Traduction assistée par ordinateur et corpus comparables : contributions à la traduction compositionnelle

Estelle Delpech

UNIVERSITÉ DE NANTES
FACULTÉ DES SCIENCES ET DES TECHNIQUES

ÉCOLE DOCTORALE STIM
« **Sciences et Technologies de l'Information et Mathématiques** »

Année 2013

N° attribué par la bibliothèque

Traduction assistée par ordinateur et corpus comparables

Contributions à la traduction compositionnelle

THÈSE DE DOCTORAT

Discipline : Informatique

Spécialité : Traitement Automatique des Langues

Présentée
et soutenue publiquement par

Estelle DELPECH

Le 2 juillet 2013, devant le jury ci-dessous

Président	Nabil Hathout, Directeur de recherche, Centre National de la Recherche Scientifique
Rapporteurs	Élisabeth Lavault-Olléon, Professeure, Université Stendhal Grenoble 3
	Michel Simard, Agent de recherche principal, Conseil National de Recherches du Canada
Examinateurs	Béatrice Daille, Professeure, Université de Nantes
	Nabil Hathout, Directeur de recherche, Centre National de la Recherche Scientifique
	Emmanuel Morin, Professeur, Université de Nantes
Invité	Emmanuel Planas, Maître de Conférences, Université Catholique de l'Ouest

Directrice de thèse : Prof. Béatrice Daille
Co-encadrant de thèse : Prof. Emmanuel Morin

À Élia


DID YOU KNOW THAT SUBURBAN WHITE MALES HAVE OVER 100 WORDS FOR "LAWN"?
7-29
©2009 BEVERLY
SPEEDBUMP.COM
DIST. BY CREATORS SYND. INC.

Remerciements

Je tiens à remercier de tout cœur Béatrice Daille et Emmanuel Morin d'avoir respectivement dirigé et co-encadré ce travail de thèse. J'ai été trés honorée de travailler et d'apprendre à leurs côtés. Ils ont tous deux fait preuve d'un savant mélange d'exigence académique et de pédagogie qui m'a permis de progresser durant ces trois années. Je les remercie tous deux de m'avoir proposé un sujet de thèse si intéressant et d'avoir su se rendre disponibles malgré leurs emplois du temps chargés.

Je remercie chaleureusement Nabil Hathout, Élisabeth Lavault-Olléon, Emmanuel Planas et Michel Simard de m'avoir fait l'honneur d'être membres de mon jury. Leurs remarques constructives m'ont été particulièrement utiles. Je suis heureuse d'avoir pu bénéficier de points de vue autant complémentaires sur mon travail. Merci spécialement à Michel Simard d'avoir fait le déplacement jusqu'à Nantes depuis le Canada !

Je suis particulièrement reconnaissante envers Emmanuel Planas, ancien directeur scientifique de Lingua et Machina, pour m'avoir fait confiance et embauchée comme ingénieure de recherche. Sans cela, je n'aurais très probablement pas eu l'opportunité d'effectuer une thèse au LINA ni de travailler sur un sujet de recherche dans un cadre industriel aussi stimulant.

Plusieurs personnes ont contribué, de près ou de loin, au travail présenté dans ce document. Je remercie en premier lieu Claire Lemaire de l'Université Stendhal de Grenoble, d'abord parce qu'elle a été une collègue et co-thésarde formidable ; ensuite pour son travail de qualité concernant la création des ressources pour le traitement et l'évaluation de l'allemand. Cela n'aurait pas été possible sans elle et je lui en suis très reconnaissante.

Je remercie aussi Geoffrey Williams et Pierre Zweigenbaum d'avoir accepté d'être membres de mon comité de suivi de thèse. Leurs retours et conseils avisés m'ont guidé tout au long de ce travail.

Mes remerciements vont également à Léa Laporte de l'Institut de Recherche en Informatique de Toulouse et Damien François de l'Université Catholique de Louvain pour avoir répondu à mes questions concernant le traitement statistique des données. Merci aussi à Van Dang, de l'Université du Massachusetts, pour avoir répondu à mes questions quant à l'utilisation des algorithmes de *learning-to-rank*.

J'ai beaucoup de gratitude envers Clémence de Baudus, Kiril Isakov, Mathieu Delage de l'Institut Supérieur de Traduction et d'Interprétation et Nicolas Auger qui ont effectué un minutieux travail d'annotation, ce qui a rendu possible l'évaluation du système de traduction.

J'ai une pensée pour mes collègues de Lingua et Machina, François, Étienne et Jean-François, auprès de qui j'ai beaucoup appris et que je remercie pour leurs encouragements. Les conseils et l'expérience de François m'ont été précieux pour ma dernière année de thèse.

Je n'ai malheureusement pas beaucoup eu l'occasion d'être présente au laboratoire mais

cela a toujours été un plaisir de venir aux réunions d'équipe. L'accueil et l'ambiance du LINA est formidable et j'ai beaucoup apprécié de discuter avec mes collègues, notamment Amir Hazem et Prajol Shrestha qui ont été d'agréables camarades de thèse.

Enfin, je remercie mon compagnon Nicolas pour son soutien sans faille ; mes amies Émilie et Nathalie et ma sœur Laureen pour leur compréhension quant à mon manque de disponibilité et pour leur présence et leur soutien logistique le jour de la soutenance. Merci à Loki qui est un formidable réveille-matin.

Traduction assistée par ordinateur et corpus comparables

Contributions à la traduction compositionnelle [1]

Estelle Delpech

Version finale 23 février 2014

1. Travail financé par l'Agence Nationale de la Recherche (subvention ANR-08-CORD-013), l'Association Nationale de la Recherche et de la Technologie (convention CIFRE nº 2010/270) et la société LINGUA ET MACHINA.

Résumé

Notre travail concerne l'extraction de lexiques bilingues à partir de corpus comparables, avec une application à la traduction spécialisée. Nous avons d'abord évalué les méthodes classiques d'acquisition de lexiques en corpus comparables (basées l'hypothèse distributionnelle : plus deux termes apparaissent dans des contextes similaires, plus il y a de chances qu'ils soient des traductions) d'un point de vue applicatif.

