco-latine potentielle. Nous calculons la précision manuellement et nous calculons le rappel sur des listes de mots annotés.

Nous présentons dans les sections 3.7.1 et 3.7.2 les ressources utilisées pour les expériences. Nous présentons les résultats de nos méthodes dans la section 3.7.3.

### 3.7.1 Ressources

Nous effectuons des expériences en utilisant les deux corpus comparables *cancer du sein* et *énergies renouvelables* présentés en section 1.4.1 dans le chapitre 1. Les mots de corpus sont lemmatisés et étiquetés par l'outil TermSuite (présenté en section 2.6 dans le chapitre 2). La table 3.3 liste les langues avec le nombre correspondant de noms et d'adjectifs uniques dans chaque corpus.

Nous utilisons également des listes de racines gréco-latines (présentées en section 3.7.2) ainsi que des dictionnaires bilingues (présentés en section 1.4.2 du chapitre 1).

Nous disposons aussi d'une liste pour chaque corpus français et anglais, ces listes contiennent des mots qui (a) ont des fréquences supérieures à un certain seuil (2 pour le corpus des énergies renouvelables et 5 pour le corpus du cancer du sein, cette différence de fréquence choisie est dû au fait que dans le corpus des énergies renouvelables les composés savants sont moins fréquents), (b) sont d'une longueur supérieur à 5 et qui (c) n'existent pas dans le dictionnaire bilingue disponible. Chaque liste a été annotée manuellement de manière aléatoire jusqu'à trouver 140 composés savants et termes complexes. Les tables 3.4 et 3.5 montrent les tailles de ces listes pour chaque corpus et chaque langue, ainsi que le nombre de mots annotés et le nombre de composés savants trouvés. Ces listes sont utilisées pour calculer le rappel de nos méthodes de traduction de composés savants.

| Langue | Nb. mots | Nb. mots annotés | Nb. composés savants |
|---|---|---|---|
| EN | 979 | 504 | 94 |
| FR | 586 | 401 | 138 |

TABLE 3.4 – *Listes de mots annotés pour le corpus du cancer du sein*

| Langue | Nb. mots | Nb. mots annotés | Nb. composés savants |
|---|---|---|---|
| EN | 1 442 | 700 | 20 |
| FR | 936 | 841 | 80 |

TABLE 3.5 – *Listes de mots annotés pour le corpus des énergies renouvelables*

### 3.7.2 Listes de racines gréco-latines

La méthode suivie pour construire des listes monolingues et bilingues de racines gréco-latines est présentée dans cette section.

**Listes monolingues** Les listes monolingues de racines gréco-latines sont construites manuellement et semi-automatiquement, comme suit :

- une liste de racines en français est extraite de Béchade (1992). Ensuite, à partir de cette liste, nous construisons manuellement des listes équivalentes en deux langues : l'anglais et l'allemand. L'alignement manuel est effectué à l'aide des listes de racines gréco-latines monolingues trouvées sur le Web.
  Nous obtenons donc des listes monolignues de racines gréco-latines en trois langues. La table 3.6 présente les tailles de ces listes.

| Langue | FR | EN | DE |
|---|---|---|---|
| **Nb. racines** | 113 | 100 | 99 |

TABLE 3.6 – *Tailles des listes de racines gréco-latines construites manuellement*

  Certaines racines n'ont pas d'équivalents d'une langue à une autre. Par exemple, la racine *-thèque* comme dans *sérothèque*[20] n'a pas d'équivalent en anglais.
- un algorithme est développé pour enrichir les listes de racines gréco-latines monolingues. Cet algorithme se base sur les mots existant dans le corpus (partie monolingue), les listes de racines gréco-latines (construites manuellement), une liste de préfixes[21] ainsi que sur les parties monolingues de nos dictionnaires bilingues.
  Pour extraire une liste de racines gréco-latines du type ICF, nous suivons l'algorithme 3.1. Cet algorithme exploite le fait que la plupart des racines d'origine grecque se terminent par *o*.
  Afin d'extraire une liste de racines gréco-latines du type FCF, nous suivons l'algorithme 3.2.
  Les deux algorithmes sont appelés consécutivement de manière ité-

20. Se traduit par *serum bank* en anglais.
21. Nous utilisons une liste de 3 préfixes pour le français et l'anglais qui se terminent par *o* : hypo-, FR rétro-/EN retro- et pro-.

rative jusqu'à la convergence (aucune racine de plus n'est ajoutée aux listes de racines).

**Données** : $C$ (corpus), *Dico* (Dictionnaire monolingue), $L_{pref}$ (préfixes), $L_{FCF}$ (racines finales), $L_{ICF}$ (racines initiales) ;
**Résultat** : $L_{ICF}$ ;
**début**
  **pour** *chaque adjectif ou nom ou racine finale (M) dans* $\{C \cup L_{FCF}\}$ **faire**
    **si** *il se trouve un autre adjectif ou nom M' dans le corpus (ex. hématotumoral, lymphocèle), où M' peut s'écrire de la forme suivante : [élément + M] (ex. hématotumoral, lymphocèle), et si (1) élément se termine par o ; (2) élément* $\notin \{L_{pref} \cup Dico \cup L_{ICF}\}$ *; (3)* $|élément| >= 3$ & $|élément| <= 9$ **alors**
      Ajouter *élément* (ex. hémato, lympho) à $L_{ICF}$;

**Algorithme 3.1 :** Identification des racines gréco-latines initiales (ICF)

**Données** : $C$ (corpus), *Dico* (Dictionnaire monolingue), $L_{FCF}$ (racines finales), $L_{ICF}$ (racines initiales) ;
**Résultat** : $L_{FCF}$ ;
**début**
  **pour** *chaque racine initiale R dans* $L_{ICF}$ **faire**
    **si** *il se trouve un adjectif ou nom M' dans le corpus, où M' peut s'écrire de la forme suivante : [R + élément] (ex. aneuploïde), et si (1) élément* $\notin \{Dico \cup L_{FCF}\}$ *; (2)* $|élément| >= 3$ & $|élément| <= 9$ **alors**
      Ajouter *élément* (ex. ïde) à $L_{FCF}$;

**Algorithme 3.2 :** Identification des racines gréco-latines finales (FCF)

Nous appliquons ces algorithmes sur les corpus français et anglais. Cela mène à extraire, par exemple, une liste de 154 racines pour le corpus cancer du sein français. Les listes des racines ICF ont une précision entre 83 à 85 %. Les listes du type FCF ne sont pas fiables et doivent être vérifiées manuellement.

Dans ces listes, nous trouvons des pseudo-racines, c'est-à-dire des éléments qui ne sont ni d'origine grecque ni d'origine latine (ex. la racine *hormono* en français dans *hormonothérapie* est dérivée du mot *hormone*). Cependant, nous n'ajoutons ces pseudo-racines à notre liste de racines que si elles ont des équivalents graphiquement proches dans d'autres langues (ex. hormonothérapie se traduit par ES *hormonoterapia*, où *hormono* est une pseudo-racine commune entre les deux mots).

À noter que l'algorithme 3.1 peut être adapté pour extraire des racines qui se terminent par *i* (le cas des racines d'origine latine), mais la qualité de la liste de racines extraite sera moins bonne que celle de racines se terminant par *o*.

En outre, plusieurs racines combinées peuvent être extraites automatiquement comme étant une seule racine (ex. *histopatho*). Pour cela, une condition sur la longueur d'une racine extraite est établie, nous avons fixé le minimum de la longueur d'une racine à 3 parce que nous récupérons les racines qui se terminent par *o*, celles si ne peuvent pas être de longueur 2. Cette condition ne permet pas de filtrer toutes les fausses racines, un découpage automatique de racines combinées peut être envisagé dans la tâche de l'extraction des racines gréco-latines, mais nous choisissons de nettoyer ces listes manuellement afin d'avoir une précision de 100 %.

Nous obtenons de cette manière des listes de racines gréco-latines monolingues pour chaque corpus français ou anglais. Nous détaillons les tailles de ces listes dans la table 3.7.

| | **FR** | **EN** |
|---|---|---|
| **Cancer du sein** | 154 | 94 |
| **Énergies renouvelables** | 7 | 18 |

TABLE 3.7 – *Tailles des listes de racines extraites semi-automatiquement des corpus*

Nous résumons les tailles des listes de racines monolingues construites manuellement et semi-automatiquement dans la table 3.8. Les préfixes natifs qui peuvent préfixer les composés savants et qui sont identiques dans plusieurs langues sont ajoutés à la liste de ICFs : (FR anti, EN anti, DE anti), (FR post, EN post, DE post), (FR trans, EN trans, DE trans), (FR hypo, EN hypo, DE hypo), (FR rétro, EN retro, DE retro).

| | **FR** | **EN** | **DE** |
|---|---|---|---|
| **Nb. de racines** | 274 | 201 | 105 |

TABLE 3.8 – *Tailles des listes des racines gréco-latines monolingues*

### Listes bilingues

Nous alignons d'abord manuellement les listes de racines monolingues construites manuellement (présentées dans la table 3.6). Nous résumons les tailles des listes alignées dans la table 3.9.

| | **FR-EN** | **FR-DE** |
|---|---|---|
| **Taille** | 100 | 99 |

TABLE 3.9 – *Tailles des listes des racines gréco-latines manuellement alignées*

Les racines monolingues extraites semi-automatiquement dans une langue sont alignées avec leurs équivalents identiques dans une autre langue (après avoir établi des règles simples comme le remplacement de *é* par *e* quand nous alignons les racines du français vers l'anglais, ex. FR *séro* sera aligné avec EN *sero*). En procédant de cette manière, nous trouvons que 49 racines sont identiques (ou presque) entre le français et l'anglais. Les équivalents des racines non-alignées seront retrouvés manuellement. Afin de faciliter l'alignement, pour chaque racine, nous extrayons du dictionnaire bilingue les mots qui commencent ou se terminent par cette ra-

cine, et nous nous basons sur leurs traductions pour les aligner (voir la table 3.10).

| Racine | Mot français | Mots anglais |
|---|---|---|
| hémato | hématologique | hematological, haematological |
| rhumato | rhumatologie | rheumatology |
| rhumato | rhumatologue | rheumatologist |
| chimio | chimioluminescence | chemiluminescence, chemoluminescence |

TABLE 3.10 – *Exemples qui facilitent l'alignement manuel des racines hémato, rhumato et chimio avec leurs équivalents en anglais*

Finalement, en combinant tous les alignements des racines gréco-latines trouvés pour chaque paire de langues, nous obtenons des listes bilingues dont les tailles sont présentées dans la table 3.11.

| Langue | FR-EN | FR-DE |
|---|---|---|
| Taille | 254 | 105 |

TABLE 3.11 – *Tailles des listes des racines gréco-latines alignées manuellement et semi-automatiquement*

### 3.7.3 Résultats

Nous commençons par extraire des composés savants candidats pour chaque corpus et chaque langue. Ensuite, nous appliquons les méthodes de traduction sur les composés savants extraits pour une paire de langues.

Les tables 3.12 et 3.13 présentent les résultats obtenus avec notre méthode compositionnelle pour les paires de langues FR-EN, EN-FR et FR-DE avec les corpus du cancer du sein et des énergies renouvelables.

Par exemple, en utilisant notre méthode compositionnelle sur la paire de langues FR-EN et avec le corpus cancer du sein, des traductions candidates ont été générées automatiquement pour 876 des composés savants candidats français parmi 2 380 composés savants candidats. Parmi les 876 composés savants qui ont eu des traductions générées, 295 composés savants ont été alignés avec des traductions valides (trouvées dans le corpus cible). Seulement 4 composés savants candidats ont été alignés avec des mauvaises traductions (précision de 98,64 %). Parmi les composés savants traduits pour la paire de langue FR-EN, 21 5 % n'existent pas dans la base de référence IATE (http ://iate.europa.eu/.) [22]

Nous remarquons que le nombre de composés savants traduits pour la paire de langues FR-DE est moins élevé que le nombre de composés savants traduits pour les paires de langues FR-EN et EN-FR. En effet, la langue allemande emploie moins de racines gréco-latines que les autres langues européennes (Namer et Baud 2007). De plus, nous disposons d'une liste de racines gréco-latines moins complète pour la langue allemande que les listes des racines pour l'anglais et le français.

22. Ex. cytokératine, micropapillaire, clinicopathologie, cytoréducteur, micrométastase, etc.

| | Composés savants candidats | Traductions générées | Traductions trouvées | Précision |
|---|---|---|---|---|
| FR-EN | 2 380 | 876 | 295 | 98,64 % |
| EN-FR | 3 054 | 750 | 316 | 98,73 % |
| FR-DE | 2 380 | 625 | 57 | 100,00 % |

TABLE 3.12 – *Composés savants extraits et traduits par la méthode compositionnelle pour les langues FR-EN, EN-FR et FR-DE sur le corpus du cancer du sein*

| | Composés savants candidats | Traductions générées | Traductions trouvées | Précision |
|---|---|---|---|---|
| FR-EN | 898 | 261 | 100 | 97,00 % |
| EN-FR | 1294 | 318 | 110 | 95,45 % |
| FR-DE | 898 | 235 | 66 | 98,48 % |

TABLE 3.13 – *Composés savants extraits et traduits par la méthode compositionnelle pour les langues FR-EN, EN-FR et FR-DE sur le corpus des énergies renouvelables*

Les listes de références présentées dans les tables 3.4 et 3.5 nous permettent d'évaluer les méthodes compositionnelle et semi-compositionnelle en calculant le rappel. Le rappel est le nombre de composés savants traduits par notre méthode et existant dans la liste de références divisé par le nombre de composés savants dans la liste de références.

Les tables 3.14 et 3.15 présentent les résultats obtenus avec la méthode compositionnelle (**Compo.**) et la méthode semi-compositionnelle 1 (**Semi-1**) sur les listes de références. Par exemple, pour la paire de langues FR-EN, 48 composés savants ont été traduits avec une précision de 100 % (**P1** signifie précision au top 1) et un rappel de 21 % (c'est-à-dire, 29 composés savants parmi les 48 composés savants traduits se trouvent dans la liste de références de 138 composés savants annotés). Le rappel augmente de $\approx$ 8 % en utilisant la méthode semi-compositionnelle 1. La méthode semi-compositionnelle 1 n'a pas influencé la précision pour le corpus cancer du sein. Cependant, la précision baisse en utilisant cette méthode sur le corpus des énergies renouvelables, cela est surtout à cause de l'existence de beaucoup de mots étrangers dans ce corpus.

| | Nb. mots | Compo. | P1 | R | Compo. + Semi-1 | P1 | R |
|---|---|---|---|---|---|---|---|
| FR-EN | 586 | 48 | 100 % | 21,01 % | 86 | 100,00 % | 28,98 % |
| EN-FR | 979 | 61 | 100 % | 35,10 % | 85 | 100,00 % | 39,36 % |

TABLE 3.14 – *Traduction des composés savants par la méthode Semi-1 et Compo. pour les langues FR-EN et EN-FR sur le corpus du cancer du sein*

Nous évaluons la méthode semi-compositionnelle 2 (**Semi-2**) avec le corpus cancer du sein et sur la paire de langues FR-EN dans les deux directions. Pour pouvoir appliquer cette méthode, il faut d'abord extraire des traductions candidates par une approche distributionnelle. Nous uti-

| | **Nb. mots** | **Compo.** | **P1** | **R** | **Compo. + Semi-1** | **P1** | **R** |
|---|---|---|---|---|---|---|---|
| FR-EN | 936 | 6 | 100 % | 7,50 % | 19 | 52,63 % | 12,50 % |
| EN-FR | 1442 | 9 | 100 % | 19,00 % | 16 | 81,25 % | 30,00 % |

TABLE 3.15 – *Traduction des composés savants par la méthode Semi-1 et Compo. pour les langues FR-EN et EN-FR sur le corpus des énergies renouvelables*

| | **Nb. mots** | **Compo. + Semi-1 + Semi-2** | **P** | **R** |
|---|---|---|---|---|
| FR-EN | 586 | 142 | 71,80 % | 54,00 % |
| EN-FR | 979 | 244 | 40,57 % | 47,87 % |

TABLE 3.16 – *Traduction des composés savants par les méthodes Compo., Semi-1 et Semi-2 pour les langues FR-EN et EN-FR sur le corpus du cancer du sein*

lisons la méthode distributionnelle implémentée dans TermSuite afin de produire 100 traductions candidates pour chaque composé savant source dans les listes de références présentées dans les tables 3.4 et 3.5. Pour la paire de langue FR-EN, 56 composés savants de plus ont pu être alignés avec une précision de 23,21 % (13 traductions correctes).

La table 3.16 présente les résultats obtenus en utilisant la combinaison des méthodes (compositionnelle et semi-compositionnelle) sur le corpus cancer du sein FR-EN. Le rappel augmente de $\approx$ 25 % mais la précision baisse de $\approx$ 28 % par rapport à l'utilisation de la méthode compostionnelle seule. En effet, 49 composés savants sur 138 annotés, ont été correctement traduits, 48 n'ont pas reçu de traductions et 41 ont été mal traduits. Les 41 mauvaises traductions ont été obtenues par la méthode semi-compositionnelle 2.

En général, les résultats obtenus sur le corpus cancer du sein sont meilleurs que ceux obtenus sur le corpus des énergies renouvelables. En effet, le domaine du cancer du sein emploie plus de composés savants que le domaine des énergies renouvelables. De plus à partir des listes annotées, nous remarquons que beaucoup de composés savants employés dans le domaine du cancer du sein contiennent des racines spécifiques au domaine, alors que les composés savants dans le domaine des énergies renouvelables contiennent surtout des racines gréco-latines générales comme : multi, mono, mini, poly, micro. Ajoutons à cela le fait que dans les corpus des énergies renouvelables, ils existent des composés savants étrangers, mais ces mots ont été extraits parce qu'ils contiennent des racines gréco-latines identiques parfois entre les langues (ex. *technology* a été extrait du corpus cancer du sein français).

### 3.7.4 Analyse des erreurs

Nous analysons les erreurs obtenues par la méthode proposée pour l'identification des composés savants ainsi que par les méthodes proposées pour la traduction des composés savants.

**Identification des composés savants candidats** Un mot dans un corpus de langue ($l$) peut être extrait parce qu'il contient une chaîne identique à une racine gréco-latine même si ce mot n'est pas un vrai composé savant de cette langue. Ce mot peut être de langue $l$ mais la racine qu'il contient est fausse. Par exemple, un candidat comme EN *decision* [23] sera décomposé en deux composants : le premier est *deci*, il sera considéré comme une racine gréco-latine (fausse racine). À titre d'exemple, en annotant manuellement une liste de 401 mots inconnus du dictionnaire, nous avons trouvé que 138 mots parmi les 401 mots sont des composés savants (34 % du vocabulaire de la liste). Or, notre méthode d'identification extrait 2 380 composés savants candidats du corpus, c'est-à-dire 57 % des adjectifs et des noms dans le corpus. Cela indique qu'au moins $\approx$ 25 % de composés savants candidats ne sont pas de vrais composés savants.

**Méthode compositionnelle** En utilisant la méthode compositionnelle, les faux alignements sont principalement des traductions obtenues à partir de faux composés savants extraits (du bruit ou des mots qui ne sont pas des composés savants). Par exemple le mot français *histoire* a été extrait du corpus anglais à partir de *histo* qui a été identifié comme racine gréco-latine et *ire* qui a été identifié comme étant un mot anglais. Ainsi, *histoire* (un mot étranger dans le corpus anglais) a été aligné avec FR *histoire* (*ire* est une traduction française du mot EN *ire*).

Des traductions erronées sont également obtenues du fait que les composés savants ne sont pas toujours traduits en des composés savants d'une langue à une autre, par exemple, le mot FR *télécommande* a été traduit en anglais par *telecontrol*, tandis que la traduction correcte est EN *remote control*.

Un composé savant candidat ($NC_s$) peut être identifié parce qu'il contient une racine gréco-latine. Il se peut cependant qu'il ne soit identifié comme aucune des formes néoclassiques que notre méthode traite. Dans ce cas, la génération de ses traductions candidates échoue même si $NC_s$ est un vrai composé savant qui peut être traduit par un composé savant. Cela peut avoir plusieurs raisons :

- **une racine manquante de la liste de racines alignées ($NE_A$)** :
  Supposons qu'un candidat comme FR *métronome* soit extrait et que l'équivalent de la racine *métro* se trouve dans $NE_A$, la génération échoue si *nome* (une vraie racine) n'existe pas dans $NE_A$.
- **composé savant mal lemmatisé** :
  Si un composé savant candidat comme le mot FR *aérogénérateurs* a été extrait dans la forme plurielle ou fléchie et qu'il a été mal lemmatisé,

23. *deci* existe dans la liste de racines gréco-latines disponible.

la génération pourra échouer. Par exemple, le composé savant *aéro-générateurs* peut être décomposé en *aéro* et *générateurs* : si *générateurs* n'existe pas dans le dictionnaire bilingue parce qu'il est en pluriel, la génération échouera. En examinant une liste de 138 composés savants dans le corpus français du cancer du sein, nous trouvons que $\approx$ 23 % de ces composés savants ont été mal lemmatisés (ex. unilatéraux, histologiques, etc.).

– **forme néoclassique non-traitée** :
Un vrai composé savant candidat qui appartient à une forme que nous n'avons pas traitée peut être extrait. Il existe bien évidemment d'autres formes de composés savants, par exemple, FR *annexectomie* (annexe : mot simple, ectomie : FCF) est une forme que notre méthode ne traite pas. En examinant une liste de 138 composés savants français (extraite du corpus cancer du sein français), nous trouvons que $\approx$ 5 % de ces composés savants appartiennent à une forme non-traitée (ex. FR *annexectomie*[24] est de la forme [Mot + FCF]).

Une traduction candidate générée qui ne se trouve pas dans la liste cible ne signifie pas nécessairement que c'est une mauvaise traduction, il se peut qu'une traduction générée soit absente du corpus cible et ne soit donc pas validée comme traduction pour un composé savant source.

**Méthode semi-compositionnelle 1** La plupart des erreurs induites par la méthode semi-compositionnelle 1 résultent des mots étrangers considérés comme des composés savants candidats. Par exemple, si EN *technology* est extrait du corpus français, il peut être aligné avec le composé savant FR *technologie* correctement traduit par l'approche compositionnelle. En d'autres mots, le problème de l'utilisation des mesures de similarité pour trouver les traductions des mots inconnus étant des composés savants candidats dans un corpus, c'est que ces mesures trouvent des traductions pour des mots étrangers dans le corpus (les corpus non-anglais emploient des mots anglais).

**Méthode semi-compositionnelle 2** Cette méthode dépend des résultats obtenus par la méthode distributionnelle. Afin de trouver la traduction correcte d'un composé savant candidat, il faut que cette traduction soit un mot simple dans la liste proposée par la méthode distributionnelle et qu'au moins un des composants du composé savant se transpose dans sa traduction. Ces deux hypothèses ne sont pas toujours vérifiées.

## 3.8 Conclusion

Dans ce chapitre, nous avons présenté une approche pour aligner les composés savants entre deux langues (source-cible). L'approche gère principalement deux types de composés savants et emploie des méthodes compostionnelle et semi-compositionnelle pour traduire les composés savants.

24. Le composant *annexe* n'est pas une racine gréco-latine ou une pseudo-racine.

La méthode compositionnelle utilise une liste de racines alignées et un dictionnaire bilingue. Les résultats ont montré une haute précision pour la traduction des composés savants (plus de 95 %). Les méthodes semi-compositionnelles ont été proposées pour traduire les composés savants candidats qui n'ont pas été traduits par la méthode compositionnelle. La première dépend d'une mesure de similarité de lettres et la deuxième dépend d'une méthode distributionnelle. Il est envisageable d'étendre l'approche afin de couvrir d'autres formes de composés savants. La traduction compositionnelle des composés savants en des mots simples et complexes a été largement étudiée et validée par Delpech et al. (2012).

Dans le chapitre suivant, nous étudions la traduction compositionnelle pour les termes complexes qui contiennent des adjectifs relationnels. Ces adjectifs demandent parfois un traitement spécial avant de procéder à une traduction compositionnelle. Pour cela, nous étudions également l'extraction et l'analyse des adjectifs relationnels.

# Extraction, alignement et traduction des adjectifs relationnels

4

Sommaire



Ce chapitre s'intéresse à la traduction compositionnelle de termes complexes français de la forme [Nom + Adjectif Relationnel] (ex. consommation alcoolique). Nous nous intéressons à l'extraction automatique des

adjectifs relationnels français d'un corpus puis à l'alignement de ces adjectifs avec leurs noms de base (ex. *alcoolique* sera aligné avec *alcool*). Nous utilisons des mesures de similarité de lettres et une similarité contextuelle pour établir des alignements d'adjectifs avec des noms. Ces alignements sont ensuite utilisés pour traduire des termes complexes par une méthode compositionnelle de base. Les résultats montrent que l'utilisation de ces alignements permet de traduire des termes complexes français vers l'anglais avec une bonne précision.

## 4.1 Introduction

Les chapitres précédents présentent le principe des approches compositionnelles adaptées à la traduction des termes complexes. Une approche compositionnelle qui consiste à traduire individuellement les composants d'un terme complexe source à l'aide d'un dictionnaire bilingue peut donner une bonne précision. Cependant, cette traduction échoue si l'un des composants du terme complexe n'existe pas dans le dictionnaire bilingue, ou si l'ensemble de traductions des composants ne permet pas d'obtenir la traduction correcte. La traduction compositionnelle peut donc échouer quand il existe une variation de structure inter-langue. Par exemple le terme français *consommation alcoolique* (de la forme [N + A][1]) peut être traduit en anglais par un terme d'une structure différente : *alcohol consumption* (de la forme [N + N]). Il est peu probable que FR *alcoolique* soit traduit par EN *alcohol* dans un dictionnaire bilingue puisque le premier est un adjectif et le deuxième est un nom.

Nous nous intéressons aux termes complexes de la forme [N + AdjR] (*AdjR* désigne un adjectif relationnel), ex. cancer pulmonaire. En effet, ces termes peuvent être traduits compositionnellement dans une autre langue par des termes de la forme [N + N] (ex. *cancer pulmonaire* est traduit en anglais par *lung cancer*). Si le substantif *lung* n'est pas une traduction de l'adjectif *pulmonaire* dans le dictionnaire, le lien entre *pulmonaire* et *lung* peut être établi via le substantif *poumon* dont *pulmonaire* est le dérivé et *lung* la traduction. Ainsi, nous pouvons traduire *cancer pulmonaire* par *lung cancer* en utilisant la paraphrase *cancer du poumon*. Ce processus est appelé « la traduction par paraphrase » et il est illustré par la figure 4.1.

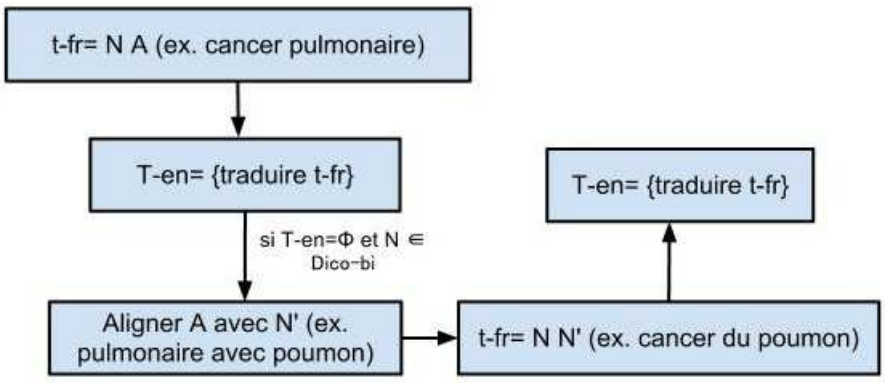


Figure 4.1 – *Traduction par paraphrase (où t-fr est le terme français, T-en est l'ensemble des traductions anglaises et Dico-bi est le dictionnaire bilingue français-anglais)*

Cette stratégie d'alignement de deux mots morphologiquement liés pour la traduction compositionnelle a été déjà explorée par Morin et Daille (2010). Des règles définies qui relient un adjectif relationnel avec son nom y ont été utilisées, nous avons présenté cette approche dans la section 2.3 du chapitre 2.

L'objectif de ce chapitre est de proposer une approche pour : (a) extraire

1. N signifie nom et A signifie adjectif.

des adjectifs relationnels automatiquement du corpus ; (b) établir un lien entre un adjectif relationnel extrait et le nom dont il est dérivé automatiquement ; (c) étudier l'influence des propriétés des adjectifs extraits et les alignements adjectif-nom sur la traduction compositionnelle des termes de la forme [N + AdjR].

Nous développons dans ce chapitre une approche qui nous permet d'extraire automatiquement des adjectifs relationnels d'un corpus français. Ensuite, nous proposons une approche afin de relier un adjectif relationnel (extrait précédemment du corpus) à un nom existant dans un dictionnaire bilingue et dans le corpus. Si la plupart des adjectifs relationnels sont dérivés par suffixation à partir de noms populaires (ex. cancéreux/cancer), il existe d'autres qui sont construits à partir de racines supplétives (gréco-latines) des noms (ex. médullaire/moelle). Nous traitons ces deux cas séparément : (a) **adjectif relationnel commun** : nous supposons qu'un adjectif relationnel partage un certain nombre de lettres avec son nom de base et que l'ordre des lettres est conservé. Ainsi, un score entre un adjectif relationnel et chaque nom du dictionnaire (et qui existe dans le corpus) sera obtenu en fonction de la similarité de lettres, nous exploitons ensuite le contexte afin que ce score soit plus représentatif ; (b) **adjectif relationnel savant** : nous vérifions si un adjectif relationnel peut être relié avec un nom à l'aide d'une racine supplétive. Nous utilisons les alignements obtenus par l'approche d'alignement adjectif-nom dans la traduction compositionnelle par paraphrase des termes de la forme [N + AdjR].

## 4.2 Adjectifs relationnels

Dans cette section, nous présentons la classe des adjectifs relationnels et ses propriétés, ainsi que des travaux qui se sont intéressés à l'identification de cette classe et les problèmes liés à cette identification.

### 4.2.1 Définition et propriétés

D'après (Dubois et Dubois-Charlier 1999, p. 128), « *un adjectif relationnel est issu d'une relative, où <de N> est caractérisé par l'absence de déterminant ; cette relative se branche directement sur l'antécédent auquel elle se rapporte, et l'ensemble formé du nom et de l'adjectif nominal suffixé forme un nom composé* ». Exemple : ce corps chimique est l'acide qui est <de nitre> ; ce corps chimique est l'acide nitrique.

Les adjectifs relationnels sont des adjectifs dénominaux (adjectifs construits sur des bases nominales) ; il ne faut pas les confondre avec les adjectifs déverbaux qui sont dérivés d'un verbe par des suffixes tels que -able, -ible, -ile, -ant, etc. (ex. dégradable/de dégrader). Alors qu'un adjectif dit *qualificatif* (*AdjQ*) peut aussi être construit sur une base nominale, la relation [N + AdjQ] est différente de la relation [N + AdjR]. Par exemple, dans la phrase *François a des jambes éléphantesques*, l'adjectif *éléphantesques*

n'établit pas une relation entre les jambes de François et la catégorie *éléphant*, il leur attribue une qualité des individus de cette catégorie : être énorme, exemple extrait de Roché (2006).

Dans (Dubois et Dubois-Charlier 1999, p. 129) et (Goes 1999, p. 251), certaines propriétés des adjectifs relationnels sont citées. Nous résumons ces propriétés et les présentons sous le titre de propriétés linguistiques dans la table 4.1. D'autres propriétés que nous appelons « opérationnelles » et qui se sont basées sur les propriétés linguistiques sont présentées également dans la table 4.1. Les propriétés opérationnelles ne sont pas toujours exclusives aux adjectifs relationnels mais elles nous permettent de repérer des adjectifs automatiquement dans un corpus. Nous avons donné des labels (mis entre parenthèses) dans la table 4.1 à chaque propriété d'adjectifs relationnels afin de faciliter la référence à ces propriétés quand elles sont utilisées par les approches proposées dans ce chapitre.

Les adjectifs relationnels sont dérivés par suffixation d'un nom. Les suffixes des adjectifs relationnels peuvent être : -ien, -ois, -ique, etc. (voir P7 (suffixes) dans la table 4.1). Toutefois, la détection automatique des noms de base dont les adjectifs relationnels sont dérivés ne peut se faire par une simple comparaison entre la base nominale et l'adjectif relationnel dé-suffixé à cause de l'allomorphie des bases ; « *l'addition d'un suffixe peut entraîner des modifications morphologiques de la base nominale, elles sont plus ou moins importantes selon la nature de N ou selon la nature du suffixe* » (Dubois et Dubois-Charlier 1999, p. 135). Par exemple, ces modifications peuvent être : la modification phonique ou graphique de *N* (tropique/tropical), l'addition de voyelles ou de syllabes (nom/nominal), la modification du radical à partir du latin (bête/bestial), etc. Par ailleurs, les adjectifs relationnels et les adjectifs déverbaux ont quelques suffixes en commun, qui sont : -if, -aire, -eux, -oire, et -é. La catégorie d'un adjectif ne peut donc pas être déterminée en ne s'appuyant que sur son suffixe.

### 4.2.2 Problèmes et travaux concernant l'identification des adjectifs relationnels

Nous présentons dans cette section des problèmes liés à l'identification des adjectifs relationnels et des travaux qui peuvent extraire de tels adjectifs d'un corpus.

**Problèmes** La tâche d'identification des adjectifs relationnels dans un corpus n'est pas évidente : d'une part la classe des adjectifs relationnels est floue, et d'une autre, il n'y a pas de règles véritablement sûres pour les identifier automatiquement (Goes 1999, Maniez 2005). De plus, les adjectifs relationnels dérivent, avec le temps, de façon régulière vers la qualification (Noailly 1999, p. 24). Par exemple, certains adjectifs peuvent jouer un rôle relationnel ou qualificatif selon le contexte (ex. le système nerveux (*AdjR*) vs. François est nerveux (*AdjQ*)). Un adjectif peut donc avoir deux interprétations dans un syntagme, l'une relationnelle et l'autre qualificative. Par exemple, *une chaise royale* : est-elle la chaise du roi ou une chaise

luxueuse ? Si l'adjectif *royale* est identifié comme relationnel et ensuite aligné avec le nom *roi* quand il s'agit d'une utilisation qualificative de cet adjectif : *chaise royale* sera paraphrasé par *chaise du roi*. Cette paraphrase peut être traduite compositionnellement en anglais par *chair of the king* qui est une traduction erronée dans ce cas.

L'alignement d'un adjectif qualificatif avec un nom peut donc introduire de mauvaises traductions quand il s'agit d'utiliser cet alignement dans une optique de traduction compositionnelle. Cependant, quand un adjectif peut avoir un emploi relationnel, l'alignement de cet adjectif avec son nom de base peut aider à la traduction des termes de la forme [N + AdjR] avec une haute précision comme il a été montré dans Morin et Daille (2010).

**Travaux connexes** Daille (2000) exploite des règles de désuffixation-recodage (définies manuellement pour le français et l'anglais) pour relier un adjectif relationnel avec son nom de base (ex. la règle (-estière, -êt) peut relier *forestière* à *forêt*). Un adjectif $A$ extrait à l'aide de ces règles et qui doit apparaître avec un nom recteur $X$ sous la forme [X + A], sera considéré comme relationnel s'il peut être paraphrasé par un groupe [PREP + DET ? + N′] [2] apparaissant dans un syntagme de la forme [X + PREP + DET ? + N′] ; où $N'$ est le nom dont $A$ est dérivé (voir P8 (paraphrases) dans la table 4.1). La recherche des paraphrases est faite à partir du corpus. Cette méthode donne une précision de 99 %, mais un faible rappel dû au nombre limité de paraphrases dans le corpus.

Maniez (2005) examine deux approches pour identifier les adjectifs relationnels dans un corpus anglais spécialisé : (a) il se penche sur l'hypothèse que dans un corpus spécialisé, la plupart des adjectifs sont relationnels. Ainsi, il exploite P1 (degré) et P4 (attribut) (voir la table 4.1) afin de filtrer les adjectifs non-relationnels dans le corpus ; (b) tous les adjectifs en deuxième position extraits à partir du motif [ADJ1 + ADJ2 + N] sont sélectionnés en tant qu'adjectifs relationnels. Ce motif peut être adapté en français par [N + ADJ1 + ADJ2] en ajoutant le critère suivant : si $ADJ2$ est relationnel alors $ADJ1$ sera également relationnel. La raison pour laquelle nous considérons que l'adjectif en première position est relationnel, c'est parce que l'adjectif relationnel suit immédiatement le nom (Pedreira 2002), et qu'on détermine avant de qualifier (ex. un discours présidentiel intéressant). Nous concluons donc qu'un adjectif qualificatif ne peut pas précéder un adjectif relationnel, cette propriété est décrite sous P9 (NAdj1Adj2) dans la table 4.1. En d'autres mots, si $ADJ2$ est relationnel dans un syntagme de la forme [N + ADJ1 + ADJ2] alors $ADJ1$ ne pourra pas être qualificatif.

Cartoni (2008) travaille sur les mots préfixés de la forme : [préfixe + Mot] (ex. antitumoral). Cartoni (2008) constate qu'avec certains préfixes (comme *post-*), si *Mot* est un adjectif alors il s'agit d'un adjectif relationnel (ex. *postopératoire*). Avec un autre groupe de préfixes (comme *anti-*), *Mot*

2. PREP signifie préposition, DET signifie déterminant, ? signifie que DET peut apparaître une ou zéro fois.

est soit un adjectif relationnel, soit un adjectif déverbal (ex. antiséparatif) (Cartoni 2008, p. 255 )(voir P11 (préfixes) et P12 (racines) dans la table 4.1).

| **Propriétés linguistiques** | |
|---|---|
| P1 (degré) | « *ils n'acceptent pas d'adverbe de degré* » (acide très nitrique, sauf cas particulier) (Dubois et Dubois-Charlier 1999). Les adjectifs relationnels « *refusent la gradation en général, et très en particulier* » (Goes 1999). |
| P2 (antéposé) | « *ils ne peuvent pas être antéposés* » (le nitrique acide). |
| P3 (adverbialisation) | « *ils ne sont pas susceptibles d'adverbialisation* » (nitriquement) « *ni de nominalisation* » (nitricité). |
| P4 (attribut) | « *ils ne s'emploient pas en fonction d'attribut* » (cet acide est nitrique, sauf cas particulier). |
| P5 (coordination) | « *la coordination d'un adjectif relationnel avec un adjectif qualificatif est impossible* ». |
| P6 (antonymes) | « *ils ne forment généralement pas de séries antonymes* ». |
| **Propriétés opérationnelles** | |
| P7 (suffixes) | les suffixes des adjectifs relationnels : *-ique, -aire, -eux, -ier, -ien, -ois, -ain, -al, -el, -estre, -il, -in, -esque, -é, -if*. |
| P8 (paraphrases) | il existe des paraphrases dans un corpus monolingue de la forme [X + PREP + DET ? + N′] : : [X + AdjR] (ex. cancer du poumon : : cancer pulmonaire) ; où *X* est un nom, N′ est le nom de base de *AdjR*. (*PREP* signifie préposition et *DET* signifie déterminant, *?* signifie que *DET* peut apparaître une ou zéro fois) |
| P9 (NAdj1Adj2) | dans les syntagmes de la forme [N + Adj1 + Adj2] (ex. rupture capsulaire ganglionnaire), si *Adj*2 est un adjectif relationnel, *Adj*1 est relationnel également. |
| P10 (NAdjEtAdj) | dans les syntagmes de la forme [N + Adj1 + et/ou + Adj2] (ex. facteurs environnementaux ou génétiques), si *Adj*2 est un adjectif relationnel, *Adj*1 est relationnel également. |
| P11 (préfixes) | ils peuvent être préfixés par les préfixes : *post-, trans-, uni-, anti-, tri-, pré-*. |
| P12 (racines) | ils peuvent être préfixés par des racines gréco-latines : *micro-, séro-, radio-, etc*. |

TABLE 4.1 – *Propriétés linguistiques et opérationnelles des adjectifs relationnels*

Nous proposons dans la section suivante une méthode pour extraire une liste des adjectifs relationnels du corpus à l'aide des propriétés présentées.

## 4.3 Extraction des adjectifs relationnels du corpus

La reconnaissance automatique des adjectifs relationnels en corpus pose un certain nombre de problèmes comme nous l'avons constaté en section 4.2 : (a) ambiguïté des suffixes ; (b) ambiguïté de la classe relationnel/qualificatif ; (c) indice de relation exprimé par des propriétés non-présentes en corpus. Dans cette section, nous développons une approche pour extraire automatiquement des adjectifs relationnels du corpus.

### 4.3.1 Méthode d'extraction

Afin d'extraire des adjectifs relationnels du corpus, nous exploitons quelques propriétés linguistiques et opérationnelles présentées dans la table 4.1. Nous partons de l'hypothèse que les racines gréco-latines et certains préfixes français préfixent des adjectifs non-qualificatifs pour extraire une liste d'adjectifs initiale (en utilisant les propriétés P11 (préfixes) et P12 (racines)). Nous étendrons cette liste en utilisant d'autres propriétés (P9 (NAdj1Adj2) et P10 (NAdjEtAdj)). La méthode d'identification automatique des adjectifs relationnels du corpus que nous proposons est présentée dans l'algorithme 4.1 (nous faisons référence dans cet algorithme aux propriétés listées dans la table 4.1). L'ensemble des listes que nous extrayons sera utilisé par l'approche d'alignement adjectif-nom que nous présentons en section 4.4.

#### Remarques sur l'algorithme

1. Des racines (ex. *bio-*) peuvent préfixer des adjectifs déverbaux (ex. biodégradable). Cependant, les adjectifs préfixés par ces racines et se terminant par un suffixe qui ne peut pas être déverbal (ex. *-ique* dans *biochimique*), sont considérés comme étant des adjectifs relationnels.
2. Afin de trouver les adverbes construits à partir d'un adjectif dans le corpus : (a) nous ajoutons le suffixe adverbial *ment* (et d'autres adaptations du suffixe) à l'adjectif ; (b) nous cherchons ces adverbes construits dans le corpus.
3. L'extraction des adjectifs relationnels par le biais de la propriété P9 (NAdj1Adj2) est plus fiable que l'extraction de ces adjectifs par P10 (NAdjEtAdj). En effet, P10 (NAdjEtAdj) peut introduire du bruit quand il s'agit d'une utilisation qualificative d'un adjectif. Pour cette raison, nous choisissons de ne l'appliquer qu'avec les adjectifs relationnels qui sont trouvés à l'aide de P11 (préfixes) et P12 (racines). Ces adjectifs sont en effet moins susceptibles d'avoir un emploi qualificatif.
4. Bien que des adjectifs déverbaux puissent être extraits par cet algorithme, ils peuvent être reliés la plupart du temps à des substantifs (ex. végétatif ; Verbe : végéter ; Nom : végétation).