L'évaluation a montré que les traducteurs sont mal à l'aise avec les lexiques extraits : la traduction correcte est trop souvent noyée dans une liste de traductions candidates et ils préfèreraient utiliser un lexique plus petit mais plus précis. Partant de ce constat, nous nous sommes orientés vers une autre approche qui a fait récemment ses preuves pour l'exploitation des corpus comparables et produit des lexiques plus adaptés aux besoins des traducteurs : la traduction compositionnelle (la traduction du terme source est fonction de la traduction de ses parties).

Nous nous sommes concentrés sur la traduction d'unités monolexicales : le terme source est découpé en morphèmes, les morphèmes sont traduits puis recomposés en un terme cible. Dans ce cadre, nous avons poursuivi trois axes de recherche : la génération de traductions fertiles (cas où le terme cible contient plus de mots lexicaux que le terme source), l'indépendance aux structures morphologiques et l'ordonnancement des traductions candidates.

Mots-clés : traduction assistée par ordinateur, corpus comparables, compositionnalité, *learning-to-rank*, évaluation centrée utilisateur, morphologie computationnelle

Abstract

Our work deals with the extraction of bilingual lexicons from comparable corpora with an application to specialized translation. We started by evaluating classical methods based on the distributional hypothesis (the more two terms appear in similar contexts, the more likely they are translations of each other) in a user-oriented fashion.

This evaluation raised the fact that translators feel very uncomfortable with this kind of lexicon: they feel correct translations are uneasy to spot in the lists of candidate translations and would rather use a smaller lexicon but with higher precision rates. Based on this observation, we turned to another approach for term translation which has been recently and successfully experimented on comparable corpora and produce lexicons that meet the demands of the translators: compositional translation. In this framework, the translation of a term is composed of the translation of its parts.

We concentrated on the translation of monolexical terms: the source term is decomposed into morphemes, morphemes are translated into the target language and recomposed as a target term. We investigated three lines of research: generation of fertile translations (cases in which the target term has more lexical words than the source term), independence to morphological structure and candidate translation ranking.

Keywords : computer-aided translation, comparable corpora, compositionality, learning-to-rank, user-centered evaluation, computational morphology

Liste des publications et communications

Extraction automatique de lexiques bilingues

E. DELPECH, B. DAILLE, E. MORIN et C. LEMAIRE : Extraction of domain-specific bilingual lexicon from comparable corpora : compositional translation and ranking. *In Proceedings of the 24th International Conference on Computational Linguistics (Long papers)*, pages 745-762, Mumbai, Inde, 2012.

E. DELPECH, B. DAILLE, E. MORIN et C. LEMAIRE : Identification of Fertile Translations in Medical Comparable Corpora : a Morpho-Compositional Approach. *In Proceedings of the 10th biennial conference of the Association for Machine Translation in the Americas*, 10 pages, San Diego, États-Unis d'Amérique, 2012.

E. DELPECH : Bilingual terminology mining. *The 4th Intensive Summer school and collaborative workshop on Natural Language Processing (Franco-Thai Workshop 2010)*, Bangkok, Thaïlande, 2010.

Évaluation applicative

E. DELPECH : Un Protocole d'Évaluation Applicative des Terminologies Bilingues Destinées à la Traduction Spécialisée. *Revue des Nouvelles Technologies de l'Information (RNTI) - Numéro spécial : Qualité des Données et des Connaissances / Evaluation des méthodes d'Extraction de Connaissances dans les Données (Eval'ECD)*, pages 23-48, 2011.

E. DELPECH : Evaluation of terminologies acquired from comparable corpora : an application perspective. *In Proceedings of the 18th Nordic Conference of Computational Linguistics*, pages 66-73, Riga, Lettonie, 2011.

E. DELPECH : Un Protocole d'Évaluation Applicative des Terminologies Bilingues Destinées à la Traduction Spécialisée. *In Actes de l'atelier Évaluation des méthodes d'Extraction de Connaissances dans les Données (EvalECD'11) – 11ème Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances EGC*, pages 37-48, Brest, France, 2011.

Outils de Traduction Assistée par Ordinateur

F. BROWN DE COLSTOUN, E. DELPECH et E. MONNERET : Libellex : une plateforme multiservices pour la gestion des contenus multilingues. *In Actes de la 18ème conférences sur le traitement automatique des langues naturelles (démonstrations logicielles)*, page 319, Montpellier, France, 2011.

E. DELPECH et B. DAILLE : Dealing with lexicon acquired from comparable corpora : validation and exchange. *In Proceedings of the 2010 Terminology and Knowledge Engineering Conference*, pages 211-223, Dublin, Irlande, 2010.

F. BROWN DE COLSTOUN et E. DELPECH : Libellex, environnement de gestion collaborative en ligne de terminologie au sein de communautés fermées. *Terminologie & Ontologie : Théories et applications (TOTh)*, Annecy, France, 2010.

Table des matières

Introduction

Enjeux socio-économiques de la gestion du multilinguisme

À l'heure de la globalisation des échanges, le multilinguisme, bien qu'étant une richesse socio-culturelle indéniable, pose de nombreux défis à notre société.

Tout d'abord, la non-connaissance d'une langue est souvent synonyme d'un accès limité à l'information et ce sont généralement les communautés linguistiques avec un faible pouvoir économique ou dont la langue manque de prestige qui souffrent de cette discrimination.

Le cas d'Internet est exemplaire : par exemple, l'anglais, langue la plus représentée sur Internet (54,8 %)[1], n'est la langue première que de 26,8% des internautes[2] alors que le chinois, langue première de 24,2% des internautes, n'est que 6ème en terme de présence sur Internet (4 %). Une vaste partie de l'information présente sur Internet reste donc inaccessible aux internautes du fait de la barrière des langues.

Dans les pays officiellement bi- ou multi- lingues ou dans les organisations supranationales comme l'Union Européenne, la gestion du multilinguisme revêt une dimension démocratique : il s'agit de garantir à chaque citoyen l'accès aux services administratifs et aux textes legislatifs dans sa langue première afin qu'il ait connaissance de ses droits et puisse bénéficier des services de l'État dans une langue qu'il maîtrise. Ceci à un coût non négligeable : la Communauté Européenne dépense chaque année 1 milliard d'euros en coûts de traduction et d'interprétation (Fidrmuc, 2011).