**Données** : $L_{prefixes}$ (préfixes français qui n'acceptent qu'une base adjectivale relationnelle ou déverbale), $L_{suffRel}$ (suffixes relationnels), $L_{racines}$ (racines gréco-latines extraites du corpus), $C_{cds_{fr}}$ (corpus français), $L_{[N+A]_{fr}}$ (termes de la forme [N + A] extraits du corpus français);

**Résultat** : $Liste_{AdjR}$-1 (contient les adjectifs qui commencent par une racine dans $L_{racines}$ ou un préfixe dans $L_{prefixes}$), $Liste_{AdjR}$-2 (contient les adjectifs qui peuvent être trouvés sous une forme préfixée par des racines dans $L_{racines}$ ou des préfixes dans $L_{prefixes}$), $Liste_{AdjR}$-3 (contient les adjectifs extraits à l'aide de P9 (NAdj1Adj2)), $Liste_{AdjR}$-4 (contient les adjectifs relationnels extraits à l'aide de P10 (NAdjEtAdj));

**début**

  **pour** *chaque A qui apparaît dans au moins un terme [N + A]* $\in L_{[N+A]_{fr}}$, *et qui se termine par un suffixe* $\in L_{suffRel}$ *(ex. tumoral)* **faire**

    **si** *il existe un autre A″ (ex. hématotumoral) dans* $C_{cds_{fr}}$ *qui a la forme [racine + A ] ou [préfixe + A] (ex. racine=hémato, A=tumoral) (où préfixe* $\in L_{prefixes}$, *racine* $\in L_{racines}$*)* **alors**

      **si** *le « préfixe » ou la « racine » accepte seulement des bases relationnelles* **alors**

        Ajouter $A''$ à $Liste_{AdjR}$-1 et $A$ à $Liste_{AdjR}$-2;

      **sinon**

        **si** *le suffixe de A est un suffixe non-commun entre les adjectifs relationnels et les adjectifs déverbaux (ex. le suffixe -ique, voir la remarque 1 en section 4.3.1)* **alors**

          Ajouter $A''$ à $Liste_{AdjR}$-1 à et $A$ à $Liste_{AdjR}$-2;

  $temp_{AdjR} \longleftarrow \{ Liste_{AdjR}\text{-}1 \cup Liste_{AdjR}\text{-}2 \}$;

  **répéter**

    **pour** *chaque adjectif AdjR dans* $temp_{AdjR}$ *qui vérifie P1 (degré) (avec très)* **faire**

      $tempList_{A''} \longleftarrow$ Trouver tous les adjectifs qui précèdent $AdjR$ immédiatement dans les syntagmes de la forme : [N + $A''$ + AdjR] (ex. profil protéique tumoral);

      **pour** *chaque* $A'' \in tempList_{A''}$ **faire**

        **si** $A''$ *respecte les propriétés P1 (degré) (avec très) et P3 (adverbialisation) (voir la remarque 2 en section 4.3.1)* **alors**

          Ajouter $A''$ à $Liste_{AdjR}$-3;

          $temp_{AdjR} \longleftarrow temp_{AdjR} \cup Liste_{AdjR}$-3;

  **jusqu'à** *Aucun nouvel adjectif ajouté à* $Liste_{AdjR}$-3;

  **pour** *chaque adjectif AdjR dans* $Liste_{AdjR}$-1 **faire**

    $tempList_{A''} \longleftarrow$ Trouver tous les adjectifs qui sont en coordination avec $AdjR$ (ex. mammaire et tumoral);

    **pour** *chaque* $A'' \in tempList_{A''}$ **faire**

      **si** $A''$ *respecte les propriétés P1 (degré) (avec très) et P3 (adverbialisation)* **alors**

        Ajouter $A''$ à $Liste_{AdjR}$-4;

**Algorithme 4.1 :** Extraction des adjectifs relationnels du corpus

### 4.3.2 Identification des racines gréco-latines

La méthode d'extraction des adjectifs relationnels que nous avons présentée dans l'algorithme 4.1 exploite une liste de racines gréco-latines. Nous rappelons que nous supposons que les racines gréco-latines préfixent les bases adjectivales non-qualificatives. Certaines racines ne peuvent préfixer que des adjectifs relationnels, alors que d'autres peuvent également préfixer des adjectifs déverbaux.

Nous avons déjà présenté dans le chapitre 3 (section 3.7.2) une méthode dont le but est d'extraire automatiquement des racines gréco-latines. Nous suivons un algorithme similaire pour extraire une liste de racines sauf que nous essayons de regrouper les racines dans deux catégories : celles qui ne préfixent que les adjectifs relationnels et celles qui peuvent aussi préfixer les adjectifs déverbaux.

Nous présentons dans l'algorithme 4.2 la méthode que nous développons pour extraire des racines et les regrouper. Nous ne filtrons pas cette liste manuellement.

**Données** : $C_{cds_{fr}}$ (corpus français), $L_{[N+A]_{fr}}$ (termes français de la forme [N + A]), $L_{suffRel}$ (suffixes relationnels) ;
**Résultat** : $L_{racines}$ ;
**début**
  **pour** *chaque adjectif A dans* $C_{cds_{fr}}$ *qui compose un terme [N + A] dans* $L_{[N+A]_{fr}}$ *(ex. barrière tumoral)* **faire**
    **si** *il existe un autre adjectif* $A'$ *dans le corpus (ex. hématotumoral), où* $A'$ *peut s'écrire de la forme suivante : [élément + A] (ex. hémato<u></u>tumoral), et si cet élément se termine par o, et s'il n'est pas l'un des préfixes français qui se terminent par o : hypo-, rétro- ou pro-* **alors**
      Ajouter *élément* (ex. hémato) à $L_{racines}$;
      **si** *élément préfixe au moins un adjectif Adj dans* $C_{cds_{fr}}$ *où Adj se termine par un suffixe* $\notin L_{suffRel}$ **alors**
        *élément* est une racine qui peut préfixer les adjectifs déverbaux (ex. bio-);
      **sinon**
        *élément* est une racine qui ne préfixe que les adjectifs relationnels (ex. micro-);

**Algorithme 4.2 :** Identification des racines gréco-latines

## 4.4 Alignement d'un adjectif relationnel avec un nom

Dans cette section, nous présentons des méthodes pour aligner un adjectif relationnel avec son nom de base. Nous commençons d'abord par introduire ces méthodes puis nous présentons des travaux connexes.

**Introduction** Nous supposons qu'un adjectif relationnel partage des lettres dans le même ordre avec son nom de base. Afin de trouver le nom $N$ dont un adjectif $A$ est dérivé, cet adjectif est comparé avec tous les noms qui existent dans un dictionnaire et un corpus source. Nous calculons d'abord un score entre un nom et un adjectif par des mesures de similarité de lettres. De nombreux algorithmes existants peuvent mesurer la similarité ou la distance entre deux chaînes. Une mesure intéressante pour notre tâche préservera l'ordre linéaire des lettres lors de la comparaison de deux chaînes. Nous utilisons des mesures de similarité afin de calculer un premier score entre un adjectif et un nom.

Nous supposons également qu'un adjectif relationnel et son nom de base partagent des paraphrases en commun comme le travail de Daille (2000) qui est présenté sous la section 4.2.2. Nous cherchons donc des paraphrases entre un couple d'un adjectif et un nom afin que son score obtenu par les mesures de similarité de lettres soit plus représentatif. Enfin, nous utilisons des racines gréco-latines pour relier les adjectifs relationnels supplétifs avec des noms.

**Travaux connexes** Les mesures de similarité de lettres peuvent être utilisées pour identifier des relations entre des mots d'une même langue, par exemple, elles peuvent regrouper les variantes graphiques d'un mot (ex. *éolien* avec *éolienne*). Les mesures de similarité peuvent aussi être utilisées pour identifier les cognats entre deux langues (mots similaires orthographiquement ayant un sens similaire, ex. FR *activiste* avec EN *activist*).

Il existe de nombreuses mesures de similarité de lettres qui ont été proposées dans la littérature. Nous décrivons ci-dessous quelques mesures de similarité de lettres. Chaque mesure retourne une valeur entre 0 et 1, cette valeur décrit la similarité entre deux chaînes.

- LCS : cette mesure consiste à trouver la sous-séquence la plus longue *Longest Common Subsequence*[3] entre deux chaînes, l'ordre des lettres est donc préservé.
  Par exemple, LCS(forestier, forêt)=fort, |LCS(forestier, forêt)|=4. On peut normaliser cette mesure comme suit (Ketkar et Youngblood 2010) :
  $$LCS_{normalise}(A,B) = \frac{|LCS(A,B)|^2}{(a \times b)} \in [0,1] \tag{4.1}$$
  où $a$ est la longueur de $A$ et $b$ la longueur de $B$. Plus ce score est élevé plus les chaînes sont similaires. Exemple : $LCS_{normalise}$(forestier, forêt)=16/45=0,35.
- Levenshtein : cette distance est définie comme le nombre minimum de modifications nécessaires pour transformer une chaîne en une autre. Les opérations autorisées sont l'insertion, la suppression et la substitution d'un seul caractère. Le coût de chaque opération est égal à 1. Par exemple, Levenshtein(forestier, forêt)=5. On peut normaliser

3. Subsequence : les lettres de la sous-séquence sont dans le même ordre que dans la chaine complète, substring : les lettres de la sous-chaîne sont consécutives et dans le même ordre que dans la chaine complète.

cette distance si on la divise par la chaîne la plus longue. Moins ce score est élevé plus les chaines sont similaires.
Exemple : $Lev_{normalise}$(forestier, forêt)=5/9=0,55.
On peut prendre (1-$Lev_{normalise}$) pour mesurer la similarité entre deux chaînes.

– DICE : cette mesure est égale au nombre de bi-grammes de lettres en commun entre deux mots divisé par le nombre total des bi-grammes dans les deux mots.

$$DICE(X,Y) = \frac{2|bi\text{-}grammes(X) \cap bi\text{-}grammes(Y)|}{|bi\text{-}grammes(X) + |bi\text{-}grammes(Y)|} \quad (4.2)$$

Il existe bien évidemment d'autres mesures de similarité : XDICE (emploie des bi-grammes étendus, c'est-à-dire des tri-grammes sans la lettre au milieu), SOUNDEX (code deux chaînes d'une prononciation similaire par la même chaîne, les consonnes à prononciation similaire ont le même code, par exemple, les lettres $f$ et $v$ en français sont codées par 9), toutes ces mesures et d'autres sont présentées dans Frunza et Inkpen (2009).

Les mesures de similarité sont souvent utilisées dans la tâche d'identification des cognats. Par exemple, Hauer et Kondrak (2011) et Frunza et Inkpen (2009) utilisent ou combinent plusieurs mesures de similarité telles que Levenshtein, LCS, SOUNDEX, Longest Prefix (le préfixe le plus long entre deux chaînes), etc. Les scores obtenus entre chaque couple de mots seront ensuite utilisés comme traits par un algorithme d'apprentissage qui les classifie comme cognats ou non. Afin d'identifier les mots qui sont des faux cognats, Frunza et Inkpen (2009) utilisent un corpus parallèle qui sert à désambiguïser les sens des mots. Les mesures de similarité ont été également utilisées par Cartoni (2008) pour relier un adjectif relationnel avec son nom de base. Cependant, Cartoni (2008) exige qu'un adjectif et un nom aient une similarité de lettres très importante afin de les relier automatiquement.

Nous présentons maintenant nos différentes méthodes d'alignement d'un adjectif relationnel avec son nom de base.

### 4.4.1 Alignement adjectif-nom par mesures de similarité de lettres

D'abord, nous essayons de relier un adjectif avec un nom en n'utilisant que des mesures de similarité.

Nous combinons les deux similarités : $similarity_{LCS}(A,B)$ et $similarity_{Lev}(A,B)$ (voir les équations 4.4 et 4.5), en prenant leur moyenne géométrique afin d'avoir un seul score $\in [0,1]$ entre un adjectif et un nom :

$$similarity_{lettres}(A,B) = (similarity_{LCS}(A,B) + similarity_{Lev}(A,B))/2 \quad (4.3)$$

$$similarity_{LCS}(A,B) = |LCS(A,B)|^2/(a \times b) \quad (4.4)$$

$$similarity_{Lev}(A,B) = \begin{cases} 1 - (Levenshtein(A,B)/a) & \text{si } a >= b \\ 1 - (Levenshtein(A,B)/b) & \text{autrement} \end{cases} \quad (4.5)$$

où $similarity_{LCS}(A, B)$ et $similarity_{Lev}(A, B) \in [0, 1]$, $a$ et $b$ sont les longueurs des chaînes $A$ et $B$ respectivement, les mesures de Levenshtein et LCS ont été présentées précédemment en section 4.4. Dans le calcul du score de Levenshtein, chaque opération a un coût égal à 1. Cependant, nous donnons une pénalité moins élevée à la substitution de deux lettres qui sont proches phonétiquement. Pour la langue française, nous nous inspirons de Dubois et Dubois-Charlier (1999) pour définir ces substitutions, puisque des adaptations générales de la langue française y sont définies. Par exemple, nous définissons la pénalité des substitutions d'une lettre par une autre lettre à 0,5 si les deux appartiennent au même groupe, les groupes sont : {e,é,è}, {a,â}, {f,v}, {s,x,z}, {c,q}. C'est-à-dire que, par exemple, la substitution de la lettre $f$ par la lettre $v$ est pénalisée à hauteur de 0,5.

Si un adjectif peut avoir un emploi nominal (considéré comme un substantif dans le dictionnaire), nous l'alignons avec lui-même (ex. clinique, esthétique, etc). De plus, nous supposons qu'un adjectif relationnel commence par la même lettre que son nom de base. En effet, nous avons trouvé, en examinant une liste de 200 adjectifs, que cette hypothèse est vérifiée dans 97 % des cas (l'adjectif *spatial* et son nom de base *espace* forment un contre-exemple).

Nous considérons qu'un adjectif est relié à un nom si le score entre les deux est supérieur ou égal à un certain seuil (le suffixe relationnel de l'adjectif est supprimé lors de la comparaison).

Plus le seuil de similarité de lettres est élevé plus le nombre d'adjectifs qui peuvent être reliés à des noms est faible. La mesure LCS favorise les noms les plus longs quand un score est calculé entre un adjectif et un nom. Par exemple, selon LCS, le nom *notion* est plus proche de *nominal* que de *nom* : |LCS(nomin, notion)|=4, |LCS(nomin, nom)|=3. Alors que les deux noms ont le même score avec *nominal* selon Levenshtein : Levenshtein(nomin, notion)=2, Levenshtein(nomin, nom)=2. De plus, si nous exigeons une similarité très importante entre un adjectif et un nom, nous pourrons perdre de nombreux alignements corrects (ex. axillaire/aisselle, germe/germinal, etc.). Pour cela, il faut choisir un seuil de similarité qui ne soit pas trop important afin de permettre à d'autres méthodes de filtrage de mieux ordonner les alignements obtenus par les mesures de similarité.

### 4.4.2 Alignement adjectif-nom par mesure de similarité contextuelle

Dans de nombreux cas, la similarité de lettres seule ne suffit pas pour trouver le nom auquel un adjectif est relié. Par exemple, comment peut-on dire que l'adjectif *sérique* est dérivé de *sérum* et non pas de *série* ?

Nous essayons donc de modifier le score entre un adjectif relationnel et un nom si le score de similarité de lettres entre le nom et l'adjectif est supérieur à un certain seuil. Le score est modifié en fonction des paraphrases monolingues dans lesquelles les deux mots apparaissent. Pour un adjectif $A$ et un nom $N$, nous cherchons une paraphrase dans le corpus

de la forme [X + A] : : [X + PREP + DET ? + N][4], où X est un nom tête. Comme par exemple, cancer pulmonaire : : cancer du poumon.

Afin de calculer un score de similarité contextuelle entre $A$ et $N$, nous représentons chacun par un vecteur où les attributs sont les noms têtes qui apparaissent avec $A$ ou $N$ (voir figure 4.2). Le score d'un attribut dans le vecteur d'un nom $N$ est calculé à l'aide de la mesure d'association $IM$ (Information Mutuelle, (Fano 1961)) entre $N$ et le nom tête $X$ :

$$IM(N, X) = \log_2 \frac{a}{(a+b)(a+c)} \tag{4.6}$$

- $a$ est le nombre d'occurrences de $N$ et $X$ ensemble.
- $b$ est le nombre d'occurrences de $N$ avec tous les autres nom têtes $\neq X$.
- $c$ est le nombre d'occurrences de $X$ avec tous les autres noms $\neq N$.

Le score de chaque attribut dans un vecteur d'adjectif $A$ est calculé de la même manière :

$$IM'(A, X) = \log_2 \frac{a'}{(a'+b')(a'+c')} \tag{4.7}$$

- $a'$ est le nombre d'occurrences de $A$ et $X$ ensemble.
- $b'$ est le nombre d'occurrences de $A$ avec tous les autres nom têtes $\neq X$.
- $c'$ est le nombre d'occurrences de $X$ avec tous les autres adjectifs $\neq A$.

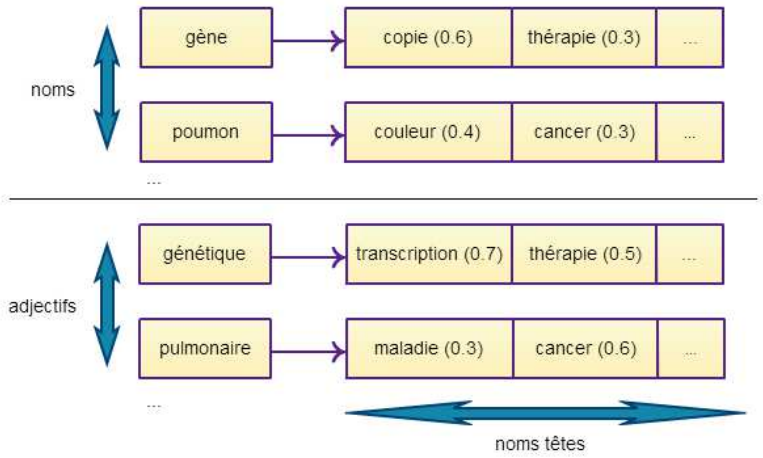


FIGURE 4.2 – *Vecteurs des adjectifs relationnels et des noms*

Les mesures d'association peuvent être utilisées pour mesurer quantitativement si deux mots ($m_1$ et $m_2$) ont plus tendance à apparaître ensemble que séparément (Prochasson 2010). Ces mesures combinent la fréquence de cooccurrences de deux mots avec les fréquences d'apparitions de l'un sans l'autre. Elles peuvent aussi exploiter le nombre de mots total dans le corpus ou tout autre indice pour bien caractériser l'association entre deux mots. Il existe plusieurs mesures d'association proposées dans la littérature et largement utilisées (quelques mesures sont présentées dans Streiter

4. Nous pouvons aussi inclure d'autres variantes, comme par exemple les formes [X + $A_1$ + A] (ex. région ganglionnaire axillaire) et [X + $A_1$ + PREP + DET ? + N] (ex. balayage lent de l'aisselle).

et al. (2003)). Nous avons choisi la mesure d'association $IM$ parce que cette mesure n'utilise pas le nombre d'occurrences de tous les couples de mots où, ni le mot $m_1$, ni le mot $m_2$ apparaissent ; parce que les mesures utilisant cette information ont donné de moins bons résultats dans nos expériences.

Après avoir calculé les valeurs des attributs dans les vecteurs, nous calculons un score entre les deux vecteurs (nom N et adjectif A) en utilisant le cosinus.

$$similarity_{paraphrases}(A, N) = \cos(A, N) = \frac{\sum_{i=1}^{n} IM.IM'}{\sum_{i=1}^{n} IM^2 . \sum_{i=1}^{n} IM'^2} \quad (4.8)$$

Où $n$ est le nombre de noms têtes communs entre $A$ et $N$.

### 4.4.3 Combinaison des mesures de similarité de lettres et de similarité contextuelle

Pour un adjectif $A$ et un nom $N$, nous calculons leur score final en combinant leur similarité de lettres selon l'équation 4.3 et la similarité contextuelle selon l'équation 4.8 par une interpolation linaire, comme suit :

$$score(A, N) = \alpha \times similarity_{lettres}(A, N) + \beta \times similarity_{paraphrases}(A, N) \quad (4.9)$$

où $\alpha$+$\beta$=1.

Le score de similarité contextuelle doit avoir un poids moins important que le poids de la similarité de lettres afin d'obtenir une meilleure qualité d'alignement.

Un adjectif $A$ sera relié au nom avec lequel il a le score le plus élevé. C'est-à-dire que nous alignons un adjectif avec un seul nom.

### 4.4.4 Alignement adjectif-nom en utilisant des racines supplétives

Certains adjectifs contiennent des racines supplétives et dans certains cas il n'est pas possible de les relier à leurs noms en s'appuyant sur la similarité de lettres. Cela est surtout vrai quand la modification du nom de base par la racine supplétive est importante (Ex. hépatique/foie, médullaire/moelle).

Nous considérons qu'un adjectif $A$ qui commence par une séquence de lettres identique à une racine supplétive est relationnel s'il remplit une des conditions suivantes :

– sa forme peut être identifiée comme étant : [racine + suffixe] (où *racine* est une des racines supplétives dans une liste des racines alignées avec des noms ($L_{racines-noms}$)[5], et *suffixe* est un des suffixes relationnels dans une liste de suffixes relationnels. Par exemple, l'adjectif *pulmonaire* peut être décomposé en *pulmon* (une racine) et *aire* (un suffixe).

5. Ex. hépat/foie, pulmon/poumon, médull/moelle, etc.

– il construit avec le nom $N$ associé à la racine supplétive (selon $L_{racines-noms}$) au moins une paraphrase de la forme [X + A] : : [X + PREP + DET ? + N], où $X$ est un nom tête. Par exemple, biopsie de moelle : : biopsie médullaire.

### 4.4.5 Combinaison des méthodes d'alignement

L'approche décrite en section 4.4.4 et celle présentée dans la section 4.4.3 sont combinées comme suit : pour un adjectif $A$, nous vérifions s'il peut être relié à un nom à l'aide des racines supplétives, sinon, un score sera calculé (selon l'équation 4.9) entre $A$ et chaque nom dans le dictionnaire (qui existe dans le corpus).

De plus, si $A$ (ex. oncogénique) commence par un préfixe ou une ou plusieurs racines supplétives, et s'il peut s'écrire de la forme : [(racine | préfixe)$_+$ + A′] (ex. oncogénique), où $A'$ (ex. génique) est un adjectif dans le corpus : nous relions $A'$ à un nom $N'$ (ex. gène). Ensuite, nous cherchons le nom $n$ = [(racine | préfixe)$_+$ + N′] (ex. oncogène) dans le corpus, si ceci est trouvé, le nom de base avec lequel $A$ est aligné sera $n$.

Nous supposons aussi que deux adjectifs qui partagent la même base (ex. sérique/séreux (sér), soigneux/soigné (soign), cellulaire/celluleux (cellul), etc.), doivent être alignés avec le même nom de base, sinon nous considérons que les alignements sont mauvais et nous les supprimons de la liste.

## 4.5 Traduction des termes [N + AdjR] en utilisant des alignements adjectif-nom

Nous utilisons les alignements adjectif-nom produits à l'aide de l'approche d'alignement d'un adjectif avec un nom, présentée dans la section précédente, dans la tâche de traduction compositionnelle de termes complexes de la forme [N + AdjR]. L'approche compositionnelle pour la traduction que nous suivons est une approche de base qui consiste à traduire un terme complexe mot-à-mot.

### 4.5.1 Approche

Pour chaque terme $t_s$=[N + AdjR], nous remplaçons $AdjR$ par le nom $N'$ avec lequel il a été aligné. Ce remplacement donne un nouveau terme complexe $t'_s$=[N + PREP + DET ? + N′], nous supposons que sa traduction est équivalente à celle de $t_s$. Nous suivons la méthode présentée sous le nom algorithme 4.3.

**Données** : $L_{[N+A]_{fr}}$ (termes français de la forme [N + A]), $L_{[A+N]_{en}}$ (termes anglais de la forme [A + N]), $L_{[N+N]_{en}}$ (termes anglais de la forme [N + N]), $L_{alignements}$ (alignements adjectif-nom), $Dico_{fr-en}$ (dictionnaire bilingue français-anglais) ;
**Résultat** : Liste de traductions;
**début**
  **pour** *chaque terme de la forme [N + A]* $\in L_{[N+A]_{fr}}$ *(ex. FR concentration plasmatique)* **faire**
    $N_{en} \longleftarrow$ traduire $N$ par $Dico_{fr-en}$ (ex. $N_{en} \longleftarrow$ EN concentration);
    $A_{en} \longleftarrow$ traduire $A$ par $Dico_{fr-en}$ (ex. $A_{en} \longleftarrow$ EN plasmatic);
    **si** *[*$A_{en}$ + $N_{en}$*] (ex. plasmatic concentration) existe dans* $L_{[A+N]_{en}}$ **alors**
      [$A_{en}$ + $N_{en}$] est la traduction de [N + A];
    **sinon**
      Prendre le nom $N'$ (ex. FR plasma) avec lequel l'adjectif $A$ (ex. FR plasmatique) est aligné de $L_{alignements}$ ;
      $N'_{en} \longleftarrow$ traduire $N'$ par $Dico_{fr-en}$ (ex. $N_{en} \longleftarrow$ EN plasma);
      **si** *[*$N'_{en}$ + $N_{en}$*] (ex. EN plasma concentration) existe dans* $L_{[N+N]_{en}}$ **alors**
        [$N'_{en}$ + $N_{en}$] est la traduction de [N + A];

**Algorithme 4.3 :** Traduction des termes en utilisant les alignements adjectif-nom

## 4.6 Evaluation

Dans cette section, nous évaluons les approches que nous proposons pour (a) extraire des adjectifs relationnels ; (b) aligner un adjectif extrait avec son nom de base ; (c) traduire un adjectif relationnel dans une structure [N + AdjR] en le remplaçant par le nom avec lequel il est aligné.

### 4.6.1 Ressources utilisées

Nous disposons des ressources suivantes :
- termes complexes extraits à l'aide de TermSuite (outil présenté en section 2.6 du chapitre 2) du corpus comparable cancer du sein français-anglais ($C_{cds_{fr}}$ et $C_{cds_{en}}$) présenté en section 1.4 du chapitre 1. Nous extrayons des syntagmes selon les motifs dont nous avons besoin, comme suit :
  - 12 991 syntagmes français ($L_{[N+A]_{fr}}$) du motif [N + A], et 11 941 syntagmes anglais ($L_{[A+N]_{en}}$) du motif [A + N].
  - 12 954 syntagmes français ($L_{[N+N]_{fr}}$) du motif [N + PREP + DET ? + N], et 10 069 syntagmes anglais ($L_{[N+N]_{en}}$) du motif [N + N].

- liste de préfixes $L_{prefixes}$ en français qui n'acceptent qu'une base adjectivale relationnelle ou déverbale, cette liste a été établie à l'aide des travaux de Cartoni (2008) (voir P11 (préfixes) dans la table 4.1).
- liste de 15 suffixes relationnels en français $L_{suffRel}$ (voir P7 (suffixes) dans la table 4.1).
- deux listes d'adjectifs français extraites automatiquement du corpus $C_{cds_{fr}}$ :
  - $LAdjR_{Classes}$ : cette liste comprend 361 adjectifs extraits automatiquement du corpus. Elle correspond à l'ensemble des listes extraites en suivant l'algorithme 4.1.
  - $LAdjR_{Base}$ : cette liste contient tous les adjectifs extraits à partir de la propriété P7 (suffixes) et qui composent au moins un terme français de la forme [N + A]. Cette liste contient 1 346 adjectifs, elle est considérée comme la liste de base et les résultats de l'alignement adjectif-nom sur cette liste seront comparés avec ceux obtenus sur la liste $LAdjR_{Classes}$.
- dictionnaire bilingue français-anglais ($Dico_{fr-an}$) de 145 542 entrées de mots simples (présenté en section 1.4.2 du chapitre 1).
- liste de 66 racines supplétives françaises ($L_{racines-noms}$) alignées avec des noms communs (ex. hépat/fois, pulmon/poumon, etc.) (Cottez 1982).
- liste de 100 racines $L_{racines}$ extraite automatiquement du $C_{cds_{fr}}$ en appliquant l'algorithme 4.2 présenté en section 4.3.2.

### 4.6.2 Résultats de l'extraction automatique des adjectifs relationnels

En appliquant l'algorithme d'extraction des adjectifs présenté en section 4.3 sur $C_{cds_{fr}}$, nous obtenons quatre listes d'adjectifs. Un adjectif extrait appartient à une ou plusieurs classes d'adjectifs : qualificative, relationnelle et composée. Par exemple, l'adjectif *sérologique* est composé et relationnel, car il peut être relié à *sérologie* et il se compose de deux éléments : *séro* et *logique*. Les adjectifs de la classe *composée* ont des emplois non-qualificatifs, mais dans certains cas, ils ne peuvent pas être reliés à un seul substantif, mais à un syntagme. Par exemple, *unilatéral* ('un seul côté') ou *infraclinique* ('se dit d'un trouble ou d'une maladie qui ne provoque pas de manifestation décelable à l'examen') ont été formés par préfixation.

Les listes extraites sont présentées dans la table 4.2. Nous appelons l'ensemble de ces listes par $LAdjR_{Classes}$. Cette liste comprend 361 adjectifs.

Les adjectifs ont été classés manuellement et à l'aide de l'outil Dérif (Namer 2003). Nous remarquons que 198 adjectifs dans $LAdjR_{Classes}$ peuvent être classifiés comme relationnels et qu'il y a beaucoup d'adjectifs composés qui ne sont ni relationnels ni qualificatifs (ex. carcino-embryonnaire, infraradiologique, pluridisciplinaire, périmammaire, etc.).

La liste $LAdjR_{Classes}$ se compose de 54 % d'adjectifs relationnels. En effet, beaucoup d'adjectifs composés qui ne peuvent pas être reliés à des noms se trouvent dans la liste $Liste_{AdjR}$-1. La liste $Liste_{AdjR}$-2 se compose de 93 % d'adjectifs relationnels. Les listes $Liste_{AdjR}$-2, $Liste_{AdjR}$-3 et $Liste_{AdjR}$-4 se composent de 82 % d'adjectifs relationnels.

Pour avoir une idée du rappel, nous utilisons l'outil Dérif pour aligner les adjectifs de la liste $LAdjR_{Base}$ avec des noms. Dérif est capable d'aligner 558 adjectifs avec des noms par la relation « en rapport avec ». Nous appelons cette liste par $Liste_{Derif}$. Nous trouvons que la liste $LAdjR_{Classes}$ couvre 141 adjectifs de $Liste_{Derif}$. Cependant, 57 des adjectifs que nous avons jugés comme relationnels dans $LAdjR_{Classes}$ n'ont pas pu être alignés par Dérif. De plus, il existe des adjectifs dénominaux mais non relationnels dans $Liste_{Derif}$ (ex. original/origine, critique/crise, etc.).

| | **Nb. d'adjectifs** | **Nb. classe qualificative** | **Nb. classe relationnelle** | **Nb. classe composée** |
|---|---|---|---|---|
| $Liste_{AdjR}$-1 | 154 | 0 | 28 | 153 |
| $Liste_{AdjR}$-2 | 103 | 8 | 96 | 19 |
| $Liste_{AdjR}$-3 | 47 | 3 | 34 | 18 |
| $Liste_{AdjR}$-4 | 57 | 6 | 40 | 27 |
| Total ($LAdjR_{Classes}$) | 361 | 17 | 198 | 217 |

TABLE 4.2 – *Listes d'adjectifs extraits par l'algorithme 4.1*

Les listes d'adjectifs extraites seront utilisées par la méthode de l'alignement d'un adjectif relationnel avec un nom.

### 4.6.3 Résultats de l'alignement adjectif-nom sur les listes d'adjectifs

Nous appliquons la méthode d'alignement que nous avons proposée en section 4.4.5 sur les listes des adjectifs extraits automatiquement ($LAdjR_{Classes}$ et $LAdjR_{Base}$). Nous fixons par observation les poids des deux similarités dans l'équation 4.9 : $\alpha$=0,70 et $\beta$=0,30. Un adjectif et un nom doivent avoir une similarité minimale de $similarity_{Lev}$ à 0,6 et une similarité minimale de $similarity_{LCS}$ à 0,7.

Ainsi, 157 adjectifs de la liste $LAdjR_{Classes}$ (parmi 361) ont été alignés avec une précision de 89,8 %. De la liste $LAdjR_{Base}$, 582 adjectifs (parmi 1 346) ont été alignés avec une précision de 84,53 %. Nous avons évalué les alignements manuellement et à l'aide de l'outil Dérif (Namer 2003). Nous considérons qu'un alignement est correct si l'adjectif a été aligné avec lui-même ou avec son nom de base. La liste $LAdjR_{Base}$ contient plus d'adjectifs non-relationnels et du bruit (des mots non-français) que $LAdjR_{Classes}$, ce qui explique le taux plus élevé des mauvais alignements.

Les mauvais alignements ont plusieurs causes, par exemple, nous supposons que le nom de base d'un adjectif doit être présent dans le corpus alors que ce n'est pas toujours le cas. Cependant, cette hypothèse est aussi importante pour filtrer des mauvais alignements.

Nous calculons également le rappel, qui est le nombre d'adjectifs alignés avec des noms divisé par le nombre d'adjectifs dans la liste. Cependant, il faut noter qu'il y a de nombreux adjectifs dans $LAdjR_{Base}$ et $LAdjR_{Classes}$ qui ne peuvent pas être reliés à des noms. Par exemple, les adjectifs composés sont parfois reliés à des phrases comme nous l'avons

déjà mentionné. Nous résumons les résultats de l'alignement dans la table 4.3.

| | **Nb. d'alignements Adj-Nom** | **Précision** | **Rappel** |
|---|---|---|---|
| $\text{LAdjR}_{Classes}$ | 157 | 89,80 % | 43,49 % |
| $\text{LAdjR}_{Bases}$ | 582 | 84,53 % | 43,23 % |

TABLE 4.3 – *Résultats des méthodes d'alignement adjectif-nom sur $LAdjR_{Classes}$ et $LAdjR_{Base}$*

Nous présentons, dans la section suivante, les résultats de la traduction des termes de la forme [N + AdjR] en utilisant les alignements adjectif-nom présentés dans cette section.

### 4.6.4 Résultats de la traduction des termes [N + AdjR]

Une méthode compositionnelle de base qui consiste à traduire des termes français de la forme [N + A] en termes anglais de la forme [A + N] nous a permis de traduire 2 039 termes (ex. FR *adaptation psychologique* traduit par EN *psychological adaptation*) dont les adjectifs sont issus de la liste $LAdjR_{Base}$. Elle nous a aussi permis de traduire 574 termes dont les adjectifs sont issus de la liste $LAdjR_{Classes}$. Cette méthode a donné une précision de 79,5 % sur une liste de 200 termes traduits qui a été examinée manuellement. Les termes français [N + A] non-traduits par la méthode précédente sont ensuite traduits par paraphrase ex. les termes *prélèvement tumoral*, *catégorie cellulaire*, *pathologie pulmonaire* n'ont pas été traduits par la méthode compositionnelle de base.

Nous suivons l'algorithme 4.3 afin d'évaluer l'impact des alignements adjectif-nom sur la traduction des termes [N + A], voir la table 4.4. Nous utilisons les 157 alignements adjectif-nom obtenus de la liste $LAdjR_{Classes}$ et nous trouvons que 42 alignements adjectif-nom de cette liste ont aidé à traduire 172 termes [N + A] distincts avec une précision de 91,86 %. En appliquant l'algorithme 4.3 sur les 582 alignements adjectif-nom obtenus de $LAdjR_{Base}$, nous trouvons que 92 de ces alignements ont participé à traduire 250 termes distincts avec une précision de 86 %. Les traductions ont été vérifiées à l'aide du dictionnaire rédactionnel Linguee [6] et de la banque de données Termium [7]. La précision des traductions est égale au nombre de termes distincts qui ont au moins une traduction correcte parmi les 5 premières traductions proposées divisé par le nombre de termes distincts qui ont été traduits. Les traductions proposées ont été classées par leurs fréquences dans le corpus cible.

Les alignements des adjectifs qualificatifs (ex. originale/origine, formel/forme) avec des noms ont donné de mauvaises traductions. Des adjectifs déverbaux qui peuvent être reliés à un nom, ont donné des bonnes et/ou des mauvaises traductions. Par exemple, l'adjectif *étudié* est dérivé du verbe *étudier*, il a été relié au nom *étude* par la méthode d'alignement

6. http ://www.linguee.fr/
7. http ://www.termiumplus.gc.ca/tpv2alpha/alpha-fra.html ?lang=fra

adjectif-nom. Cet alignement a donné de bonnes traductions (ex. *population étudiée* a été traduit par *study population*), ainsi que de mauvaises traductions (ex. *cellule étudiée* a été traduit par *study unit*).

Parfois nous pouvons trouver des mauvaises traductions malgré l'utilisation d'un alignement correct d'un adjectif relationnel avec un nom. Ces mauvaises traductions sont plutôt obtenues à cause des problèmes liés à la méthode compositionnelle et au corpus comparable. Par exemple, l'adjectif relationnel *génétique* a été relié au nom *gène*, cet alignement a participé à la traduction de *mutation génétique* par *gene transfer* (*gène* traduit par *gene*) tandis que la bonne traduction est *gene mutation*. Ainsi, la mauvaise traduction n'a pas été obtenue à cause de l'alignement de *génétique* avec *gène*, mais parce que soit *mutation* n'a pas été traduit par *mutation* dans le dictionnaire bilingue, soit *gene mutation* n'existe pas dans le corpus anglais.

| | **Nb. d'alignements Adj-Nom** | **Nb. de termes** [N + A] **traduits** | **Précision** |
|---|---|---|---|
| $\text{LAdjR}_{Classes}$ | 157 | 172 | 91,86 % |
| $\text{LAdjR}_{Bases}$ | 582 | 250 | 86,00 % |

TABLE 4.4 – *Résultats de la traduction en utilisant les alignements adjectif-nom*

## 4.7 ÉVALUATION AVEC UNE LISTE D'ADJECTIFS ALIGNÉS

Afin d'avoir une idée de l'utilité de l'alignement automatique d'un adjectif avec un nom pour la traduction des termes de la forme [N + A], nous construisons une liste ($Liste_{Derif}$) de 558 adjectifs à l'aide de l'outil Dérif. Cette liste contient les adjectifs dans la liste $LAdjR_{Base}$ qui ont pu être reliés à des noms à l'aide de Dérif (en exploitant la relation « en rapport avec »).

Les adjectifs dans cette liste ne sont pas forcément des adjectifs relationnels, mais ce sont plutôt des adjectifs dénominaux (ex. original/origine, critique/crise, etc.) qui peuvent être relationnels ou qualificatifs.

En appliquant la même approche de traduction suivie dans la section précédente : nous utilisons les alignements fournis par Dérif (pour la liste $Liste_{Derif}$) et nous trouvons que 74 alignements adjectif-nom (où le nom existe dans le corpus français) permettent d'aider à traduire par paraphrase 251 termes de la forme [N + A] distincts avec une précision de 83,26%.

Afin d'évaluer les différentes méthodes d'alignement d'un adjectif avec un nom que nous avons proposé dans la section 4.4, nous essayons d'aligner les adjectifs dans la liste $Liste_{Derif}$ automatiquement en utilisant les méthodes d'alignement que nous avons présentées : la similarité de lettres, la similarité contextuelle, les racines supplétives ainsi que des différentes combinaisons des méthodes d'alignement. Les résultats sont présentés dans la table 4.5. La précision et le rappel ont été calculés par rapport aux alignements obtenus par Dérif.

| | **Nb. d'alignements Adj-Nom** | **Précision** | **Rappel** |
|---|---|---|---|
| similarité de lettres (1) | 333 | 76,57% | 59,67 % |
| similarité contextuelle (2) avec condition (similarité$_{lettres}$>0.65) | 76 | 77,63% | 13,62 % |
| racines supplétives (3) | 12 | 100% | 2,15% |
| (1) + (2) | 337 | 77,15% | 60,39 % |
| (1) + (3) | 338 | 77,21 % | 60,57 % |
| (1) + (2) + (3) | 341 | 77,71 % | 61,11 % |

TABLE 4.5 – *Résultats des méthodes d'alignement des adjectifs-noms avec la liste* $L_{Derif}$

| | **Nb. d'alignements Adj-Nom** | **Nb. termes traduits** | **Précision** |
|---|---|---|---|
| similarité de lettres (1) | 333 | 141 | 86,52% |
| similarité contextuelle (2) avec condition (similarité$_{lettres}$>0.65) | 76 | 163 | 92,63% |
| racines supplétives (3) | 12 | 32 | 93,75% |
| (1) + (2) | 337 | 158 | 86,70% |
| (1) + (3) | 338 | 164 | 87,19% |
| (1) + (2) + (3) | 341 | 177 | 87,00% |

TABLE 4.6 – *Résultats de la traduction par paraphrase en utilisant les alignements adjectif-nom sur la liste* $L_{Derif}$

Ensuite, nous évaluons l'utilisation des alignements obtenus par chaque méthode d'alignement d'un adjectif avec un nom pour la traduction des termes de la forme [N + A]. La table 4.6 présente les résultats obtenus.

Nous pouvons remarquer à partir de ces résultats que la combinaison de trois méthodes d'alignement que nous avons proposées donne les meilleurs résultats en termes de nombre de traductions obtenues. Nous remarquons aussi que les alignements de la similarité contextuelle aident à obtenir un nombre plus important de traductions que le nombre de traductions obtenu par les alignements de la similarité de lettres.

Nous pouvons aussi constater que seulement 177 termes ont été traduits par la combinaison de méthodes d'alignement sur la liste $L_{Derif}$ et que 251 termes ont pu être traduits en utilisant les alignements fournis par Dérif. Cependant, en alignant tous les adjectifs de la liste LAdjR$_{Bases}$ avec la combinaison de méthodes d'alignement, nous avons pu trouver 250 termes avec une précision de 86% (voir la table 4.4). C'est-à-dire que des termes traduits en plus ont pu être obtenus et qu'ils ne sont pas inclus dans les 251 termes traduits à l'aide des alignements obtenus par Dérif.

## 4.8 Synthèse et Discussion

Nous avons développé une approche qui exploite plusieurs propriétés des adjectifs relationnels pour les identifier en se basant sur un corpus monolingue. Nous avons extrait par cette approche une liste d'adjectifs $LAdjR_{Classes}$. Une autre liste d'adjectifs $LAdjR_{Base}$ a été extraite en utilisant une liste de suffixes relationnels. Nous avons trouvé que la liste $LAdjR_{Classes}$ contient très peu d'adjectifs qualificatifs et moins de bruit que la liste $LAdjR_{Base}$.

Ensuite, nous avons développé une méthode afin d'aligner les adjectifs relationnels extraits avec leurs noms de base à partir d'un corpus monolingue. Nous nous sommes appuyés sur la similarité de lettres, la similarité contextuelle et des racines gréco-latines afin de relier un adjectif à un nom. Nous avons appliqué la méthode d'alignement sur les deux listes $LAdjR_{Base}$ et $LAdjR_{Classes}$, et nous avons acquis des couples d'adjectif-nom avec une précision supérieure à 84 %.

Enfin, nous avons exploité les alignements adjectif-nom obtenus pour traduire compositionnellement des termes de la forme [N + AdjR]. Le qualité des alignements adjectif-nom obtenus à partir de $LAdjR_{Classes}$ et des traductions des termes [N + AdjR] obtenues en utilisant ces alignements, a été meilleure que celle des alignements et des traductions obtenues avec $LAdjR_{Base}$. Par contre, nous obtenons plus d'alignements avec $LAdjR_{Base}$ et donc plus de traductions par rapport à l'utilisation de $LAdjR_{Classes}$. Les mauvais alignements adjectif-nom n'ont pas beaucoup influencé la précision des traductions de ces termes qui est de 86 % en utilisant $LAdjR_{Base}$. La traduction compositionnelle permet, en effet, de filtrer les mauvais alignements adjectif-nom pour la traduction des termes [N + AdjR]. Cela est dû au fait que même si un adjectif est aligné avec un nom dont il n'est pas dérivé, il faut que la traduction générée en utilisant cet alignement soit présente dans le corpus pour qu'elle soit considérée comme correcte.

## 4.9 Conclusion

Dans ce chapitre, nous nous sommes intéressés à l'identification des adjectifs relationnels et à l'alignement de ces adjectifs avec leurs noms de base. Nous avons également essayé de traduire des termes qui se composent d'un nom et d'un adjectif relationnel [N + AdjR] en remplaçant $AdjR$ par son nom de base.

Nous nous sommes concentrés sur la traduction des termes [N + AdjR] pour le couple de langues français-anglais. Le principe de traduction par paraphrase de ces termes pour d'autres couples de langues devrait être étudié pour en démontrer la généricité. De plus, la méthode d'extraction des adjectifs relationnels que nous avons proposée devrait être plus efficace pour les domaines spécialisés (ex. les domaines médicaux) qui emploient beaucoup de racines gréco-latines.

Le principe de la traduction compositionnelle a été appliqué dans ce

chapitre ainsi que dans le chapitre précédent sur des termes ayant une propriété compositionnelle. Les traductions obtenues par la méthode compositionnelle sont de bonne qualité et elles sont donc utiles pour les traducteurs automatiques ou humains. Afin d'étudier la possibilité d'obtenir des traductions dont la qualité est proche de celle des traductions obtenues de manière compositionnelle, nous nous intéressons dans le chapitre suivant au reclassement des traductions candidates proposées dans un lexique bilingue obtenu par une approche distributionnelle.

# 5 Reclassement des traductions candidates à partir d'un corpus comparable

Sommaire



Les lexiques bilingues extraits avec une approche distributionnelle, à partir des corpus comparables spécialisés, fournissent pour un terme une liste de traductions candidates ordonnées. Nous proposons, dans ce chapitre, une technique pour ré-ordonner ces traductions candidates afin d'améliorer la qualité d'un lexique. Nous supposons qu'un terme et sa traduction apparaissent dans des phrases comparables qui peuvent être extraites à partir des corpus comparables spécialisés. Pour un terme source et une liste de traductions candidates, nous proposons une méthode permettant d'identifier et d'aligner des phrases privilégiées sources et phrases cibles qui contiennent le terme et ses traductions candidates. Nous présentons les expériences réalisées avec trois paires de langues (français-anglais, français-espagnol et français-allemand) et deux corpus spécialisés. Les résultats obtenus montrent que notre approche améliore la précision au top 1, top 5 et top 10 d'un lexique bilingue spécialisé et donne donc des résultats plus adaptés aux besoins des utilisateurs.