Le multilinguisme a aussi un impact sur notre économie : le rapport ELAN (Hagen *et al.*, 2006) estimait en 2006 que le manque de compétences linguistiques avait fait perdre, sur une période de trois ans, une moyenne de 325 000 euros par PME européenne.

Pour répondre à ce coût social et économique, des recherches se sont développées dans le but d'accélerer et d'améliorer le processus de traduction humaine. Aujourd'hui, il existe toute une industrie dédiée à cette problématique. L'industrie des langues offre à la fois des services de traduction humaine mais aussi toute une palette de logiciels visant à réduire les coûts de traduction : mémoires de traduction, logiciels d'extraction et de gestion de terminologies bilingues, logiciels de localisation, etc. C'est dans ce cadre de recherche et développement en *traduction assistée par ordinateur* que s'est inscrit notre travail de thèse, travail qui a été partiellement financé par la société LINGUA ET MACHINA[3], spécialiste de la gestion des contenus multilingues en entreprise, et par le projet ANR METRICC[4], dédié à l'exploitation des

1. En Mai 2011, d'après WEB TECHNOLOGY SURVEYS
http://w3techs.com/technologies/overview/content_language/all
2. http ://www.internetworldstats.com/stats7.htm
3. http://www.lingua-et-machina.com
4. http://www.metricc.com

corpus comparables.

Motivations et objectifs

La traduction assistée par ordinateur (TAO) a, depuis toujours, eu recours à des historiques de traductions : elle nécessite que le traducteur ait à sa disposition un ensemble de traductions passées sur lesquelles le logiciel de TAO va pouvoir s'appuyer pour produire, par exemple, des lexiques bilingues. Cet état de fait est problématique lorsque le traducteur ne dispose pas d'un tel corpus. Cette situation se produit lorsque les textes à traduire appartiennent à un domaine émergent ou à un couple de langues peu doté. Pour répondre à cette problématique, les recherches en traduction assistée par ordinateur se sont orientées vers l'exploitation de *corpus comparables*, c'est-à-dire un ensemble de textes, dans deux ou plusieurs langues, qui traitent d'une même thématique mais ne sont pas des traductions les uns des autres.

Les corpus comparables font l'objet de recherches académiques depuis les années 90 (Fung, 1995; Rapp, 1999) et l'existence du *Workshop on Building and Using Comparable Corpora (BUCC)*, organisé annuellement depuis 2008 en périphérie de grandes conférences, montre le dynamisme de cette thématique de recherche.

Les recherches actuelles visent principalement à extraire des paires de termes ou de phrases alignés qui sont ensuite utilisés dans des systèmes de recherche d'information cross-lingue (Renders *et al.*, 2003; Chiao, 2004; Li *et al.*, 2011) ou dans des systèmes de traduction automatique (Rauf et Schwenk, 2009; Carpuat *et al.*, 2012). Si la traduction assistée par ordinateur est souvent évoquée comme un potentiel domaine applicatif, l'apport des corpus comparables n'a, à notre connaissance, pas encore été réellement étudié dans ce cadre d'application. Or, ceci soulève de nombreuses problématiques comme le passage à l'échelle ou encore l'adaptation aux besoins des utilisateurs finaux.

Notre thèse poursuit deux objectifs. Le premier sera d'évaluer l'apport des lexiques extraits de corpus comparables dans le cadre d'une tâche de traduction spécialisée. Nous nous soucierons de mettre au jour les besoins des traducteurs et de comprendre comment les corpus comparables peuvent être exploités au mieux pour la traduction assistée par ordinateur.

Notre second objectif sera d'identifier des méthodes d'extraction de lexiques bilingues qui répondent au mieux aux besoins des traducteurs. Nous tâcherons de déterminer les limites actuelles de ces techniques et d'en proposer des améliorations. Nous nous pencherons notamment sur l'identification de traductions fertiles (cas où le terme cible contient plus de mots que le terme source), la gestion de multiples structures morphologiques et l'ordonnancement de traductions candidates (les algorithmes proposent généralement plusieurs traductions candidates pour un même terme source).

Nos expériences seront menées sur deux couples de langues (anglais-français et anglais-allemand) et sur des textes spécialisés traitant du cancer du sein. Notre travail possèdera une dimension applicative forte et nos choix méthodologiques seront guidés par les besoins des utilisateurs finaux.

Plan de lecture

La présente thèse est organisée en deux parties :

La première partie pose le contexte applicatif et scientifique de nos recherches. **Dans le**

premier chapitre, nous faisons un retour historique sur les débuts de la traduction automatique et montrons comment les recherches se sont peu à peu orientées vers la traduction assistée par ordinateur puis l'exploitation des corpus comparables. Nous présentons les techniques actuelles d'extraction de lexiques bilingues et détaillons la façon dont nous avons prototypé un outil de TAO destiné à l'exploitation de corpus comparables. **Le deuxième chapitre** est consacré à l'évaluation applicative de cet outil : nous observons dans quelle mesure les lexiques extraits permettent aux traducteurs d'être plus efficaces dans leur travail. Cette évaluation met au jour des besoins spécifiques à la traduction humaine qui ne sont pas traités par les techniques classiques d'alignement de termes. C'est pourquoi nous nous orientons vers un autre type de méthode qui vise à *générer* des traductions de termes qui peuvent ensuite être filtrées grâce au corpus plutôt qu'à *aligner* des termes préalablement extraits des corpus. Ces techniques sont détaillées **dans le chapitre 3**. Dans ce chapitre, nous nous intéressons principalement aux approches dites *compositionnelles*. Nous en détaillons les limites et concluons cette première partie sur plusieurs perspectives de recherche.

La seconde partie de la thèse est consacrée à la recherche d'améliorations de la traduction compositionnelle. **Le chapitre 4** présente le cadre méthodologique de nos recherches : nous décrivons le principe de notre approche et tentons de mettre en avant nos contributions à la traduction compositionnelle (fertilité, variété des structures morphologiques traitées, ordonnancement des traductions candidates). Nous exposons également notre méthodologie d'évaluation. **Le chapitre 5** décrit les données avec lesquelles nous avons expérimenté notre méthode de traduction : origine, nature, taille, mode d'acquisition. **Le chapitre 6** donne les détails de notre implémentation : nous y donnons notre algorithme de génération de traduction. La méthode de génération de traduction est ensuite évaluée sous plusieurs angles (apport des ressources, apport des stratégies de traduction, des traductions fertiles...). Enfin, **le chapitre 7** formalise et expérimente plusieurs méthodes d'ordonnancement des traductions générées.