## 5.1 Introduction

Nous avons introduit dans le chapitre 2 des approches distributionnelles qui s'intéressent à l'extraction de lexiques bilingues à partir de corpus comparables. Nous avons présenté le travail de Rapp (1995), qui suppose que si un mot $A$ cooccurre fréquemment avec un autre mot $B$ dans une langue, la traduction de $A$ et la traduction de $B$ devraient cooccurrer fréquemment dans une autre langue. Les approches qui sont développées en s'appuyant sur le travail de Rapp (1995) sont toutes basées sur l'hypothèse qu'une paire de traductions partage un contexte similaire dans un corpus comparable.

Nous avons discuté aussi dans le chapitre 2 les résultats obtenus à partir des approches distributionnelles. Nous rappelons que ces résultats varient en fonction de nombreux paramètres. Par exemple, un des paramètres qui a une influence sur la performance des approches distributionnelles est la façon dont le contexte d'un mot est défini. Le contexte d'un mot peut être défini de différentes manières : fenêtres, phrases, paragraphes, ou à partir des dépendances syntaxiques. En effet, le contexte d'un mot ($m$) a été souvent représenté par les mots dans des fenêtres centrées sur $m$ (Laroche et Langlais 2010), ces fenêtres sont souvent de petite taille (par exemple une fenêtre de taille 3 est utilisée par Rapp (1999)).

Les corpus comparables spécialisés ont été utilisés pour l'extraction de termes bilingues. Ces corpus sont de taille modeste puisque des grands corpus spécialisés ne sont pas disponibles pour de nombreux domaines (Morin et al. 2007; 2008). Les approches distributionnelles donnent de meilleures performances pour de grands corpus comparables. Cependant, ces performances sont moins bonnes pour des corpus comparables spécialisés (Chiao et Zweigenbaum 2002, Hazem et Morin 2013a).

Le but du travail que nous présentons dans ce chapitre est d'améliorer la qualité d'un lexique bilingue spécialisé. En partant d'une liste de traductions candidates ordonnées (fournie par une approche distributionnelle standard) pour un terme donné : nous visons à améliorer le classement des traductions correctes qui ne sont pas au premier rang dans la liste. Évidemment, plus le nombre de traductions candidates proposées pour un terme est élevé, plus la probabilité de trouver la ou les traduction(s) correcte(s) est élevée. Par exemple, Rapp (1999) obtient une précision de 72 % lorsque seule la première traduction candidate est retenue pour chaque mot source, et une précision de 89 % lorsque les 10 premières traductions candidates sont retenues pour chaque mot source.

Nous proposons de ré-ordonner les meilleurs traductions candidates fournies par une approche distributionnelle afin d'améliorer la précision au top 1, top 5 et top 10 d'un lexique bilingue. Nous faisons l'hypothèse qu'un terme source et sa traduction correcte apparaissent dans des phrases comparables. Selon nous, deux phrases sont comparables si elles partagent des données parallèles (ex. des mots, des segments alignés, etc.). Les phrases comparables sont utilisées pour ré-ordonner les traductions candidates d'un terme source. De plus, les phrases comparables qui contiennent un terme et sa traduction sont prometteuses car elles peuvent

être des exemples utiles à un utilisateur ou à un traducteur humain pour vérifier une paire de traductions.

Ce chapitre décrit principalement une approche qui consiste à : (a) extraire des phrases pour un terme et sa traduction candidate ; (b) trouver des phrases comparables pour une paire de traductions ; (c) donner un score pour une paire de traductions afin de ré-ordonner les traductions candidates dans un lexique bilingue extrait automatiquement.

## 5.2 Hypothèses

Certains contextes dans lesquels un terme apparaît sont plus intéressants et plus informatifs que d'autres contextes. La table 5.1 présente un exemple de deux phrases dans lesquelles le terme EN *tumor* apparait. Ces phrases ont été extraites du corpus de cancer du sein présenté dans le chapitre 1 sous la section 1.4. La phrase (A) est considérée comme plus informative et plus représentative du contexte du terme EN *tumor* que la phrase (B) ; elle contient des termes qui sont fortement spécifiques au sujet du cancer du sein (ex. chemotherapy, histological) et reliés au terme EN *tumor*.

| | |
|---|---|
| **(A)** | **Chemotherapy** was also administered to patients with smaller primary tumors with **histological** grade 2 or 3 or with negative hormone receptors. |
| **(B)** | The size of any captured image corresponding to the tumor was estimated. |

Table 5.1 – *Phrases (A) et (B) contenant le terme EN tumor*

Plusieurs hypothèses sont émises :

– pour chaque terme, il existe un contexte plus informatif que les autres : c'est ce qu'on appelle le contexte privilégié du terme. Il peut être représenté par certains mots dans les phrases d'un corpus comparable.

– les mots du contexte privilégié d'un terme cooccurrent souvent avec ce terme et sont spécifiques au domaine du corpus.

– les phrases comparables contenant un terme et sa traduction candidate peuvent être trouvées à partir des contextes privilégiés. (Ces phrases sont ensuite utilisées pour ré-ordonner les traductions candidates d'un terme.)

– plus un terme et sa traduction candidate partagent des phrases comparables plus cette traduction est susceptible d'être correcte.

Nous nous basons donc sur la similarité entre le contexte privilégié d'un terme et celui de ses traductions candidates afin de trouver le meilleur classement. Nous supposons que les phrases comparables nous permettent d'acquérir des types de similarité lexicale qu'une méthode distributionnelle de base ne peut pas trouver. Par exemple, si nous trouvons des chaînes longues de mots alignés entre une phrase source (contenant un terme $t_s$) et une phrase cible (contenant une traduction candidate $t_c$ de $t_s$) nous pouvons mieux ré-ordonner $t_c$ dans la liste de traductions proposées pour $t_s$.

## 5.3 Approche

Après avoir appliqué une approche distributionnelle pour trouver des traductions candidates pour un terme à partir d'un corpus comparable (source-cible), nous calculons un score entre un terme source ($t_s$) et chaque traduction candidate ($t_c$) [1] comme suit :

D'abord, les $n$ phrases privilégiées qui représentent $t_s$ sont extraites du corpus source. De même, les $n$ phrases privilégiées sont extraites pour $t_c$ du corpus cible. Ensuite, chaque phrase extraite pour $t_s$ est alignée avec au maximum une phrase extraite pour $t_c$. Les phrases sont alignées à l'aide d'un dictionnaire bilingue. Enfin, un score est attribué à la paire de traductions ($t_s$,$t_c$) en fonction des scores des phrases alignées pour cette paire.

Cette approche est illustrée dans la figure 5.1. Nous combinons le score résultant de l'alignement des phrases avec le score obtenu de l'approche distributionnelle pour ($t_s$,$t_c$). Les scores combinés sont utilisés pour réordonner les traductions candidates d'un terme.

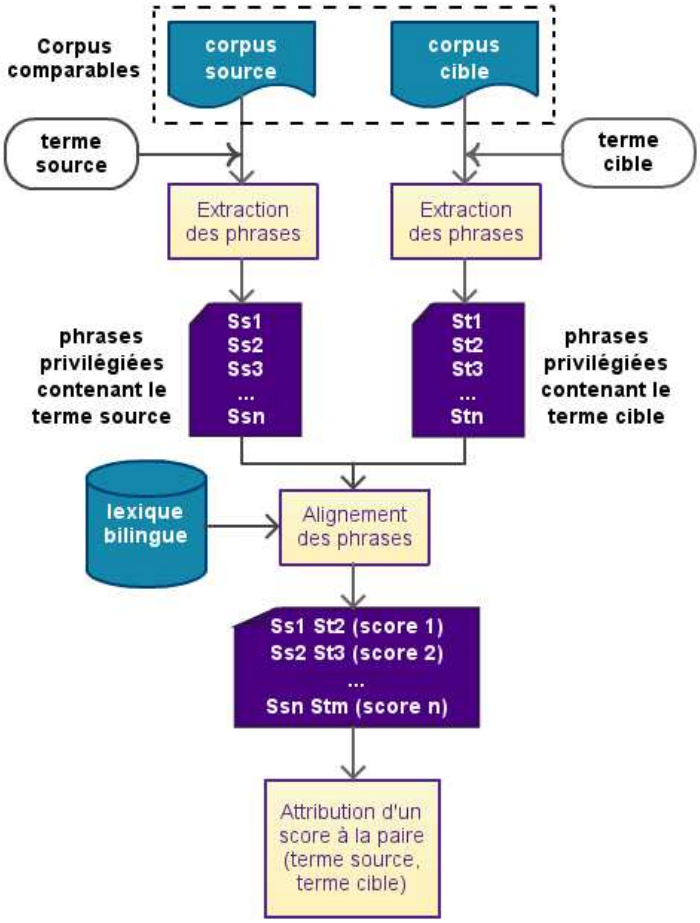


Figure 5.1 – *Approche pour attribuer un score à une paire de traductions (terme source et terme cible)*

L'extraction de phrases ou de segments parallèles à partir de corpus comparables a reçu l'attention de plusieurs chercheurs. En effet, de telles

1. Proposée par l'approche distributionnelle comme l'une des meilleures candidates.

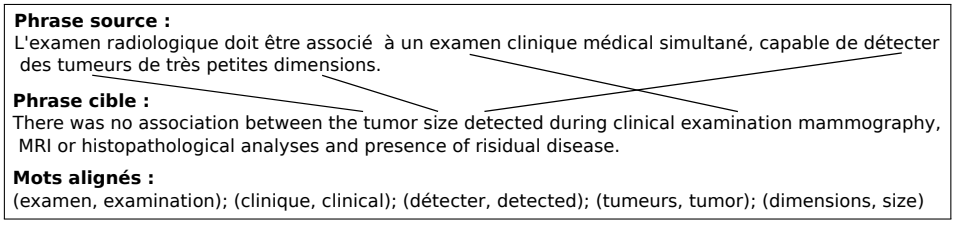


FIGURE 5.2 – *Exemple d'une phrase source et une autre cible contenant la paire de traductions (FR clinique, EN clinical)*

phrases parallèles permettent d'enrichir des textes parallèles utilisés par exemple dans les systèmes de traduction automatique statistique. Nous avons présenté quelques travaux concernant l'extraction des phrases parallèles à partir de corpus comparables dans le chapitre 2 sous la section 2.4. Ces travaux mènent des expériences avec de grands corpus (principalement des corpus de journaux). L'alignement est effectué d'abord au niveau des documents pour réduire l'espace de recherche des phrases parallèles. Or, les corpus spécialisés dont nous disposons contiennent au maximum quelques centaines de documents et comprennent peu de phrases parallèles. Ils sont aussi de taille modeste (environ 0,3 à 0,5 millions de mots). Néanmoins, nous supposons que certains traits utilisés dans la littérature de l'extraction des phrases parallèles peuvent être utilisés pour identifier des phrases comparables contenant une paire de traductions (terme source et son équivalent cible).

Nous soulignons que notre objectif n'est pas d'extraire des phrases parallèles, mais nous voulons trouver pour une paire de traductions des phrases qui sont le plus possible comparables. Par exemple, supposons que nous devions donner un score pour la paire de traductions (FR *clinique*, EN *clinical*) et que nous ayons deux phrases : la première contient FR *clinique* et la deuxième contient EN *clinical* (voir la figure 5.2). Les deux phrases ne sont pas parallèles, cependant, toutes les deux contiennent l'information suivante : un examen clinique détecte la taille d'une tumeur (quatre mots alignés entre les deux phrases). Ces phrases qui contiennent une paire de traductions candidate et qui ont des mots ou des segments alignés peuvent être considérées comme comparables. Les phrases comparables sont des indices fiables et a priori disponibles pour tous les termes.

### 5.3.1 Extraction des phrases candidates privilégiées pour un terme

Nous supposons que les mots spécifiques au domaine du corpus qui cooccurrent avec un terme ($t$) peuvent être exploités pour extraire les phrases privilégiées de $t$.

Chaque mot dans une phrase contenant $t$ est donc pondéré par : (a) son association avec $t$ ; (b) sa spécificité dans le corpus par rapport à un corpus de langue générale.

Nous calculons le score d'un mot dans une phrase contenant $t$ à partir des éléments suivants [2] :

1. **Association avec** $t$ :
   Nous choisissons de calculer l'association entre le terme et un mot ($m$) par la mesure de taux de vraisemblance (*logliklihood ratio*, Dunning (1993)) au niveau du corpus. Cette mesure est basée sur les cooccurrences de mots dans une fenêtre de taille ($w$= 7) autour de $t$. Les ($k$ = 30) [3] mots qui ont les scores d'association les plus élevés avec $t$ sont dénotés par $v_k$ (vecteur de contexte de $t$ de taille $k$). L'association entre $m$ et $t$ est calculée à partir d'occurrences qui sont présentées dans la table de contingence (voir la table 5.2), où occ(t, m) est le nombre d'occurrences de $t$ et $m$, et $\neg m$ signifie tous les mots sauf $m$.

| | **m** | **¬m** |
|---|---|---|
| **t** | a=occ(t,m) | b=occ(t,¬m) |
| **¬t** | c=occ(¬t,m) | d=occ(¬t,¬m) |

TABLE 5.2 – *Table de contingence pour t et w*

   La mesure d'association est calculée comme suit :

$$\begin{aligned} \text{association}(t,m) = {} & a\log(a) + b\log(b) + c\log(c) + d\log(d) \\ & + (N)\log(N) - (a+b)\log(a+b) - (a+c)\log(a+c) \\ & - (b+d)\log(b+d) - (c+d)\log(c+d) \end{aligned} \tag{5.1}$$

   où $N$=$a$+$b$+$c$+$d$. L'association entre $t$ et $m$ est ensuite divisée par le score d'association le plus élevé avec $t$ afin d'obtenir un score $\in$[0,1] ($\text{association}_{\text{normalisé}}$).

2. **Spécificité au domaine** :
   Le calcul de la spécificité d'un mot à un domaine s'inscrit dans la problématique de l'extraction terminologique en corpus. L'extraction des termes en corpus peut être effectuée soit à l'aide des informations linguistiques prédéfinies (ex. morphologique, syntaxique), soit à l'aide des traits statistiques dans le corpus (ex. fréquences des mots), voir (Streiter et al. 2003) pour plus de détails. Certaines mesures statistiques proposées dans la littérature (ex. l'information mutuelle) ne peuvent être utilisées que pour les composés syntagmatiques ou les termes complexes car elles se basent sur l'association entre les unités/composants de ces composés/termes.
   Dans ce travail, nous voulons calculer le score de spécificité d'un mot simple dans une phrase. Pour cela, nous choisissons d'utiliser la mesure de *weirdness ratio* proposée dans (Khurshid et al.

2. Tous les seuils et les valeurs ont été fixés de manière empirique sur le corpus cancer du sein français-anglais et une liste de termes existant dans ce corpus.

3. Les contextes privilégiés doivent contenir des mots qui apparaissent souvent avec le terme en question. Nos expériences ont montré que des valeurs de $k$ supérieures à 15 et inférieurs à 50 donnent de meilleurs résultats.

1994). Elle est définie comme étant la fréquence relative d'un mot $m$ dans un corpus spécifique au domaine ($dc$=$\{m_1,m_2,..,m_{n_1}\}$) divisée par sa fréquence relative dans un corpus issu de langue générale ($gc$=$\{m_1{}',m_2{}',..,m_{n_2}{}'\}$) :

$$\mathrm{ds}(m) = \frac{rvf_{dc}(m)}{rvf_{gc}(m)} \tag{5.2}$$

où $rvf_{dc}$=$\frac{freq_{dc}(m)}{\sum_{m_i \in dc} freq_{dc}(m_i)}$ est la fréquence relative dans un corpus spécifique au domaine, $rvf_{gc}(m)$=$\frac{freq_{gc}(m)}{\sum_{m_i' \in gc} freq_{gc}(m_i')}$ est la fréquence relative dans un corpus de langue générale et $freq$ signifie fréquence.

La spécificité d'un terme est ensuite divisée par la valeur de spécificité la plus élevée parmi les spécificités de tous les mots dans le corpus ($\mathrm{ds}_{\text{normalisé}}$).

Afin d'extraire les $n$ phrases privilégiées pour un terme $t$, un score est donné à chaque phrase $S$ contenant $t$ et des mots (non vides) $m1, m2, ..., m_n$ comme suit :

$$score(S) = \sum_{i=1}^{n} \Big( \mathrm{association}_{\text{normalisé(si } m_i \in v_k)}(m_i, t) + \mathrm{ds}_{\text{normalisé}}(m_i) \Big) / |S| \tag{5.3}$$

où $|S|$ est le nombre de mots dans $S$. Toute phrase de longueur inférieure à 5 mots n'est pas retenue [4]. Toutes les phrases contenant $t$ sont ensuite ordonnées en fonction de leurs scores.

Pour une paire de traduction ($t_s$, $t_c$), les $n$ phrases privilégiées pour $t_s$ ainsi que pour $t_c$ sont extraites en suivant la méthode expliquée ci-dessus.

La prochaine étape consiste à aligner les $n$ phrases privilégiées d'un terme source avec les $n$ phrases privilégiées de chacune de ses traductions candidates.

### 5.3.2 Alignement des phrases pour une paire de traductions

Nous faisons l'hypothèse que si un terme source ($t_s$) se traduit par un terme cible ($t_c$), les deux doivent partager quelques phrases comparables. Plus $t_s$ et $t_c$ partagent des phrases comparables, plus leur score devrait être élevé.

Nous supposons également que le nombre de mots alignés [5] entre deux phrases comparables devraient être supérieur ou égal à 2. Comme les approches d'extraction de phrases parallèles dans la littérature, notre approche repose principalement sur la similarité lexicale entre deux phrases bilingues.

4. Nous supposons qu'une phrase très courte ne peut pas être représentative du contexte d'un terme. Nos expériences ont montré qu'une phrase de longueur inférieure à 5 mots est trop courte pour être alignée avec d'autres phrases.

5. Trouvés en utilisant un dictionnaire bilingue.

Supposons que nous ayons une phrase source $S_s$={$m_1$,$m_2$,$t_s$,...,$m_n$}[6] et une phrase cible $S_c$={$m_1^{'}$,$m_2^{'}$,$t_c$...,$m_n^{'}$}[7] (après avoir enlevé les mots vides), avec un ensemble de mots alignés $M$={($m_1$,$m_1^{'}$),($m_2$,$m_2^{'}$),..,($m_n$,$m_n^{'}$)}. Un alignement optimal 1 :1 $A$ (chaque mot dans la phrase $S_s$ est aligné avec un mot au maximum dans la phrase $S_t$) est estimé selon une fonction linéaire que nous expliquons dans la suite.

Supposons que nous ayons calculé l'alignement optimal $A$ : nous pouvons utiliser des fonctions (chacune $\in$ [0,1]) pour calculer un score entre deux phrases en exploitant cet alignement. Nous choisissons quatre fonctions définies à partir des traits, comme suit :

1. la similarité cosinus entre les deux phrases (Fung et Cheung 2004) : chaque mot dans $S_s$ (respectivement $S_c$) est pondéré par son score dans le vecteur de contexte $v_k$[8] (respectivement $v_k^{'}$) de $t_s$ (respectivement $t_c$). Si un mot est absent du $v_k$, un score minimal fixe lui sera attribué.
   La première fonction est définie de la manière suivante :

$$f_1(S_s, S_c) = \frac{\cos(S_s, S_c)}{|\text{Mots non-alignés}|} \tag{5.4}$$

   où |Mots non-alignés| est le nombre de mots non-alignés entre les deux phrases.

2. les positions des mots alignés existant dans la phrase source (respectivement cible) par rapport à la position du terme source (respectivement cible) : plus les mots alignés sont proches du terme $t$ dans la phrase, plus le score de cette fonction sera élevé. En outre, nous supposons que pour deux mots alignés ($m_i$,$m_i^{'}$), la distance entre $m_i$ et $t_s$ devrait être proche de la distance entre $m_i^{'}$ et $t_c$ (voir l'exemple dans figure 5.3). En effet, il a été empiriquement observé par (Rapp 1999, p. 522) que l'ordre des mots non vides est souvent similaire entre les langues (même entre les langues éloignées telles que l'anglais et le chinois). Nous définissons la distance des positions comme suit :

$$pos_{\text{distance}}(S_s, S_c) = \sum_{m_i, m_i^{'} \in A} \frac{(pos_s + pos_c + |pos_s - pos_c|)}{|S_s| + |S_c| + |S_s - S_c|} \tag{5.5}$$

   où $pos_s$=|$pos(m_i)$-$pos(t_s)$| et $pos_t$=|$pos(m_i^{'})$-$pos(t_c)$|.
   La distance $pos_{\text{distance}}$ est ensuite divisée par $|A|$ (le nombre de mots alignés). La similarité des positions est calculée comme suit :

$$f_2(S_s, S_c) = 1 - \frac{pos_{\text{distance}}}{|A|} \tag{5.6}$$

6. $t_s$ pourrait être dans n'importe quelle position dans $S_s$.
7. $t_c$ pourrait être dans n'importe quelle position dans $S_c$.
8. Vecteur de contexte de $t_s$ de taille $k$.

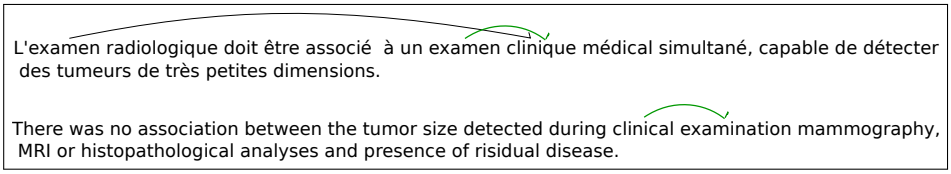


FIGURE 5.3 – *Exemple de deux phrases (source et cible) pour la paire de traductions (FR clinique, EN clinical). La première phrase contient deux occurrences de FR examen, les deux peuvent être alignées avec EN examination, la deuxième occurrence de FR examen est plus proche de FR clinique*

**Phrase source :**
... notamment la recommandation du diagnostic préopératoire systématique de cancer du sein invasif...

**Phrase cible :**
...with a diagnosis of invasive breast cancer, were entered into a database following local ethics approval...

**Plus longue séquence bilingue :**
(cancer, sein, infasif); (cancer, breast, invasive)

FIGURE 5.4 – *Exemple de deux phrases (source et cible) pour la paire de traductions (FR diagnostic, EN diagnosis) avec une séquence bilingue contigüe de longueur 3*

3. la séquence bilingue contigüe[9] la plus longue est définie par Munteanu et Marcu (2005) comme étant la plus longue paire de « chaînes où les mots dans une chaîne sont alignés uniquement avec des mots dans l'autre chaîne ». Un exemple de deux phrases qui partagent une séquence bilingue contiguë de longueur 3 est donné dans figure 5.4. Nous supposons que la longueur d'une séquence alignée doit être supérieure à 2 (une séquence égale à 2 constitue un bi-gramme ; les bi-grammes seront pris en compte dans le calcul d'une autre fonction). Elle est ensuite divisée par la longueur de la phrase la plus courte, cela donne la fonction suivante :

$$f_3(S_s, S_c) = \frac{\text{span}(S_s, S_c)}{\min(|S_s|, |S_c|)} \tag{5.7}$$

4. les bi-grammes en commun : cette fonction est définie comme le nombre de bi-grammes alignés divisé par la longueur de $A$ (en nombre de mots), comme suit :

$$f_4(S_s, S_c) = \frac{\text{bi-grammes}(S_s, S_c)}{|A|} \tag{5.8}$$

L'alignement optimal $A$ est l'alignement qui minimise la distance euclidienne au carré entre les deux phrases et la fonction $pos_{\text{distance}}$ (voir équation 5.5). En effet, nous choisissons cette fonction linéaire de minimisation afin de pouvoir utiliser l'algorithme hongrois[10] (Kuhn 1955).

Nous considérons le score final entre une paire de phrases comme la somme pondérée des fonctions proposées (comme dans Stefanescu et al. (2012)) :

$$score(S_s, S_c) = \sum_{i=1}^{4} \left(r_i \times f_i(S_s, S_c)\right) \tag{5.9}$$

9. Traduit du terme anglais *contiguous span*.
10. Un algorithme d'optimisation combinatoire qui permet de résoudre le problème d'affectation en temps polynomial.

où $\sum_{i=1}^{4}(r_i) = 1$.

Contrairement à d'autres travaux de la littérature utilisant des corpus parallèles pour entraîner leurs modèles et pour définir les poids des fonctions, nous définissons les poids de manière empirique sur le corpus cancer du sein français-anglais car nous n'avons pas de corpus parallèles annotés. Néanmoins, l'impact sur nos résultats ne devrait pas être important parce que notre but n'est pas d'extraire des phrases parallèles mais de réordonner les traductions candidates d'un terme en exploitant des phrases comparables.

### 5.3.3 Attribution de scores aux paires de traductions

Pour une paire de traductions ($t_s$,$t_c$), chacune des $n$ phrases privilégiées représentant $t_s$ est alignée avec au maximum une des $n$ phrases privilégiées représentant $t_c$. Une phrase cible peut être alignée avec plusieurs phrases sources. Le score entre $t_s$ et $t_c$ est la moyenne des scores des alignements de leurs phrases comparables ($P_{comparables} = \{(p_1, p_1^{'}), (p_2, p_2^{'}), ...\}$) :

$$score_{phrases}(t_s, t_c) = \frac{\sum_{p_i, p_i{}' \in P_{comparables}} score(p_i, p_i^{'})}{|P_{comparables}|} \tag{5.10}$$

Le reclassement est effectué en combinant le score obtenu par la méthode d'alignement des phrases pour une paire de traductions avec son score initial qui est obtenu par une approche distributionnelle ($score_{distributionnel}$). Les scores sont combinés par la moyenne géométrique pondérée, comme suit :

$$score_{final}(t_s, t_c) = score_{phrases}(t_s, t_c)^{\alpha} * score_{distributionnel}(t_s, t_c)^{\beta} \tag{5.11}$$

où $\alpha + \beta = 1$.

Nos expériences ont montré que la moyenne géométrique pondérée donne de meilleurs résultats que la moyenne arithmétique pondérée (ou la somme pondérée). Il y a une différence pratique entre les deux moyennes. La moyenne arithmétique pondérée permet à un élément qui a une valeur faible d'être compensé par d'autres éléments. Un élément qui a une valeur très faible va avoir un impact important sur le résultat obtenu par la moyenne géométrique pondérée et ne peut pas être compensé par d'autres éléments. En utilisant la moyenne géométrique pondérée, une traduction candidate ayant un score minimal (attribué par la méthode distributionnelle) va avoir un score $score_{final}$ très faible qui ne va pas pouvoir être compensé par le score d'alignement de phrases $score_{phrases}(t_s, t_c)$.

Dans la suite de ce chapitre, nous référencerons cette méthode de reclassement par la méthode d'alignement des phrases.

## 5.4 Évaluation

Afin d'évaluer notre approche, nous devons préalablement extraire des traductions pour une liste de termes spécifiques au domaine du corpus. Pour ce faire, nous alignons une liste de termes sources avec la méthode distributionnelle implémentée par TermSuite [11]. TermSuite fournit un nombre défini de meilleures traductions pour un terme. Nous choisissons la configuration par défaut de TermSuite : une fenêtre de taille 3 avec la mesure de taux de vraisemblance pour calculer les vecteurs de contexte de mots, et la mesure de Jaccard (Grefenstette 1994) pour calculer les similarités entre les vecteurs de contexte de mots.

Notre but est d'améliorer les classements des traductions correctes en se basant sur les $n$ meilleures traductions proposées pour chaque terme source.

### 5.4.1 Ressources

Pour réaliser nos expériences, nous avons besoin de corpus comparables, de dictionnaires bilingues, de corpus monolingues de langue générale et d'une liste de termes sources à traduire.

– **Corpus comparables** : nous effectuons des expériences avec des corpus comparables français-anglais et français-allemand du domaine du cancer du sein et des énergies renouvelables. Nous réalisons aussi des expériences avec la paire de langues français-espagnol dans le domaine des énergies renouvelables (voir le chapitre 1 section 1.4 pour plus de détails sur les corpus utilisés). Les corpus sont lemmatisés à l'aide de TermSuite. Nous réalisons les expériences sur les lemmes des mots dans les corpus.
– **Dictionnaires bilingues** : nous utilisons les dictionnaires présentés dans le chapitre 1 sous la section 1.4.2, pour les paires de langues FR-EN, FR-DE et FR-ES.
– **Corpus de langue générale** : pour chaque langue, un corpus de langue générale est disponible dans TermSuite pour le calcul des spécificités de mots. Ces corpus sont collectés à partir des textes de journaux et contiennent 12 003, 3 903, 44 365 et 149 472 mots simples uniques pour le français, l'anglais, l'allemand et l'espagnol respectivement.
– **Listes de termes sources** : nous avons à notre disposition, pour chaque corpus et pour chaque paire de langues, une liste qui se compose de termes simples à traduire. Chaque terme dans une liste est spécifique au domaine avec une fréquence supérieure à 5 dans le corpus source. Un terme est aligné manuellement avec une traduction de référence (un terme simple) qui existe dans le corpus cible avec une fréquence également supérieure à 5.
Pour le corpus de cancer du sein, une liste qui contient 122 paires de traductions est disponible pour chaque paire de langues. La liste FR-EN

11. Outil d'extraction et d'alignement de termes, présenté en section 2.6 du chapitre 2.

a été extraite du méta-thésaurus UMLS [12] et du *Grand dictionnaire terminologique* [13], cette liste a été utilisée dans les expériences menées par Morin et Prochasson (2011) et Hazem et al. (2011). La partie française de cette liste a ensuite été traduite manuellement en allemand en utilisant des dictionnaires en ligne et le moteur de recherche de traductions Linguee [14].
Quant au corpus des énergies renouvelables, pour chaque paire de langues, nous construisons manuellement une liste de références à l'aide des listes monolingues de termes extraits automatiquement du corpus. D'abord, des listes de termes sont extraites à partir de chaque corpus en utilisant les informations sur leurs fréquences et leurs spécificités calculées par la mesure *weirdness ratio* (voir l'équation 5.2). Dans le cadre du projet TTC, des listes de références de termes d'une langue source alignés avec leurs équivalents dans une langue cible (en comparant les listes monolingues extraites) sont disponibles. La construction des listes de références est détaillée dans (Loginova et al. 2012). Cependant, il existe au maximum seulement 35 termes simples alignés avec des termes simples dans chacune de ces listes. Pour construire des listes de plus grande taille, nous avons étendu les listes disponibles en alignant manuellement des termes entre les listes monolingues de termes (à l'aide des dictionnaires bilingues en ligne et le moteur de recherche de traductions Linguee). Des extraits des listes de références FR-EN pour chaque corpus sont donnés dans la table 5.3.

| **Cancer du sein** | | **Énergies renouvelables** | |
|---|---|---|---|
| **FR** | **EN** | **FR** | **EN** |
| kinase | kinase | arbre | shaft |
| cancer | cancer | nacelle | nacelle |
| tamoxifène | tamoxifen | aérodynamique | aerodynamic |
| antécédent | history | portance | lift |
| curage | dissection | tour | tower |

TABLE 5.3 – *Extraits de listes de références FR-EN*

### 5.4.2 Paramètres d'évaluation

Pour la méthode d'alignement des phrases, nous définissons de manière empirique les mêmes paramètres pour le corpus cancer du sein et le corpus des énergies renouvelables. Tous les paramètres ont été fixés de manière empirique sur le corpus cancer du sein français-anglais et la liste de références respective.

**Extraction des phrases** Pour chaque terme et chacune de ses traductions candidates [15], les 70 phrases privilégiées sont extraites, les phrases parta-

12. http ://www.nlm.nih.gov/research/umls
13. http ://www.granddictionnaire.com/
14. http ://www.linguee.fr/
15. Pour simplifier l'explication, nous considérons qu'une traduction candidate est un terme.

geant le même score auront le même rang. Nous retenons 200 phrases au maximum pour un terme ou une traduction candidate. Si un terme apparaît moins de 70 fois dans le corpus, toutes les phrases contenant ce terme seront retenues. Nous n'avons pas besoin d'extraire un grand nombre de phrases pour un terme parce que le processus d'alignement des phrases sera coûteux. En outre, notre hypothèse est que si une paire de traductions est correcte, cela signifie que ses phrases privilégiées sont comparables.

Nous considérons qu'une phrase est simplement délimitée par des ponctuations ( ?, !,.). Nous notons que les mots dans une phrase contenant un terme $t$, qui sont utilisés dans le calcul du score de cette phrase et considérés comme contexte pour $t$, sont les mots apparaissant au maximum dans une fenêtre de taille $n = 10$ autour de $t$ (10 mots ou moins qui précèdent $t$ dans la phrase et 10 mots ou moins qui suivent $t$ dans la phrase [16]). C'est-à-dire qu'en prenant 20 mots au maximum autour d'un terme, certains mots dans une phrase qui contient plus de 10 mots après ou avant $t$ seront négligés.

**Alignement des phrases** Pour donner un score à une paire de traductions en alignant ses phrases (voir l'équation 5.9), nous avons besoin de définir les poids. Nous fixons le poids de la première fonction à 0,4. Chaque autre poids dans l'équation est fixé à 0,2. Nous avons essayé plusieurs valeurs et nous avons conclu à partir des résultats que la première fonction est celle qui devrait avoir le poids le plus important et que les autres fonctions ont un impact plus ou moins égal sur les résultats. Nous avons donc choisi ces poids à partir des expériences et par soucis de simplicité.

**Attributions des scores** En combinant les scores de l'approche distributionnelle et la méthode d'alignement par la moyenne géométrique pondérée, le poids de la première est fixée à 0,3 et le poids de la seconde est fixé à 0,7.

### 5.4.3 Mesures d'évaluation

La précision d'un lexique bilingue est calculée à différents niveaux après avoir pris les $n$ meilleures traductions candidates pour chaque terme (top 1, top 5, etc). La précision est le nombre des traductions correctes trouvées divisé par le nombre de termes sources dans la liste de références.

La mesure du rang réciproque moyen [17] (MRR) est également utilisée pour évaluer les résultats obtenus. Le score du rang réciproque d'un terme source donné est l'inverse du rang de sa première traduction cible correcte dans la liste de traductions candidates. Le MRR est la moyenne des rangs réciproques des termes sources de références alignés. Les valeurs de

16. Nos expériences ont montré qu'il fallait prendre une taille de fenêtre plus grande que celle utilisée pour l'approche distributionnelle qui nous a permis d'obtenir le lexique bilingue. La taille de cette fenêtre ne doit pas être très grande non plus.

17. Traduit de l'anglais *Mean Reciprocal Rank*.

MRR sont comprises entre 0 et 1 ; des valeurs plus élevées indiquent de meilleurs résultats.

$$\text{MRR} = \frac{1}{Q}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \tag{5.12}$$

où $|Q|$ est le nombre de termes sources à aligner. Si la traduction correcte du terme n'a pas été trouvée, le score $\frac{1}{rank_i}$ correspondant au terme source sera égal à 0.

### 5.4.4 Expériences

Les résultats de l'approche distributionnelle de base avec trois paires de langues et deux corpus sont donnés dans la table 5.4 (P1 signifie la précision lorsque la première traduction candidate est proposée pour un terme). Nous pouvons remarquer que les résultats sur le corpus du cancer du sein sont meilleurs que ceux sur le corpus des énergies renouvelables (en comparant les mêmes paires de langues). Ceci peut être justifié par le fait que les corpus des énergies renouvelables sont d'une moins grande taille.

Les résultats sont également nettement meilleurs avec les paires de langues français-anglais et français-espagnol qu'avec la paire de langues français-allemand. En effet, les corpus spécialisés contiennent en général de nombreux termes complexes. Pour la langue allemande, de nombreux mots composés peuvent être écrits comme des unités simples (par exemple, le terme allemand *Produktionsstandort* est traduit en français par *site de production*). Par conséquent, l'approche distributionnelle peut considérer de tels termes allemands comme un seul mot lors du calcul de cooccurrences. Afin de résoudre ce problème, nous pouvons effectuer un découpage de tels termes avant d'appliquer l'approche distributionnelle (Weller et al. 2011).

Pour comprendre l'écart des résultats obtenus en fonction de paires de langues, nous avons calculé la comparabilité des corpus en utilisant la mesure de comparabilité présentée dans Li et Gaussier (2010) (nous avons présenté cette mesure en section 2.5 du chapitre 2). Pour le corpus du cancer du sein, nous avons obtenu une valeur de comparabilité de 0,79 pour le corpus français-anglais et 0,52 pour le corpus français-allemand. Pour le corpus des énergies renouvelables, nous avons obtenu une valeur de comparabilité de 0,81 pour le corpus français-anglais, 0,82 pour le corpus français-espagnol et 0,70 pour le corpus français-allemand. La mesure de Li et Gaussier (2010) ne nous permet pas de comparer les comparabilités des corpus des domaines différents. Elle peut nous donner une idée sur les comparabilités des corpus dans le même domaine avec les différentes paires de langues. En utilisant le dictionnaire français-anglais sur le corpus de cancer du sein français-anglais ou le corpus des énergies renouvelables, nous avons obtenu un score de comparabilité plus élevé que le score de comparabilité obtenu avec la paire de langues français-allemand.

Cela peut justifier en partie les mauvais résultats obtenus avec la paire de langues français-allemand par l'approche distributionnelle.

| | **Cancer du sein** | | **Énergies renouvelables** | | |
|---|---|---|---|---|---|
| | **FR-EN** | **FR-DE** | **FR-EN** | **FR-DE** | **FR-ES** |
| **P1** | 26,22 % | 9,16 % | 16,66 % | 3,12 % | 18,75 % |
| **P5** | 45,08 % | 18,85 % | 38,54 % | 9,37 % | 29,16 % |
| **P10** | 53,27 % | 26,22 % | 45,83 % | 10,41 % | 39,58 % |
| **P15** | 59,01 % | 29,50 % | 50,00 % | 12,50 % | 40,62 % |
| **P20** | 60,65 % | 31,96 % | 57,29 % | 14,58 % | 43,75 % |
| **P25** | 61,47 % | 32,78 % | 59,37 % | 14,58 % | 45,83 % |

TABLE 5.4 – *Résultats obtenus avec l'approche distributionnelle de base (Baseline)*

Afin d'améliorer ces résultats, en particulier les précisions top 1, top 5 et top 10, nous essayons de ré-ordonner les traductions candidates pour chaque terme source.

Supposons que pour un terme source $t_s$, nous voulons ré-ordonner ses 5 premières traductions candidats $L_{top5}$=$\{t_{c_1}$,$t_{c_2}$,$t_{c_3}$,$t_{c_4}$,$t_{c_5}\}$ fournies par l'approche distributionnelle. En suivant l'approche présentée dans la section 5.3.1, nous extrayons les phrases privilégiées pour $t_s$ . De même, nous extrayons des phrases pour chaque traduction candidate dans $L_{top5}$. Ensuite, pour chaque paire de traductions (par exemple $t_s$ et $t_{c_1}$), nous essayons d'aligner chaque phrase extraite pour $t_s$ avec une phrase qui partage le meilleur score avec elle parmi les phrases extraites pour $t_{c_1}$, selon la méthode décrite dans la section 5.3.2. Une phrase source peut être alignée avec au maximum une phrase cible. Si une phrase source n'a pas pu être alignée avec une phrase cible, elle sera alignée avec $\phi$. Le score donné pour $t_s$ et $t_{c_1}$ est la moyenne des scores des alignements de leurs phrases.

En suivant la procédure décrite ci-dessus, nous prenons les meilleures $n$ = 20 traductions candidates proposées par l'approche distributionnelle pour chaque terme et nous essayons de les ré-ordonner. Cette stratégie d'évaluation est appelée **RR1** dans les tables 5.5, 5.6 et 5.7. Ces dernières présentent les résultats obtenus sur les corpus avec trois paires de langues. Par exemple, sur la liste de cancer du sein français-anglais, nous constatons que le reclassement (par RR1) améliore la précision top 1 d'environ 5 % (ex. la traduction correcte de FR *exérèse* a été re-ordonné du 4ème au 1èr rang, la traduction correcte de FR *génomique* a été re-ordonné du 5ème au 1èr rang). Avant le reclassement, 43,24 % des traductions correctes trouvées dans toutes les listes de top 20 ont été classées au premier rang, après le reclassement, ce pourcentage augmente à 52,70 %. Ce qui signifie que le reclassement a amélioré les rangs des traductions correctes. Une amélioration d'environ 6 % de précision top 1 est obtenue avec la liste de références pour le corpus des énergies renouvelables français-anglais. Cependant, nous obtenons de moins bons résultats avec la paire de langues français-allemand. Cela est surtout dû au fait qu'il n'y a pas beaucoup de traductions correctes dans les listes de 20 premières traductions fournies pour chaque terme par la méthode distributionnelle.

En effectuant les expériences, nous avons remarqué que le reclassement des 5 premières traductions candidates pour chaque terme peut per-

| | Cancer du sein | | | Énergies renouvelables | | |
|---|---|---|---|---|---|---|
| FR-EN | **Baseline** | **RR1** | **RR2** | **Baseline** | **RR1** | **RR2** |
| **P1** | 26,22 % | 31,96 % | **35,24 %** | 16,66 % | **23,95 %** | 22,91 % |
| **P5** | 45,08 % | **52,45 %** | **52,45 %** | 38,54 % | **45,83 %** | 44,79 % |
| **P10** | 53,27 % | **57,37 %** | **57,37 %** | 45,83 % | 48,95 % | **52,08 %** |
| **MRR** | 0,338 | 0,396 | **0,419** | 0,249 | **0,324** | 0,319 |

TABLE 5.5 – *Résultats obtenus sur la paire de langues français-anglais en utilisant les corpus du cancer du sein et des énergies renouvelables*

| | Cancer du sein | | | Énergies renouvelables | | |
|---|---|---|---|---|---|---|
| FR-DE | **Baseline** | **RR1** | **RR2** | **Baseline** | **RR1** | **RR2** |
| **P1** | 9,16 % | **11,47 %** | **11,47 %** | 3,12 % | **7,29 %** | 5,20 % |
| **P5** | 18,85 % | **21,31 %** | **21,31 %** | 9,37 % | **10,41 %** | **10,41 %** |
| **P10** | 26,22 % | **27,04 %** | **27,04 %** | 10,41 % | **13,51 %** | **13,51 %** |
| **MRR** | 0,139 | 0,160 | **0,162** | 0,051 | **0,088** | 0,075 |

TABLE 5.6 – *Résultats obtenus sur la paire de langues français-allemand en utilisant les corpus du cancer du sein et des énergies renouvelables*

| | Énergies renouvelables | | |
|---|---|---|---|
| FR-ES | **Baseline** | **RR1** | **RR2** |
| **P1** | 18,75 % | **23,95 %** | 21,87 % |
| **P5** | 29,16 % | **35,41 %** | **35,41 %** |
| **P10** | 39,58 % | 40,62 % | **41,66 %** |
| **MRR** | 0,235 | **0,287** | 0,269 |

TABLE 5.7 – *Résultats obtenus sur la paire de langues français-espagnol en utilisant le corpus des énergies renouvelables*

mettre d'augmenter la précision top 1 plus que si, par exemple, nous ré-ordonnons les 20 premières traductions candidates pour chaque terme. Pour cela, nous avons décidé de suivre une stratégie différente (indiquée par **RR2**) pour le reclassement de traductions. Pour déterminer quelle traduction candidate sera classée au rang $n$ (à partir de 1) pour un terme, nous ré-ordonnons d'abord la liste de top m=$\big($(2(n-1)+5) arrondit au multiple de 5 le plus proche$\big)$ traductions proposées pour chaque terme. La traduction candidate au rang 1 aura le rang $n$ dans la liste ré-ordonnée et le nouveau rang de cette traduction sera fixe. Ensuite, nous déterminons la traduction candidate qui sera classée au rang $(n+1)$ dans la liste ré-ordonnée. Nous répétons ce processus jusqu'à l'obtention de 10 traductions candidates pour chaque terme dans la liste ré-ordonnée.

Par exemple, prenons une liste de 25 traductions candidates proposées pour un terme. Afin de déterminer quelle traduction candidate sera classée au premier rang dans la liste ré-ordonnée : nous ré-ordonnons d'abord la liste des top 5 ($L_{top5}$) traductions candidates proposées pour ce terme. La traduction désormais classée au premier rang sera ajoutée à une liste que nous appelons $L_{pris}$. Pour déterminer quelle traduction candidate sera au deuxième rang, nous ré-ordonnons la liste ($\{L_{top5}\} - \{L_{pris}\}$) et nous ajoutons la traduction classée au premier rang à $L_{pris}$. Maintenant, pour déterminer quelle traduction sera classée au troisième rang, nous ré-ordonnons

la liste ({liste des top 10} − {$L_{pris}$}), et nous ajoutons la traduction classée au premier rang à $L_{pris}$, et ainsi de suite. Les résultats obtenus en utilisant cette stratégie sont présentés dans les tables 5.5,5.6 et 5.7 sous les colonnes nommées RR2.