Nous concluons la présente thèse en faisant le bilan de nos travaux et en proposant plusieurs perspectives de recherche. Les annexes comprennent un index des mesures employées tout au long du document ainsi que des extraits des données expérimentales.

Première partie

Contexte scientifique et applicatif

Chapitre 1

Exploitation des corpus comparables pour la traduction assistée par ordinateur

Sommaire

Introduction

Ce chapitre débute par une mise en perspective historique de la traduction assistée par ordinateur (1.1) : nous retraçons les débuts de la traduction automatique et expliquons comment la traduction assistée par ordinateur s'est développée jusqu'à ce que récemment se pose la question de l'exploitation des corpus comparables. La section 1.2 explique les techniques actuelles d'extraction de lexiques bilingues à partir de corpus comparables. Nous donnons un aperçu des performances qui peuvent en être attendues et nous en discutons les limites. La dernière section (1.3) décrit le prototypage d'outil de TAO destiné aux corpus comparables et basé sur les techniques décrites en 1.2.

1.1 Perspective historique : des premiers traducteurs automatiques aux corpus comparables

1.1.1 Premières recherches en traduction automatique

Dès les débuts de l'informatique, les recherches scientifiques ont cherché à exploiter la machine pour accélérer voire remplacer la traduction humaine. D'après Hutchins (2005), c'est aux États-Unis, entre 1959 et 1966, que les premières recherches en traduction automatique - c'est-à-dire la traduction d'un texte par une machine sans intervention humaine - ont été menées. Jusqu'en 1966, divers groupes de recherches se forment et deux familles d'approches se distinguent déjà :

- D'un côté, se trouvaient des approches pragmatiques mêlant informations statistiques et méthode de développement par essai-erreur [1] et dont le but était de produire au plus vite un système opérationnel (Université de Washington, Rand Corporation, Université de Georgetown). Ces recherches appliquaient la méthode de la traduction directe [2] et donnèrent lieu à la première génération de traducteurs automatiques.
- De l'autre, s'étaient constituées des approches théoriques impliquant la linguistique fondamentale et envisageant des recherches sur le long terme (MIT, Cambridge Research Language Unit). Ces projets plus théoriques mirent au point les premières versions des systèmes interlingues [3] et de traduction par transfert [4].

En 1966, un rapport de l'Automatic Language Processing Advisory Commitee (ALPAC, 1966), qui évalue la traduction automatique uniquement à l'aune des besoins du gouvernement américain – c'est-à-dire la traduction de documents scientifiques russes – annonce qu'après

1. Plusieurs règles heuristiques sont implantées et testées sur les données jusqu'à l'obtention d'un résultat jugé satisfaisant.

2. Il s'agit d'une stratégie de traduction n'impliquant aucune couche de traitement intermédiaire : les tout premiers traducteurs utilisant cette approche faisaient un découpage en mots, une neutralisation des flexions, cherchaient la traduction des mots dans un dictionnaire bilingue puis les mots traduits étaient réordonnés à l'aide de quelques règles. Il n'y avait donc aucune analyse syntaxique ou sémantique.

3. La méthode interlingue analyse le texte source de façon à en produire une représentation sémantique abstraite, indépendante des langues. Le texte cible est ensuite généré à partir de cette représentation. Le module de génération en langue cible a uniquement accès à la représentation interlingue.

4. La traduction par transfert utilise des représentations intermédiaires. Contrairement à l'approche interlingue, les réprésentations sont propres au couples de langues en jeu.

plusieurs années de recherche, il n'est pas possible d'obtenir une traduction totalement automatique et de qualité humaine. Seule la post-édition permet d'obtenir des traductions de bonne qualité [5]. Or, l'intérêt de la post-édition n'est pas évident. Une étude décrite en annexe du rapport indique que la plupart des traducteurs soumis à des travaux de post-édition trouvent le processus laborieux et frustrant *mais* remarquent que les sorties du système sont utiles en tant qu'*aide à la traduction*, en particulier en ce qui concerne les termes techniques [6].

Bien que l'étude ne permette pas de décider de l'intérêt de la post-édition par rapport à la traduction 100 % manuelle (sur vingt-deux traducteurs, huit trouvent la post-édition plus facile, huit autres la trouvent plus difficile, et six ne tranchent pas), le rapport met surtout en avant les aspects négatifs, citant un des traducteurs :

> ≪ *I found that I spend at least as much time in editing as if I had carried out the entire translation from the start. Even at that, I doubted if the edited translation reads as smoothly as one which I would have started from scratch* ≫ [7] cité par Hutchins (1996, p. 13)

Le rapport s'appuie sur les propos de V. Yngve, directeur des projets de recherche en traduction automatique au MIT, pour affirmer que la traduction automatique est inutile sans la post-édition et qu'avec la post-édition, le processus devient trop lent et perd tout intérêt économique [8].

Le rapport conclut sur le fait que, si les recherches en traduction automatique étaient essentielles du strict point de vue de l'avancée scientifique, elles ne présentaient par contre aucun intérêt du point de vue économique. En conséquence, les financements furent arrêtés aux États-Unis. La recherche continua malgré tout en Europe (projet de recherche EUROTRA) et au Canada. Ces recherches firent par exemple émerger le système TAUM (traduction des bulletins météo du français vers l'anglais) et le logiciel de traduction SYSTRAN.

1.1.2 Le développement de l'aide à la traduction

Alors qu'il a signé la fin du financement public de la recherche en traduction automatique aux États-Unis, le rapport ALPAC a encouragé la poursuite d'un objectif plus réaliste, celui de la traduction assistée par ordinateur [9]. Le rapport encensait les glossaires produits par l'Agence de traduction de l'armée allemande ainsi que la base terminologique de la Communauté européenne du charbon et de l'acier - ressource précurseure d'EURODICAUTOM et de IATE - et concluait que ces ressources constituaient une réelle aide à la traduction. Les recommandations finales encouragaient clairement le développement de la traduction assistée par ordinateur, notamment l'exploitation des glossaires initialement créés pour la traduction automatique [10].

5. ≪ *MT "presumably means going by algorithm from machine-readable source text to useful target text, without recourse to human translation or editing"* ≫ - cité par Hutchins (1996, p. 11)

6. ≪ *Most translators "found postediting tedious and even frustrating", but many found "the output served as an aid... particularly with regard to technical terms"* ≫ cité par Hutchins (1996, p. 13)

7. ≪ *Finalement, j'ai passé autant de temps à éditer que si j'avais fait la traduction en entier depuis le début. Je ne suis même pas sûr(e) que la traduction éditée ait un rendu aussi naturel que si j'avais effectué la traduction directement.* ≫ (notre traduction).

8. ≪ *it quote Victor Yngve, head of the MT project at MIT that MT "serves no useful purpose without postediting, and that with postediting the over-all process is slow and probably uneconomical"* ≫ cité par Hutchins (1996, p. 12)

9. ≪ *Machine-aided translation may be an important avenue toward better, quicker and cheaper translation* ≫ cité par Hutchins (1996, p. 14)

10. ≪ *research should be supported on :* [...] *2. means for speeding up the human translation process ;* [...] *6. evaluation of the relative speed and costs of various sorts of machine-aided translation ; 7. adaptation of existing mechanized editing and production processes in translation ;* [...] *9. production of adequate reference works for the translator, including the adaptation of glossaries that now exist primarily for automatic dictionary look-up in machine*

Se développe alors tout un panel d'outils destinés à assister le traducteur dans son travail et non à le remplacer. Les premiers programmes de gestion terminologique voient le jour dans les années 60 (Hutchins, 2005) et évoluent vers des banques terminologiques multilingues comme TERMIUM ou UNTERM. Les concordanciers bilingues sont également une aide précieuse : ils permettent d'accéder aux contextes d'un mot ou d'un terme et mettent en regard la traduction de ses contextes dans la langue cible. Selon Somers (2005), l'essor de la traduction assistée par ordinateur se produit dans les années 70 avec la création des logiciels de *mémoires de traduction* qui permettent de recycler les traductions passées : lorsqu'un traducteur doit traduire une nouvelle phrase, le logiciel parcourt la mémoire à la recherche de phrases similaires ayant déjà été traduites et, le cas échéant, propose la traduction passée comme modèle de traduction. Le gain de temps est d'autant plus grand que les textes à traduire sont redondants, ce qui est fréquemment le cas avec certains documents spécialisés comme les manuels techniques.

Ces ensembles de documents traduits constituent ce que l'on appelle des *corpus parallèles*[11] (Véronis, 2000) et leur exploitation s'intensifie dans les années 80, permettant un retour en force de la traduction automatique. Alors que les systèmes de traduction à base de règles avaient dominé le domaine jusque là, l'accès à de larges bases d'exemples de traductions permet de développer des systèmes fondés sur les données. Les deux paradigmes issus de ce tournant sont la traduction par l'exemple (Nagao, 1984) et la traduction automatique statistique (Brown *et al.*, 1990) qui reste le courant dominant actuel. La qualité de la traduction automatique s'améliore. Aujourd'hui, elle donne des résultats exploitables dans les domaines spécialisés où le vocabulaire et les structures sont assez répétitifs. Le dernier bastion concerne les textes tout-venant : la traduction automatique offre, au mieux, une aide à la compréhension.

Durant les années 90, la traduction assistée par ordinateur bénéficie des apports croisés de la traduction automatique et de la terminologie computationelle (Bourigault, 1994; Daille, 1994; Enguerard et Pantera, 1995; Jacquemin, 1996). Apparaissent alors des algorithmes d'*alignement de termes* à partir de corpus parallèles (Daille *et al.*, 1994; Melamed, 1999; Gaussier *et al.*, 2000). Les listes terminologiques bilingues produites sont particulièrement utiles dans le cas de la traduction spécialisée.

Gestion et extraction automatique de terminologie, concordanciers bilingues, pré-traduction et mémoires de traduction, aide à la compréhension : aujourd'hui, le poste de travail du traducteur est un environnement complexe et fortement informatisé. L'industrie de l'aide à la traduction s'est fortement développée, donnant lieu à la création de nombreux logiciels de TAO (Traduction Assistée par Ordinateur) : TRADOS[12], WORDFAST[13], DÉJÀ VU[14], SIMILIS[15] pour en citer quelques uns. Le grand public n'est pas en reste : d'une part, Google a largement démocratisé la traduction automatique tout-venant grâce à son outil GOOGLE TRANSLATE[16] et d'autre part, des concordanciers bilingues libres d'accès ont vu le jour assez récemment sur Internet (BAB.LA[17], LINGUEE[18]) et ont très vite gagné en popularité - LINGUEE, par exemple, totalisait, un an après sa fondation en 2008, 600 000 requêtes par jour pour sa version anglais-allemand (Perez, 2010).

translation ≫ cité par Hutchins (1996, p. 14)

11. ≪ *texts accompanied by their translation in one or more languages* ≫ (Véronis, 2000, p. 1) (*ensemble de textes accompagnés de leurs traductions dans une ou plusieurs langues*, notre traduction).
12. www.trados.com
13. www.wordfast.com
14. www.atril.com
15. www.lingua-et-machina.com
16. www.translate.google.com
17. www.en.bab.la
18. www.linguee.com

1.1.3 Limites des corpus parallèles et avantages des corpus comparables

Toutes utiles qu'elles soient, ces technologies ont une limite majeure : elles nécessitent l'existence d'un historique de traduction. Que faire dans le cas de langues peu dotées ou lorsque l'on aborde des domaines de spécialité émergents ? Une solution est alors d'avoir recours à ce que l'on appelle des *corpus comparables*.