La stratégie RR2 donne une meilleure précision top 1 et une meilleure MRR que RR1 avec le corpus cancer du sein français-anglais. Elle donne aussi une meilleure précision top 10 avec le corpus des énergies renouvelables. Alors que la stratégie RR1 donne une meilleure MRR sur le corpus des énergies renouvelables. En général, les résultats des deux stratégies sont comparables. RR1 et RR2 améliorent les résultats de l'approche de base significativement sur les trois paires de langues français-anglais, français-espagnol et français-allemand.

## 5.5 Discussion

La stratégie RR2 est théoriquement une méthode plus fiable pour ré-ordonner les traductions candidates d'un terme que la stratégie RR1, notamment pour les termes qui partagent des phrases comparables avec beaucoup de leurs traductions candidates. Il est plus probable que RR2 garde aux premiers rangs les traductions correctes qui sont déjà bien classées. Par exemple, selon la première étape de RR2, une traduction correcte qui est classée au deuxième rang par la méthode distributionnelle sera en concurrence avec quatre autres traductions pour être classée au premier rang [18]. Si nous utilisons RR1, cette traduction correcte (classée au deuxième rang par l'approche distributionnelle) devra être en concurrence avec $n$=(20-1) traductions pour être au premier rang dans la liste ré-ordonnée.

En comparant manuellement les classements des traductions correctes dans la liste de base et la liste ré-ordonnée (sur le corpus de cancer du sein français-anglais), nous remarquons que tous les composés savants existants dans la liste de base (19 composés savants dont certains sont peu fréquents dans le corpus, ex. *oncologue*) ont soit conservé leur rang, soit été mieux classés dans la liste ré-ordonnée. Cela peut indiquer que cette méthode de reclassement peut également fonctionner sur les termes complexes car les termes complexes posent certains problèmes de traduction en commun avec ceux des composés savants (ex. peu fréquents dans un corpus, voir les sections 2.2 et 3.4 dans les chapitres 2 et 3 respectivement pour d'autres problèmes de traduction de tels termes).

Nous donnons un exemple (voir figure 5.5) des deux premières paires de phrases comparables trouvées par l'approche que nous proposons pour la paire de traductions (FR kinase, EN kinase), en utilisant le corpus cancer du sein français-anglais. Les phrases comparables de cette paire de traductions nous ont permis de ré-ordonner la traduction correcte EN *kinase* du quatrième rang au premier rang. Chaque paire de phrases dans l'exemple partage 3 mots alignés en commun. Prenons la première paire de phrases :

18. RR2 choisira l'une des 5 meilleures traductions candidates d'un terme au premier rang dans la liste re-ordonnée.

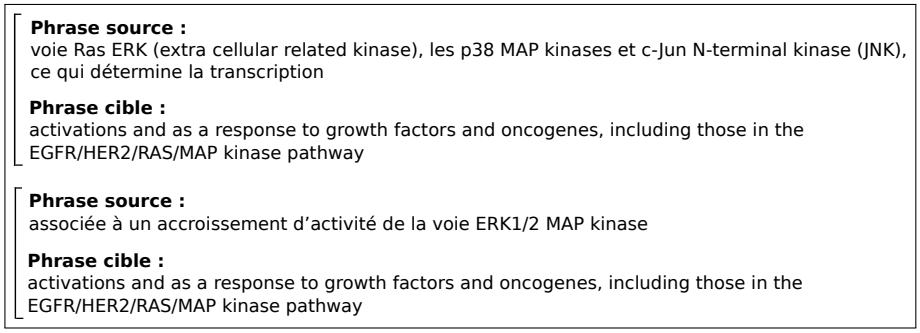


FIGURE 5.5 – *Exemple des résultats de notre approche : deux phrases sources alignées avec une phrase cible pour la paire de traductions (FR kinase, EN kinase)*

il y a 9 mots qui séparent le mot FR *voie* et le mot FR *kinase* dans la phrase en français et il y a aucun mot qui sépare le mot EN *pathway* (traduction de FR *voie*) et le mot EN *kinase*. À partir de la deuxième paire de phrases, nous pouvons remarquer que les mots FR *voie* et FR *kinase* sont séparés par deux mots dans la phrase en français.

En général, une paire de phrases comparables trouvée par notre approche (à partir du corpus cancer du sein français-anglais), contient 3 à 4 mots alignés en moyenne. Les positions de ces mots alignés peuvent être proches ou éloignées de la position du terme en question dans chaque phrase.

Nous estimons que les phrases comparables (extraites par notre approche) ne peuvent pas améliorer la performance des systèmes de traduction automatique statistique (TAS). En effet, les phrases bilingues n'étant pas parallèles peuvent dégrader la performance des systèmes de TAS si elles sont ajoutées aux textes parallèles (Munteanu et Marcu 2006). Puisque notre approche n'est pas destinée à extraire des phrases parallèles, nous envisageons de tester l'impact du lexique bilingue extrait avant et après le reclassement sur un système de TAS.

## 5.6 Conclusion

Dans ce chapitre, une méthode a été proposée pour ré-ordonner les traductions candidates acquises par une approche distributionnelle à partir de corpus comparables.

Nous avons supposé que certaines phrases étaient plus représentatives du contexte d'un terme que d'autres, et qu'un terme et sa traduction partagent des phrases comparables qui peuvent être extraites à partir de corpus comparables.

Des phrases représentant les contextes privilégiés d'un terme ont été alignées avec des phrases représentant les contextes privilégiés de ses traductions candidates pour ré-ordonner ces dernières. Nos expériences ont montré des améliorations de la qualité d'un lexique bilingue pour trois paires de langues et deux domaines.

La méthode de reclassement a été appliquée sur des termes simples.

Nous cherchons à évaluer davantage cette méthode sur des termes complexes. En outre, il est envisageable d'étudier la possibilité de proposer des phrases comparables (d'une paire de traductions) pour aider le travail d'un traducteur humain qui souhaite vérifier si cette paire est correcte ou non.

# Conclusion et perspectives

Cette thèse a porté principalement sur l'enrichissement de lexiques bilingues spécialisés et l'amélioration de leur qualité à partir des corpus comparables spécialisés.

Notre premier objectif a été d'étudier les propriétés de certains termes dans plusieurs langues. Pour cela, nous avons adapté des approches d'alignement de l'état de l'art à deux types de termes : les composés savants et les adjectifs relationnels. Notre méthode de traduction de composés savants a été testée sur le français, l'anglais et l'allemand, mais celle d'adjectifs relationnels a été vérifiée sur une seule paire de langue (français-anglais). Notre deuxième objectif a été d'améliorer la qualité d'un lexique bilingue. Pour cela, nous avons étudié l'exploitation de contextes bilingues riches en terminologies et en termes reliés dans le corpus afin de réordonner des traductions candidates proposées pour un terme. Nos expériences ont été menées pour deux domaines : le cancer du sein et les énergies renouvelables et sur trois paires de langues : français-anglais, français-allemand et français-espagnol.

## Synthèse des travaux réalisés

Pour donner une vue d'ensemble des travaux effectués, nous résumons chaque chapitre de manière individuelle.

Le **Chapitre 1** a décrit les concepts de base relatifs à cette thèse : les dictionnaires/lexiques bilingues, les termes (simples et complexes) et les corpus (parallèles et comparables). Ce chapitre a également présenté les ressources linguistiques à notre disposition pour réaliser nos expériences.

Le **Chapitre 2** a été consacré à l'état de l'art de l'extraction des correspondances bilingues en exploitant des corpus comparables. Il a mis en évidence des **approches distributionnelles** qui s'appuient sur les cooccurrences des mots dans un corpus afin de trouver des couples de traductions. Il a également présenté des **approches compositionnelles** qui s'appuient sur une propriété compositionnelle des composés syntagmatiques ou des termes complexes pour trouver leurs traductions (c'est-à-dire trouver la traduction de l'ensemble à partir des traductions des parties). On trouve également parmi les travaux de l'état de l'art présentés dans ce chapitre : (a) des approches qui concernent la traduction de certains termes complexes ne pouvant pas être traités par les approches compositionnelles classiques ; (b) des approches pour l'extraction de phrases parallèles à partir d'un corpus comparable ; (c) des approches qui proposent des mesures pour calculer le degré de comparabilité d'un corpus comparable et l'améliorer. Enfin, nous avons introduit l'outil d'extraction et d'alignement des

termes à partir de corpus comparables *TermSuite*, que nous avons utilisé dans nos expériences pour pré-traiter les corpus, extraire des termes et les aligner pour plusieurs paires de langues.

Le **Chapitre 3** a été inspiré des approches compositionnelles pour traduire des composés savants. Le travail effectué propose d'abord une méthode pour traduire les composés savants à l'aide d'une propriété compositionnelle. Ce travail reposait sur l'hypothèse qu'un composé savant pouvait être traduit par un composé savant à partir de ses composants. Ces composants peuvent être des racines gréco-latines ou des mots. À cette fin, des listes de racines gréco-latines bilingues ont été construites manuellement et semi-automatiquement. Cette méthode n'est capable de traduire qu'une partie des composés savants présents dans un corpus. Pour essayer de traduire plus de composés savants, nous avons suivi deux méthodes : (a) des traductions candidates peuvent être proposées pour un composé savant en l'associant, à l'aide d'une mesure de similarité de lettres, à un composé savant dont la traduction est connue ; (b) des traductions candidates trouvées par une approche distributionnelle peuvent être proposées à un composé savant, si elles ont en commun au moins une racine gréco-latine avec le composé savant à traduire. Les expériences sur deux domaines et deux paires de langues ont montré que la traduction des composés savants à l'aide d'une propriété compositionnelle fournit des alignements de haute précision. Les autres méthodes de traduction que nous avons proposées de composés savants ont donné de moins bonnes précisions mais ont permis de traduire un plus grand nombre de composés savants. Ce travail a été initialement publié dans plusieurs conférences Harastani et al. (2012), Weller et al. (2011), Gornostay et al. (2012).

Le **Chapitre 4** s'est concentré sur les adjectifs relationnels. Puis une approche a été proposée pour extraire les adjectifs relationnels, les aligner avec leurs noms de base et enfin traduire ceux qui apparaissent dans des termes de la forme [Nom + AdjR]. La traduction de ces termes complexes a été effectuée à l'aide d'une approche appelée « la traduction par paraphrases ». Ce travail poursuit celui de Morin et Daille (2010) qui propose une approche compositionnelle étendue à l'aide des règles de réécriture pour l'alignement d'un adjectif relationnel avec un nom. Notre travail a consisté dans l'automatisation de l'extraction et dans l'alignement des adjectifs relationnels. Nous avons également étudié les propriétés des adjectifs extraits d'un corpus. Chaque étape de l'approche proposée a été évaluée de manière individuelle. Les adjectifs extraits du corpus cancer du sein français ont été analysés et regroupés dans différentes catégories. Nous avons montré dans ce chapitre que les alignements (adjectif-nom) trouvés automatiquement peuvent aider à traduire les termes complexes français de la forme [Nom + AdjR] vers l'anglais avec une haute précision. Ce travail a fait une publication dans la conférence nationale TALN Harastani et al. (2013a).

Le **Chapitre 5** a proposé une approche de reclassement des traductions candidates fournies pour un terme. Ces traductions candidates sont fournies par une approche distributionnelle standard. Pour ré-ordonner les traductions candidates d'un terme : le travail a consisté à extraire

des phrases représentatives contenant le terme et ses traductions candidates. Les phrases extraites ont été ensuite alignées et exploitées pour réordonner les traductions candidates d'un terme. Nous avons montré que la qualité d'un lexique bilingue peut être améliorée par notre approche de reclassement. Les expériences ont été réalisées sur deux domaines et trois paires de langues : français-anglais, français-allemand et français-espagnol. Ce travail a été publié dans la conférence internationale IJCNLP Harastani et al. (2013b).

## Améliorations et Perspectives

Nous rappelons que les racines d'une langue alignées avec leurs significations ou avec des racines dans d'autres langues aident à la traduction compositionnelle des composés savants. Par exemple, le terme complexe FR *comptabilité tissulaire* peut être traduit par le composé savant EN *histocomptability* ; la traduction compositionnelle peut réussir si la signification de la racine *histo* a été identifiée comme étant *tissu*. Les significations des racines gréco-latines peuvent également aider à la traduction des termes d'une structure [Nom + AdjR], où *AdjR* est un adjectif relationnel construit à l'aide d'une racine gréco-latine (ex. hépatique). Les racines gréco-latines constituent donc une ressource essentielle pour les approches de traduction des composés savants et des adjectifs relationnels pour les langues romanes et germaniques. La construction des listes de racines gréco-latines (monolingues et bilingues) est une tâche fastidieuse. Nous avons extrait ces racines de manière semi-automatique, mais de futurs travaux peuvent être considérés pour améliorer cette extraction. L'acquisition automatique des racines gréco-latines alignées avec leurs significations est possible comme l'a montré le travail de Claveau et Kijak (2010). Cependant, ce travail ne peut traiter toutes les racines gréco-latines présentes dans un corpus car il s'appuie seulement sur un lexique bilingue (c'est-à-dire que les racines qui existent dans un corpus et non dans un lexique bilingue ne peuvent pas être alignées). Il faut donc envisager la construction des listes de racines gréco-latines bilingues à partir des corpus comparables en exploitant par exemple les traductions qui sont des composés syntagmatiques ou termes complexes (en une langue qui emploie peu de composés savants) obtenues par une approche distributionnelle pour des composés savants.

En effet, la traduction des termes à l'aide d'une hypothèse qui repose sur une propriété compositionnelle donne une haute précision. La traduction compositionnelle peut s'étendre à encore plus de types de termes qui ont des caractéristiques différentes ou en commun entre plusieurs langues. Toutefois, d'autres types de termes simples ou complexes ne possèdent pas de propriété compositionnelle qui peuvent aider à les traduire. En utilisant un corpus comparable et un dictionnaire bilingue, ces termes peuvent être traduits à l'aide d'une hypothèse distributionnelle mais les résultats ne sont pas toujours satisfaisants. Le travail effectué dans le chapitre 5 a proposé de ré-ordonner les traductions candidates fournies par une approche distributionnelle pour un terme. Il a été inspiré des ap-

proches proposées pour l'extraction de phrases parallèles à partir des corpus comparables. La méthode de reclassement proposée n'a été appliquée que sur des termes simples. Nous visons à évaluer aussi cette méthode sur des termes complexes et à améliorer l'attribution des paramètres utilisés. Nous estimons que les phrases comparables peuvent permettre de découvrir des relations entre les traductions proposées pour un terme et de donner un indice sur la fertilité des termes. Par exemple, pour le couple de traductions (FR *ganglion*, EN *node*), on remarque à partir des meilleurs phrases comparables de cette paire que la fertilité du terme FR *ganglion* peut être découverte : lorsqu'une phrase source FR *ganglion* est alignée avec une phrase cible qui contient EN *node*, cette phrase cible contient également EN *lymph*. En effet, l'équivalent du terme *ganglion* peut être le terme complexe *lymph node* en anglais. Les phrases comparables ont donc, comme les phrases parallèles, un potentiel qui peut nous permettre de trouver les traductions fertiles. En outre, les phrases comparables pour un terme et ses traductions candidates peuvent également être proposées à un traducteur, pour pouvoir comprendre les termes dans leurs contextes. Cet axe de recherche est exploré dans le cadre du projet CRISTAL [19], qui vise à développer une technologie d'extraction de contextes riches en connaissances permettant de produire de nouveaux types de dictionnaires. Pour chaque terme et ses traductions éventuelles, un dictionnaire devrait lister une fiche terminologique de leurs contextes riches en connaissance (ex. des phrases dans un corpus pouvant être des définitions d'un terme) afin de faciliter la compréhension du terme et de ses traductions.

Dans cette thèse, nous avons exploité les corpus comparables dans l'optique d'extraire des lexiques bilingues spécialisés. Cependant, ces lexiques extraits automatiquement à partir de corpus comparables sont parfois très bruités et ne sont pas prêts à être exploités par des utilisateurs ou des traducteurs, surtout quand il s'agit de traduire des termes simples. Dernièrement, des travaux dans la littérature suivent d'autres directions pour continuer à exploiter les corpus comparables. Par exemple, ils améliorent la comparabilité de certains corpus ou utilisent les corpus comparables et parallèles de manière conjointe. D'autres travaux font recours à des corpus alignés au niveau de documents, par exemple des corpus collectés à partir d'articles extraits de Wikipedia. Dans les domaines spécialisés, les corpus parallèles ou alignés au niveau de documents peuvent être rares. La recherche est encore ouverte à des méthodes pratiques pour l'amélioration de la qualité des corpus comparables spécialisés, ce qui s'avère indispensable si on veut continuer à exploiter avec succès ces corpus pour l'extraction des correspondances bilingues.

19. http ://www.projet-cristal.org/index.php/fr/

# Bibliographie


Antar Solhy Abdellah. The problem of translating English Linguistic Terminology into Arabic. Cambridge CAMLING conference, 2003. (Cité page 27.)

Sadaf Abdul-Rauf et Holger Schwenk. On the Use of Comparable Corpora to Improve SMT performance. Dans *Proceedings of the Conference of the European chapter of the association for Computational Linguistics (EACL'09)*, pages 16–23, 2009. (Cité pages 58 et 59.)

Dany Amiot et Georgette Dal. La composition néoclassique en français et l'ordre des constituants. *La composition dans les langues, Artois Presses Université*, pages 89–113, 2008. (Cité pages 67 et 68.)

Marianna Apidianaki, Nikola Ljubesic, et Darja Fiser. Vector Disambiguation for Translation Extraction from Comparable Corpora. *Informatica (Slovenia)*, 37(2) :193–201, 2013. (Cité page 45.)

Mona Baker. *In other words : A course book on translation*. Routledge, 1998. (Cité page 27.)

Timothy Baldwin et Takaaki Tanaka. Translation by Machine of Complex Nominals : Getting it Right. Dans *Proceedings of the ACL'04 Workshop on Multiword Expressions : Integrating Processing*, pages 24–31, 2004. (Cité pages 28, 46, 47, 48, 50, 51, 52, 53 et 70.)

Laurie Bauer. *English word-formation*. Cambridge university press, 1983. (Cité page 68.)

Hervé-D. Béchade. *Phonétique et morphologie du français moderne et contemporain*. Presses Universitaires de France, 1992. (Cité page 85.)

Henri Béjoint et Philippe Thoiron. *Les dictionnaires bilingues*. Aupelf-Uref - Editions Duculot, 1996. (Cité page 23.)

Dhouha Bouamor, Nasredine Semmar, et Pierre Zweigenbaum. Context Vector Disambiguation for Bilingual Lexicon Extraction from Comparable Corpora. Dans *Proceedings of the Meeting of the Association for Computational Linguistics (ACL'13) Short Papers*, pages 759–764, 2013. (Cité page 45.)

Lynne Bowker et Jennifer Pearson. *Working with specialized language : a practical guide to using corpora*. London, Routledge, 2002. (Cité pages 23, 26, 27 et 28.)

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, et Robert L. Mercer. The mathematics of statistical machine translation : parameter estimation. *Comput. Linguist.*, 19(2) :263–311, 1993. (Cité pages 27 et 57.)

James P. Callan, W. Bruce Croft, et John Broglio. TREC and TIPSTER Experiments with INQUERY. Dans *Information Processing and Management*, pages 31–3, 1994. (Cité page 57.)

Bruno Cartoni. *De l'incomplétude lexicale en traduction automatique : vers une approche morphosémantique multilingue*. PhD thesis, Université de Genève, 2008. (Cité pages 100, 101, 106 et 112.)

Bruno Cartoni. Lexical Morphology in Machine Translation : A Feasibility Study. Dans *Proceedings of the Conference of the European chapter of the Association for Computational Linguistics (EACL'09)*, pages 130–138, 2009. (Cité pages 54, 73 et 76.)

Yun-Chuang Chiao. *Extraction lexicale bilingue à partir de textes médicaux comparables : application à la recherche d'information translangue*. PhD thesis, Université Pierre et Marie Curie - Paris VI, 2004. (Cité page 46.)

Yun-Chuang Chiao et Pierre Zweigenbaum. Looking for candidate translational equivalents in specialized, comparable corpora. Dans *Proceedings of the International Conference on Computational linguistics (COLING'02)*, volume 2, pages 1–5, 2002. (Cité page 121.)

Vincent Claveau. Translation of Biomedical Terms by Inferring Rewriting Rules. Dans *Information Retrieval in Biomedicine : Natural Language Processing for Knowledge Integration*, pages 106–123. IGI Global, 2009. (Cité pages 72 et 76.)

Vincent Claveau et Ewa Kijak. Analyse morphologique en terminologie biomédicale par alignement et apprentissage non-supervisé. Dans *Actes de la conférence sur le Traitement Automatique des Langues Naturelles (TALN'10)*, 2010. (Cité pages 11, 70 et 141.)

Monique C. Cormier et John Humbley. *La terminologie : Théorie, méthode et applications*. Les presses de l'Université d'Ottawa, 1998. (Cité page 23.)

Henri Cottez. *Dictionnaire des structures du vocabulaire savant*. Les usuels du Robert, Paris, 1982. (Cité page 112.)

Béatrice Daille. Morphological Rule Induction for Terminology Acquisition. Dans *Proceedings of the International Conference on Computational Linguistics (COLING'00)*, pages 215–221, 2000. (Cité pages 100 et 105.)

Béatrice Daille. Conceptual structuring through term variations. Dans *Proceedings of the ACL'03 Workshop on Multiword Expressions : Analysis, Acquisition and Treatment*, pages 9–16, 2003. (Cité page 25.)

Hervé Déjean, Éric Gaussier, et Fatiha Sadat. An Approach Based on Multilingual Thesauri and Model Combination for Bilingual Lexicon Extraction. Dans *Proceedings of the International Conference on Computational linguistics (COLING'02)*, pages 1–7, 2002. (Cité pages 28 et 44.)

Estelle Delpech, Béatrice Daille, Emmanuel Morin, et Claire Lemaire. Extraction of Domain-Specific Bilingual Lexicon from Comparable Corpora : Compositional Translation and Ranking. Dans *Proceedings of the*

*International Conference on Computational Linguistics (COLING'12)*, pages 745–762, 2012. (Cité pages 27, 54, 75, 76 et 93.)