Plusieurs définitions des corpus comparables existent. À un premier extrême, se trouve la définition, très stricte, donnée par Mc Enery et Xiao (2007) dans le cadre de recherches en traductologie. Selon ces auteurs, un corpus comparable contient des textes dans deux langues ou plus collectés selon les mêmes critères de genre, domaine et période de production. De plus, les corpus doivent être équilibrés [19]. À un autre extrême, nous rencontrons la définition de Déjean et Gaussier (2002), donnée dans le cadre de recherches en traitement automatique des langues, qui soulignent uniquement le fait qu'il doive exister ≪ *une sous-partie non négligeable* ≫ de vocabulaire en commun entre les textes [20].

Pour notre part, nous adoptons une position intermédiaire, considérant comme comparables des ensembles de textes dans deux langues ou plus qui traitent d'une même thématique et qui, si possible, ont été produits dans une même situation de communication si bien qu'il existe une possibilité d'y trouver des traductions utiles pour l'aide à la traduction. Nous nous intéressons uniquement aux corpus comparables *spécialisés*, c'est-à-dire que les textes ont été produits par un expert du domaine à destination d'autres experts ou du grand public (Bowker et Pearson, 2002).

En plus d'être plus facilement disponibles, les corpus comparables présentent également un intérêt qualitatif largement souligné par les traductologues. Les corpus parallèles sont connus pour leur non-fidélité aux usages linguistiques de la langue cible. Pour Mc Enery et Xiao (2007), les traductions sont au mieux une variante particulière et non représentative de la langue cible [21]. Pour Zanettin (1998), les textes traduits ne peuvent pas représenter la totalité des possibilités linguistiques de la langue cible et ils tendent à refléter les idiosyncrasies de la langue source ainsi que celles du traducteur. Baker (1996), quant à elle, nous explique que les textes produits par une traduction, comme n'importe quel autre texte, sont influencés par leur contexte de production et les buts communicatifs qu'ils servent. Par conséquent, ils possèdent des caractéristiques propres qui les distinguent des textes produits "spontanément".

Nous employons le terme de *translecte* [22] pour évoquer cette variante de langue produite en situation de traduction. L'existence du translecte a été largement étudiée et démontrée. Ses caractéristiques sont dégagées en comparant un corpus de traduction avec un corpus de textes spontanés portant sur une même thématique.

Baker (1996) synthétise les résultats de plusieurs études principalement basées sur la comparaison d'originaux et de traductions en anglais (textes journalistiques, romans).

---

19. ≪ *a comparable corpus can be defined as a corpus containing components that are collected using the same sampling frame and similar balance and representativeness (McEnery, 2003 :450) , e.g. the same proportions of the texts of the same genres in the same domains in a range of different languages in the same sampling period. However the subcorpora of a comparable corpus are not translations of each other. Rather, their comparability lies in their same sampling frame and similar balance.* ≫ (Mc Enery et Xiao, 2007, p. 20)

20. ≪ *Deux corpus de deux langues l1 et l2 sont dits comparables s'il existe une sous-partie non négligeable du vocabulaire du corpus de langue l1, respectivement l2, dont la traduction se trouve dans le corpus de langue l2, respectivement l1* ≫ (Déjean et Gaussier, 2002, p. 2)

21. ≪ *As such, translated language is at best an unrepresentative special variant of the target language* ≫ (Mc Enery et Xiao, 2007, p. 24)

22. Notre traduction du terme anglais *translationese*, par analogie avec géolecte *'variété de langue parlée dans une zone géographique donnée'*, sociolecte *'variété de langue parlée dans un milieu social donné'*, etc.

Elle met au jour quatre particularités :

Explicitation L'explicitation est la tendance à éviter l'implicite, voire à ajouter des informations supplémentaires pour recontextualiser le message. Les textes traduits sont toujours plus longs que le texte source, quel que soit le sens de traduction ; d'un point de vue lexical, on note plus de vocabulaire explicatif (*cause, reason*) et de conjonctions telles *because, consequently*.

Simplification Le langage utilisé est simplifié. Les phrases trop longues sont redécoupées en phrases plus courtes. La ponctuation est altérée : les formes faibles sont remplacées par des formes plus fortes (virgule → point-virgule → point). Les traductions sont moins variées lexicalement et comprennent une plus forte proportion de mots outils.

Normalisation / conservatisme Cet aspect concerne la conformité voire l'exagération des caractéristiques typiques de la langue cible, en particulier au niveau des structures grammaticales, de la ponctutation et des collocations.

Homogénéisation (« *levelling out* ») Sur de nombreux d'aspects, les textes traduits montrent beaucoup moins de variation que les textes spontanés. Par exemple, si on observe les variations du ratio forme:occurrence (mesure de variété lexicale) ou de la longueur des phrases sur plusieurs textes, la variance de ces caractéristiques est beaucoup plus faible pour les textes traduits.

Concernant les corpus comparables, plusieurs études soulignent leur intérêt pour la traduction.

Deux études (Friedbichler et Friedbichler, 1997; Gavioli et Zanettin, 1997), citées par Mc Enery et Xiao (2007), ont estimé que les corpus comparables spécialisés se révèlent utiles en traduction technique lorsqu'il s'agit de vérifier des hypothèses de traduction. Friedbichler et Friedbichler (1997) notent des améliorations qualitatives, que la traduction soit vers la langue première ou vers la langue seconde du traducteur. Le fait qu'il y ait une amélioration même dans le cas de la traduction vers la langue première est révélateur de la difficulté d'appréhender des textes spécialisés. En effet, le fait de maîtriser la langue courante n'implique pas que l'on connaisse la terminologie ou les usages linguistiques propres à un domaine, encore moins les notions qui y sont manipulées.

Les travaux de Zanettin (1998) sur la formation des traducteurs mettent en lumière trois usages possibles des corpus comparables :

Recherche d'équivalences traductionnelles Zanettin décrit une expérience sur l'identification d'équivalences traductionnelles dans des journaux sportifs, réputés pour contenir une grande quantité de language figuratif. L'exemple donné est celui de la traduction de l'expression *salire il gradino più alto del podio 'monter sur la plus haute marche du podium'* vers l'anglais : peut-elle être traduite littéralement ou faut-il trouver un équivalent ? L'étude en corpus des contextes d'apparition de l'expression italienne montre que cette expression a le sens de *'gagner la médaille d'or'*. L'étude des coocurrences du mot *podium* dans les textes anglais montre que, bien que le sens dénotatif soit le même que *podio*, *podium* n'apparaît pas en conjonction avec *the highest step* pour signifier *'gagner la médaille d'or'*. Une traduction littérale serait donc maladroite et la traduction retenue sera *to win the gold medal*.