Amélie Depierre. Souvent HAEMA varie : Les dérivés du grec HAEMA en anglais : Étude de cas de variation. *Terminology*, pages 155–176, 2007. (Cité page 25.)

Jean Dubois et Françoise Dubois-Charlier. *La dérivation suffixale en français*. Nathan Université, 1999. (Cité pages 98, 99, 101 et 107.)

Estelle Dubreil. *La dimension argumentative des collocations textuelles en corpus électronique spécialisé au domaine du TALN*. PhD thesis, Université de Nantes, 2006. (Cité page 26.)

Ted Dunning. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1) :61–74, 1993. (Cité pages 40 et 125.)

Rosa Estopa, Jordi Vivaldi, et M. Teresa Cabre. Use of Greek and Latin forms for term detection. Dans *Proceedings of the International Conference on Language Resources and Evaluation (LREC'oo)*, volume 78, pages 885–859, 2000. (Cité page 77.)

Stefan Evert. *The Statistics of Word Cooccurrences. Word Pairs and Collocations*. Universität Stuttgart, 2005. (Cité page 44.)

Robert M. Fano. *Transmission of Information : A Statistical Theory of Communications*. PhD thesis, 1961. (Cité pages 39 et 108.)

Oana Frunza et Diana Inkpen. Identification and Disambiguation of Cognates, False Friends, and Partial Cognates Using Machine Learning Techniques. *International Journal of Linguistics*, 1(1), 2009. (Cité page 106.)

Pascale Fung. Compiling Bilingual Lexicon Entries From a Non-Parallel English-Chinese Corpus. Dans *Proceedings of the Workshop on Very Large Corpora*, pages 173–183, 1995. (Cité pages 35, 37, 38, 42 et 43.)

Pascale Fung et Percy Cheung. Mining Very-Non-Parallel Corpora : Parallel Sentence and Lexicon Extraction via Bootstrapping and EM. Dans *In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'04)*, pages 57–63, 2004. (Cité pages 28, 60 et 127.)

Pascale Fung et Kathleen Mckeown. Finding Terminology Translations from Non-parallel Corpora. Dans *Proceedings of the Workshop on Very Large Corpora*, pages 192–202, 1997. (Cité pages 28, 35, 36, 38, 39, 40, 42, 43 et 44.)

Jan Goes. *L'adjectif entre nom et verbe*. De Boeck and Larcier Département Duculot, 1999. (Cité pages 99 et 101.)

Lorriane Goeuriot. *Découverte et caractérisation des corpus comparables spécialisés*. PhD thesis, Université de Nantes, 2009. (Cité page 29.)

Tatiana Gornostay, Anita Gojun, Ulrich Heid, Emmanuel Morin, Rima Harastani, et Emmanuel Planas. Terminology Extraction from Comparable Corpora for Latvian. Dans *Proceedings of Baltic HLT'12*, pages 66–73, 2012. (Cité page 140.)

Gregory Grefenstette. Corpus-Derived First, Second and Third-Order Word Affinities. Dans *Proceedings of Euralex*, pages 279–290, 1994. (Cité pages 64 et 130.)

Gregory Grefenstette. The World Wide Web as a resource for example-based machine translation tasks. Dans *Translating and the Computer 21 : Proceedings of the International Conference on Translating and the Computer*, 1999. (Cité pages 47, 48, 49, 50, 51, 52, 53, 64 et 70.)

Clément De Groc. Babouk : Focused Web Crawling for Corpus Compilation and Automatic Terminology Extraction. Dans *The IEEE/WIC/ACM International Conferences on Web Intelligence*, pages 497–498, 2011. (Cité page 30.)

Benoît Habert. Des corpus représentatifs : de quoi, pour quoi, comment ? Dans *Linguistique sur corpus. Études et réflexions*, numéro 31 dans Cahiers de l'université de Perpignan, pages 11–58. Presses Universitaires de Perpignan, 2000. (Cité page 26.)

Thierry Hamon et Adeline Nazarenko. Detection of synonymy links between terms : experiment and results. *Recent Advances in Computational Terminology*, pages 185–208, 2001. (Cité page 25.)

Rima Harastani, Béatrice Daille, et Emmanuel Morin. Neoclassical Compound Alignments from Comparable Corpora. Dans *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'12)*, pages 72–82, 2012. (Cité page 140.)

Rima Harastani, Béatrice Daille, et Emmanuel Morin. Identification, alignement, et traductions des adjectifs relationnels en corpus comparables. Dans *Actes de la conférence sur le Traitement Automatique des Langues Naturelles (TALN'13)*, pages 313–326, 2013a. (Cité page 140.)

Rima Harastani, Béatrice Daille, et Emmanuel Morin. Ranking Translation Candidates Acquired from Comparable Corpora. Dans *Proceedings of the Sixth International Joint Conference on Natural Language Processing (IJCNLP'13)*, pages 401–409, 2013b. (Cité page 141.)

Bradely Hauer et Grzegorz Kondrak. Clustering Semantically Equivalent Words into Cognate Sets in Multilingual Lists. Dans *The International Joint Conference on Natural Language Processing (IJCNLP'11)*, pages 865–873, 2011. (Cité page 106.)

Amir Hazem et Emmanuel Morin. A Comparison of Smoothing Techniques for Bilingual Lexicon Extraction from Comparable Corpora. Dans *Proceedings of the Workshop on Building and Using Comparable Corpora (BUCC'13)*, pages 24–33, 2013a. (Cité page 121.)

Amir Hazem et Emmanuel Morin. Extraction de lexiques bilingues à partir de corpus comparables par combinaison de représentations contextuelles. Dans *Actes de la conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, pages 243–256, 2013b. (Cité page 46.)

Amir Hazem, Emmanuel Morin, et Sebastian Pena Saldarriaga. Bilingual Lexicon Extraction from Comparable Corpora As Metasearch. Dans *Proceedings of the Workshop on Building and Using Comparable Corpora : Comparable Corpora and the Web (BUCC'11)*, pages 35–43, 2011. (Cité page 131.)

Azniah Ismail et Suresh Manandhar. Bilingual lexicon extraction from comparable corpora using in-domain terms. Dans *Proceedings of the International Conference on Computational Linguistics (COLING'10) Posters*, pages 481–489, 2010. (Cité page 45.)

Sittichai Jiampojamarn, Grzegorz Kondrak, et Tarek Sherif. Applying Many-to-Many Alignments and Hidden Markov Models to Letter-to-Phoneme Conversion. Dans *Proceedings of the Conference on Human Language Technologies : The Conference of the North American Chapter of the Association for Computational Linguistics (HLT'07)*, pages 372–379, 2007. (Cité page 71.)

Nikhil S. Ketkar et G. Michael Youngblood. A Largest Common Subsequence-based Distance Measure for Classifying Player Motion Traces in Virtual Worlds. Dans *FLAIRS Conference*, 2010. (Cité page 105.)

Ahmad Khurshid, Davies Andrea, Fulford Heather, et Rogers Margaret. What is a term ? The semi-automatic extraction of terms from text. Dans *Translation Studies : An Interdiscipline (John Benjamins Publishing Company)*, pages 267–278, 1994. (Cité page 125.)

Adam Kilgarriff et Gregory Grefenstette. Web as corpus. Dans *Lancaster University*, pages 342–344, 2001. (Cité page 53.)

Philipp Koehn. Europarl : A parallel corpus for statistical machine translation. Dans *Proceedings of Machine Translation Summit*, pages 79–86, 2005. (Cité page 27.)

Philipp Koehn et Kevin Knight. Learning a Translation Lexicon from Monolingual Corpora. Dans *In Proceedings of ACL'02 Workshop on Unsupervised Lexical Acquisition*, pages 9–16, 2002. (Cité page 45.)

Harold W. Kuhn. The Hungarian Method for the Assignment Problem. *Naval Research Logistics Quarterly*, 2(1–2) :83–97, 1955. (Cité page 128.)

Audrey Laroche et Philippe Langlais. Revisiting context-based projection methods for term-translation spotting in comparable corpora. Dans *Proceedings of the International Conference on Computational Linguistics (COLING'10)*, pages 617–625, 2010. (Cité pages 44 et 121.)

Alise Lehmann et Françoise Martin-Berthet. *Introduction à la lexicologie*. Armand Colin, 2008. (Cité page 23.)

Marie-Claude L'Homme. *La terminologie : principes et techniques*. Les Presses de l'Université de Montréal, 2004. (Cité pages 24, 25 et 27.)

Bo Li et Éric Gaussier. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. Dans *Proceedings of the International Conference on Computational Linguistics (COLING'10)*, pages 644–652, 2010. (Cité pages 44, 60, 61, 62 et 133.)

Sa Liu et Chengzhi Zhang. Termhood-Based Comparability Metrics of Comparable Corpus in Special Domain. Dans *Proceedings of the Chinese conference on Chinese Lexical Semantics (CLSW'12)*, pages 134–144, 2012. (Cité page 62.)

Tie-Yan Liu. *Learning to Rank for Information Retrieval*. Springer, 2011. (Cité page 75.)

Elizaveta Loginova, Anita Gojun, Helena Blancafort, Marie Guégan, Tatiana Gornostay, et Ulrich Heid. Reference Lists for the Evaluation of Term Extraction Tools. Dans *Terminology and Knowledge Engineering (TKE'12)*, 2012. (Cité pages 30 et 131.)

Belinda Maia. What are comparable corpora. Dans *Proceedings of the Corpus Linguistics workshop on Multilingual Corpora : Linguistic requirements and technical perspectives*, 2003. (Cité page 28.)

François Maniez. Identification automatique des adjectifs relationnels : une étude sur corpus. *De la mesure dans les terme, Presses Universitaires de Lyon*, 2005. (Cité pages 99 et 100.)

Alexa McCray, Allen Browne, et Dorothy Moore. The Semantic Structure of Neo-Classical Compounds. Dans *Proceeding of the Symposium on Computer Applications in Medical Care (SCAMC'88)*, pages 165–168, 1988. (Cité page 77.)

Tony McEnery et Costas Gabrielatos. English corpus linguistics. *The Handbook of English Linguistics*, pages 33–71, 2006. (Cité page 26.)

Charles F. Meyer. *English Corpus Linguistics : An Introduction*. Cambridge University Press, 2002. (Cité page 26.)

Robert C. Moore. Improving IBM Word Alignment Model 1. Dans *Proceedings of the meeting of Association for Computational Linguistics (ACL'04)*, pages 518–525, 2004. (Cité page 58.)

Emmanuel Morin et Béatrice Daille. Compositionality and lexical alignment of multi-word terms. *Language Resources and Evaluation*, 44(1-2) : 79–95, 2010. (Cité pages 46, 54, 64, 97, 100 et 140.)

Emmanuel Morin et Béatrice Daille. Compositionnalité et contextes issus de corpus comparables pour la traduction terminologique. Dans *Actes de la conférence sur le Traitement Automatique des Langues Naturelles (TALN'12)*, pages 141–154, 2012a. (Cité page 83.)

Emmanuel Morin et Béatrice Daille. Revising the Compositional Method for Terminology Acquisition from Comparable Corpora. Dans *Proceedings of the International Conference on Computational linguistics (COLING'12)*, pages 1797–1810, 2012b. (Cité pages 55, 56, 64 et 83.)

Emmanuel Morin, Béatrice Daille, Koichi Takeuchi, et Kyo Kageura. Bilingual terminology mining - using brain, not brawn comparable corpora. Dans *Proceedings of the meeting of Association for Computational Linguistics (ACL'07)*, 2007. (Cité page 121.)

Emmanuel Morin, Béatrice Daille, Koichi Takeuchi, et Kyo Kageura. Brains, not brawn : The use of smart comparable corpora in bilingual terminology mining. volume 7, pages 1–23. ACM, 2008. (Cité pages 28, 29, 44 et 121.)

Emmanuel Morin et Emmanuel Prochasson. Bilingual Lexicon Extraction from Comparable Corpora Enhanced with Parallel Corpora. Dans *Proceedings of the Workshop on Building and Using Comparable Corpora : Comparable Corpora and the Web (BUCC'11)*, pages 27–34, 2011. (Cité page 131.)

Dragos Stefan Munteanu et Daniel Marcu. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31 :477–504, 2005. (Cité pages 11, 27, 29, 57, 58, 59, 60 et 128.)

Dragos Stefan Munteanu et Daniel Marcu. Extracting parallel subsentential fragments from non-parallel corpora. Dans *Proceedings of the Meeting of the Association for Computational Linguistics (ACL'06)*, pages 81–88, 2006. (Cité pages 29, 59 et 137.)

Fiametta Namer. *Morphologie, lexique et traitement automatique des langues*. Lavoisier, 2009. (Cité page 68.)

Fiammetta Namer. Automatiser l'analyse morphosémantique non affixale : le système DériF. *Cahiers de Grammaire*, 28 :31–48, 2003. (Cité pages 112 et 113.)

Fiammetta Namer et Robert H. Baud. Defining and relating biomedical terms : Towards a cross-language morphosemantics-based system. *I. J. Medical Informatics*, 76(2-3) :226–233, 2007. (Cité pages 74, 76 et 88.)

Michèle Noailly. *L'adjectif en français*. Editions Ophrys, 1999. (Cité page 99.)

Franz Josef Och et Hermann Ney. Improved statistical alignment models. Dans *Proceedings of Meeting of the Association for Computational Linguistics (ACL'oo)*, pages 440–447, 2000. (Cité page 59.)

Maeve Olohan. *Introducing Corpora in translation studies*. Routledge, 2004. (Cité page 27.)

Christopher Olston et Marc Najork. Web Crawling. *Foundations and Trends in Information Retrieval*, 4(3) :175–246, 2010. (Cité page 30.)

Pablo Gamallo Otero. Learning Bilingual Lexicons from Comparable English and Spanish Corpora. Dans *Proceedings of Machine Translation Summit*, pages 191–198, 2007. (Cité page 45.)

Jennifer Pearson. *Terms in context*. John Benjamins Publishing Company, 1998. (Cité page 24.)

Nuria Rodriguez Pedreira. De la grammaire traditionnelle à la morphologie dérivationnelle : retour sur l'adjectif de relation. *VERBA*, 29 :421–434, 2002. (Cité page 100.)

Emmanuel Prochasson. *Alignement multilingue en corpus comparables spécialisés*. PhD thesis, Université de Nantes, 2010. (Cité page 108.)

Emmanuel Prochasson et Pascale Fung. Rare Word Translation Extraction from Aligned Comparable Documents. Dans *Proceedings of the Meeting on Association for Computational Linguistics (ACL'11)*, pages 1327–1335, 2011. (Cité page 46.)

Emmanuel Prochasson et Emmanuel Morin. Anchor points for bilingual extraction from small specialized comparable corpora. *TAL*, 50(1) :283–304, 2009. (Cité page 45.)

Reinhard Rapp. Identifying word translations in non-parallel texts. Dans *Proceedings of the Meeting on Association for Computational Linguistics (ACL'95)*, pages 320–322, 1995. (Cité pages 12, 35, 36, 37, 38, 40, 42, 43, 44, 46 et 121.)

Reinhard Rapp. Automatic Identification of Word Translations from Unrelated English and German Corpora. Dans *Proceedings of the Meeting of the Association for Computational Linguistics on Computational Linguistics (ACL'99)*, pages 519–526, 1999. (Cité pages 35, 36, 40, 42, 43, 44, 63, 83, 121 et 127.)

Xavier Robitaille, Yasuhiro Sasaki, Masatsugu Tonoike, Satoshi Sato, et Takehito Utsuro. Compiling French-Japanese Terminologies from the Web. Dans *Proceedings of the Conference of the European chapter of the Association for Computational Linguistics (EACL'06)*, pages 225–232, 2006. (Cité pages 47, 49, 50, 51, 52, 53, 55 et 70.)

Michel Roché. Comment les adjectifs sont sémantiquement construits. *Cahier de Grammaire 30*, 2006. (Cité page 99.)

Matthew S. Ryan et Graham R. Nudd. The Viterbi Algorithm. Rapport technique, 1993. (Cité page 71.)

Fatiha Sadat, Masatoshi Yoshikawa, et Shunsuke Uemura. Learning bilingual translations from comparable corpora to cross-language information retrieval : hybrid statistics-based and linguistics-based approach. Dans *Proceedings of the International Workshop on Information Retrieval with Asian Languages (AsianIR'03)*, pages 57–64, 2003. (Cité page 45.)

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, et Dan Flickinger. Multiword Expressions : A Pain in the Neck for NLP. Dans *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'02)*, pages 1–15, 2001. (Cité pages 46 et 54.)

Juan C. Sager. *A Practical Course in Terminology Processing*. Amesterdam : John Benjamins Publisher Company, 1990. (Cité page 24.)

Xabier Saralegi, Inaki Vicente, et Antton Gurrutxaga. Automatic Extraction of Bilingual Terms from Comparable Corpora in a Popular Science Domain. Dans *Proceedings of the LREC'08 Workshop on Comparable Corpora*, 2008. (Cité page 60.)

Li Shao et Hwee Tou Ng. Mining new word translations from comparable corpora. Dans *Proceedings of the International Conference on Computational Linguistics (COLING'04)*, 2004. (Cité page 45.)

John Sinclaire. *Preliminary recommendations on corpus typology*. Expert Advisory Group on Language Engineering Standards (EAGLE), 2003. (Cité page 25.)

Jason R. Smith, Chris Quirk, et Kristina Toutanova. Extracting parallel sentences from comparable corpora using document level alignment. Dans *Proceedings of the Conference on Human Language Technologies (HLT'10)*, pages 403–411, 2010. (Cité page 60.)

Dan Stefanescu, Radu Ion, et Sabine Hunsicker. Hybrid Parallel Sentence Mining from Comparable Corpora. Dans *Proceedings of the Conference of the European Association for Machine Translation (EAMT'12)*, pages 137–144, 2012. (Cité pages 59 et 128.)

Oliver Streiter, Daniel Zielinski, Isabella Ties, et Leonhard Voltmer. Term Extraction for Ladin : An Example-based Approach. Dans *Actes de la conférence sur le Traitement Automatique des Langues Naturelles (TALN'03)*, 2003. (Cité pages 108 et 125.)

Fangzhong Su et Bogdan Babych. Measuring comparability of documents in non-parallel corpora for efficient extraction of (semi-)parallel translation equivalents. Dans *Proceedings of the Joint EACL'12 Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, pages 10–19, 2012. (Cité pages 61 et 62.)

Takaaki Tanaka et Timothy Baldwin. Translation Selection for Japanese-English Noun-Noun Compounds. Dans *Proceedings of the ACL'03 Workshop on Multiword Expressions : Analysis Acquisition and Treatment*, pages 17–24, 2003. (Cité pages 47, 48, 50, 51, 52 et 53.)

Charlotte Taylor. What is corpus linguistics ? What the data says. *ICAME Journal*, pages 179–200, 2008. (Cité page 26.)

Trimble. *The Development of EFL Materials for Occupational English : The Technical Manual*. Trimble, 1978. (Cité page 24.)

Mark Tuttle, David Sherertza, et Nels Olsona. Using META-1, the first version of the UMLS Metathesaurus. Dans *Proceedings of the Symposium on Computer Applications in Medical Care (SCAMC'90)*, pages 131–135, 1990. (Cité pages 71 et 73.)

Spela Vintar. Bilingual term recognition revisited the bag-of-equivalents term alignment approach and its evaluation. *Terminology*, 16 :141–158, 2010. (Cité pages 47, 49, 51, 52 et 53.)

Marion Weller, Anita Gojun, Ulrich Heid, Béatrice Daille, et Rima Harastani. Simple methods for dealing with term variation and term alignment. Dans *Proceedings of the Conference of Terminology and Artificial Intelligence (TIA'11)*, 2011. (Cité pages 133 et 140.)

Geoffrey C. Williams. *Les réseaux collocationnels dans la construction et l'exploitation d'un corpus dans le cadre d'une communauté de discours scientifique*. PhD thesis, Université de Nantes, 1999. (Cité page 29.)

Pierre Zweigenbaum et Benoît Habert. Les corpus naissent tous comparables en droit : apports méthodologiques de l'acquisition lexicale en contexte multilingue. *Glottopol*, (8) :22–44, 2006. (Cité page 28.)

**Titre** Alignement lexical en corpus comparables : le cas des composés savants et des adjectifs relationnels

**Résumé** Notre travail concerne l'extraction automatique d'une liste de termes alignés avec leurs traductions (c'est-à-dire un lexique bilingue spécialisé) à partir d'un corpus comparable dans un domaine de spécialité. Un corpus comparable comprend des textes écrits dans deux langues différentes sans aucune relation de traduction entre eux mais dont les textes appartiennent à un même domaine. Les contributions de cette thèse portent sur l'amélioration de la qualité d'un lexique bilingue spécialisé extrait à partir d'un corpus comparable. Nous proposons des méthodes consacrées à la traduction de deux types de termes, qui ont des caractéristiques en commun entre plusieurs langues ou qui posent par leur nature des problèmes pour la traduction : les composés savants (termes contenant au moins une racine gréco-latine) et les termes composés d'un nom et un adjectif relationnel. Nous développons également une méthode, qui exploite des contextes riches en termes spécifiques au domaine du corpus, pour réordonner dans un lexique bilingue spécialisé des traductions candidates fournies pour un terme. Les expériences sont réalisées en utilisant deux corpus comparables spécialisés (dans les domaines du cancer du sein et des énergies renouvelables), sur les langues français, anglais, allemand et espagnol.

**Mots-clés** corpus comparables, langue de spécialité, alignement multilingue, composés savants, adjectifs relationnels.

**Title** Lexical Alignment from Comparable Corpora : The case of Neoclassical Compounds and Relational Adjectives

**Abstract** Our work concerns the automatic extraction of a list of aligned terms with their translations (i.e. specialized bilingual lexicon) from comparable corpora belonging to a specific domain. Comparable corpora include texts written in two languages which are not mutual translations but belong to the same domain. This thesis contributes to the improvement of the quality of an extracted bilingual lexicon. We propose methods dedicated to the translation of two types of terms that have common characteristics among many languages or that cause specific problems for translation due to their nature. These types of terms are the neoclassical compounds (terms containing at least one root borrowed from Greek or Latin) and the terms composed of one noun and one relational adjective. We also propose a method that exploits contexts rich in domain-specific terms to re-rank some provided translations in a bilingual lexicon for a given term. The experiments are performed using two specialized comparable corpora (in the domains of Breast Cancer and Renewable Energy), on the French, English, German and Spanish languages.

**Keywords** comparable corpora, specialized language, multilingual alignment, neoclassical compounds, relational adjectives.