Apprentissage de la terminologie Zanettin souligne la forte proportion d'équivalences traductionnelles entre termes aux graphies similaires dans les corpus médicaux (termes ayant des origines gréco-latines communes, ex : *hépatique* ↔ *hepatic*). Il explique que l'observation des collocations de ces termes similaires peut servir à acquérir des connaissances terminologiques sur le domaine. L'exemple donné est celui de la traduction

de *biopsia epatica* dont une traduction intuitive en anglais serait *hepatic biopsy*. Pourtant, les contextes de *biopsy* ne font jamais état de l'expression *hepatic biospy* alors que *liver biopsy* apparaît 39 fois. Une étude plus avant des contextes de *liver vs. fegato* (formes populaires) et *hepatic vs. epatico/a* (formes savantes) montrent qu'anglais et italien ne recourrent pas aux formes populaires et savantes de la même façon : en anglais, *hepatic* co-occurre seulement avec des mots génériques comme *lesion* ou exedisease alors qu'en italien, la forme savante est employée sans restriction particulière.

Exploration des textes en post- ou pré- traduction Il s'agit ici d'utiliser les corpus comparables pour examiner les usages propres à un domaine ou un genre. L'expérience décrite concerne une étude comparée des contextes d'apparition du mot *Mitterand* dans les journaux anglais et italiens. Cette étude révèle des usages stylistiques propres à chaque langue : l'italien a tendance à appeler les politiciens par leur prénom et nom (*François Mitterand*) alors que l'anglais a plutôt recours à un titre (*Mr. Mitterand, President Mitterand*). Les usages sont également différents en ce qui concerne l'introduction du discours rapporté : en anglais, peu de verbes différents sont utilisés (*say* et *add* sont utilisés dans 60 % des cas) alors qu'en italien, les verbes employés pour rapporter un discours sont beaucoup plus variés.

1.1.4 Difficultés de la traduction technique

Pour expliquer les difficultés de la traduction technique, nous nous appuierons sur l'ouvrage de Christine Durieux (2010) qui s'inscrit dans le cadre de la théorie interprétative de la traduction (ou théorie du sens) de Danica Seleskovitch.

De prime abord, on pourrait croire que la traduction spécialisée est uniquement concernée par l'acquisition d'équivalences traductionnelles entre termes (apprentissage de la terminologie). Or, comme l'explique Durieux (2010), la traduction technique ne peut être réduite à un processus de production d'équivalences terminologiques, démarche qu'elle appelle « *transcodage* » et qui consisterait simplement à transposer les termes en langue cible sans vraiment les comprendre. L'auteure estime qu'une bonne traduction technique ne peut être produite que si le traducteur est réellement à l'aise avec les notions désignées par les termes : « *on ne traduit pas une suite de mots, mais un message dont on a auparavant appréhendé le sens* » (*op. cit.*, p. 42). Le travail du traducteur implique donc une part d'auto-formation au domaine technique par une recherche documentaire préalable permettant de découvrir la terminologie du domaine en contexte. Durieux conseille d'effectuer la recherche documentaire dans des œuvres de vulgarisation comme les encyclopédies où les notions sont décrites avec un vocabulaire facile d'accès plutôt que dans des ressources spécialisées (revues scientifiques). Les ressources spécialisées ne sont employées que dans un second temps, pour affiner certaines notions.

Il en est de même pour les usages stylistiques : Durieux note qu'il existe des « *tournures particulières* » propres à chaque domaine. Certaines constructions syntaxiques, certaines collocations peuvent être plus fréquentes dans le discours spécialisé qu'en langue générale. Souvent, les collocations propres à un domaine de spécialité impliquent une traduction différente d'un des collocats. Par exemple, *to spread* se traduit par *répandre* en langue générale mais *to spread insecticide* se traduit par *traiter à l'insecticide* ; *unscheduled* se traduit par *imprévu* en langue générale mais *unscheduled maintenance* se traduit par *entretien curatif*. La traduction des prépositions est délicate car la préposition peut changer le sens d'un terme : *exception détectée par le programme* a un sens différent de *exception détectée dans le programme*. De plus, le choix de la préposition peut être idiosyncratique au terme : on parle de *perçage par laser*,

de $soudure\ \underline{par}\ laser$ mais de $découpe\ \underline{au}\ laser$. Ici aussi, Durieux préconise une recherche documentaire systématique permettant de relever les usages linguistiques propres au domaine.

À la lecture de Durieux, on comprend qu'un traducteur spécialisé passe une partie de son temps à effectuer de la recherche documentaire dans le but de constituer manuellement des fiches terminologiques qui mettent en correspondance non seulement des termes mais aussi des contextes (contextes définitoires pour le sens de termes, contextes "langagiers" mettant en lumière les collocations et aspects stylistiques).

D'autres études viennent appuyer les constations de Durieux. Ainsi, Darbelnet (1979) considère qu'une langue de spécialité est spécifique par sa « *nomenculature* » mais aussi par (ce qu'il nomme) son « *discours* » :

> « *Dans l'emploi de ce qu'il est convenu d'appeler les langues de spécialité, il y a d'une part les choses techniques qu'il faut pouvoir désigner exactement et d'autre part le texte qui véhicule et actualise ces notions et qui doit répondre à certaines exigences de forme. Il s'ensuit que l'auteur du texte doit posséder une double compétence : bien connaître la nomenclature du sujet et être capable de tirer pleinement parti, dans un certain registre, des ressources langagières propres à mettre en valeur les éléments de la nomenclature.* [...] *Dans cette perspective, on peut considérer que chaque langue de spécialité se présente sous ce double aspect, que nomenclature et discours ne peuvent aller l'un sans l'autre et qu'il est souvent plus facile, grâce à la documentation appropriée, d'accéder à la nomenclature qu'aux ressources du discours spécialisé.* » (op. cit.)

Reprenant la distinction de Darbelnet, Scurtu (2008) nous livre une analyse fine des difficultés de la traduction des textes juridiques français vers le roumain. Un point notable est qu'elle considère que les éléments de la nomenculature (i.e. les termes techniques) ne posent pas nécessairement de difficulté de traduction. Scurtu décompose la nomenculature en trois catégories :

Les mots d'appartenance juridique exclusive Il s'agit des termes techniques, employés par les initiés. Certains peuvent ne poser aucune difficulté de traduction parce qu'ils ont un correspondant direct en langue cible (voire ils sont un emprunt à la langue source) et qu'ils possèdent une ressemblance formelle avec le terme source, ex : $abrogatif \rightarrow abrogativ$. Les termes pouvant poser une difficulté de traduction correspondent à des termes n'ayant pas de ressemblance formelle avec le terme source (ex. :$prononcé \rightarrow pronunţare$) et/ou désignant une notion qui n'existe pas dans la culture associée à la langue cible (ex. : $communauté \rightarrow regim\ matrimonial\ legal$).

Les mots à double appartenance Il s'agit de termes que le droit emploie dans une acception qui lui est propre. Parmi ceux-ci, on retrouve :

- Les termes d'appartenance juridique principale : il s'agit de termes juridiques passés dans la langue courante avec un sens secondaire, ex. : $arbitre,\ témoin,\ garantie$.
- Les termes d'appartenance juridique secondaire : il s'agit de termes dont le sens principal est en langue courante et qui ont acquis un sens particulier dans le domaine juridique, ex. : $acte,\ mobile,\ jouissance$.

La difficulté de traduction des mots à double appartenance vient du fait que ceux-ci sont partagés avec la langue courante : leur traduction n'est possible qu'en contexte.

Le discours, quant à lui, regroupe divers éléments. On y retrouve les éléments stylistiques, les formulations spécifiques et les choix syntaxiques déjà mis en exergue par Durieux mais également ce que Darbelnet et Scurtu nomment le « *vocabulaire de soutien* ». Darbelnet (1979) définit le vocabulaire de soutien comme « *les mots qui, étant d'une technicité moindre ou nulle,*

servent à actualiser les mots spécialisés et à donner ainsi au texte son organicité. ». Il nous donne l'exemple, pour le domaine juridique, des mots $rupture$ (de la vie commune), $entendre$ (un témoin), $exorbiter$, $dépérir$, $supporter$ (au sens fiscal).

De même, Scurtu (2008) indique que, les mots à double appartenance mis à part, il reste un certain nombre de termes qui, sans avoir un sens juridique, apparaissent toutefois dans les textes avec une valeur spécifique, différente de celle qu'ils ont dans la langue "commune". Par exemple, $affaire$ n'a pas le sens, dans les textes juridiques, qui est rendu par sa traduction littérale en roumain ($afacere$). En contexte juridique il sera tradui par $cauză$ ($porter\ une\ affaire\ devant\ la\ Cour \rightarrow a\ duce\ o\ cauză\ înaintea\ Curţii$ vs. $faire\ des\ affaires \rightarrow a\ face\ afaceri$).

Notons Scurtu et Darbelnet déplorent tous deux que les ressources à disposition du traducteur ne prennent pas en compte le vocabulaire de soutien :

> « *on peut considérer que chaque langue de spécialité se présente sous ce double aspect, que nomenclature et discours ne peuvent aller l'un sans l'autre et qu'il est souvent plus facile, grâce à la documentation appropriée, d'accéder à la nomenclature qu'aux ressources du discours spécialisé.* » (Darbelnet, 1979)

> « *Les ouvrages en question n'incluent souvent que les termes du domaine proprement dit et excluent les termes de la langue courante qui, ayant acquis un sens particulier, échappent à la compréhension du néophyte.* » Scurtu (2008, p. 88)

D'après Darbelnet, l'absence du vocabulaire de soutien dans les glossaires techniques s'explique par le fait que ces ressources sont plus orientées vers l'aide à la compréhension que vers l'aide à la rédaction. De plus, comme les termes techniques frappent par leur technicité, ils s'imposent naturellement comme comme nécessaires à répertorier dans un glossaire technique. *A contrario*, le vocabulaire de soutien, qui semble plus transparent, sera plus facilement négligé. Pourtant, il n'en est pas moins indispensable. Cette vue est également supportée par Scurtu (2008) :

> « *Paradoxalement, pour rédiger ou traduire un texte, souvent ce n'est pas le mot technique qui constitue le problème le plus important (ces mots techniques ont fait et continuent de faire l'objet de lexicographies terminologiques). On constate, en feuilletant des répertoires de la langue juridique, que nombre de termes utilisés dans la rédaction de textes juridiques et administratifs n'ont pas été retenus. Cela est d'autant plus valable si on prend en considération la situation des dictionnaires bilingues dans le domaine. Il est vrai qu'en général les répertoires visent plutôt à la compréhension qu'à la rédaction. Au contraire, les termes du vocabulaire de soutien, bien qu'apparaissant comme marginaux, parce que transparents, s'avèrent d'un maniement plus délicat, car ils sont nécessaires pour passer de simples listes de termes au texte : c'est au moment où il faut rédiger, précise encore Darbelnet* [1979]*, et, en l'occurrence, complétons-nous, traduire, que ce vocabulaire prend effectivement toute sa valeur.* » (*op. cit.*, p. 892)

Comme nous nous plaçons dans une optique d'aide à la traduction et non d'ingénierie des connaissances, notre travail ne sera pas focalisé sur l'extraction d'équivalences traductionnelles entre termes. Nous nous attacherons plutôt à identifier les traductions de tout élément lexical susceptible de poser des difficultés de traduction. Nous écartons donc de notre sujet de recherche toute information relative à la syntaxe, à la stylistique ou à la structuration du texte. Nous considérons comme « *susceptible de poser des difficultés de traduction* » toute unité lexicale qui n'est pas présente dans le dictionnaire généraliste. De part cette définition, nous excluons certains termes techniques couramment employés dans la langue courante et dont la

traduction est nécessairement connue des traducteurs (ex. $chimiothérapie$ est un terme médical mais sa traduction ne posera pas de problème à un traducteur professionnel). Par contre, nous incluons des éléments tels que $patient\text{-}centred$ qui n'auraient pas leur place dans une terminologie mais qui peuvent poser des difficultés de traduction.

Ainsi, dans la suite du mémoire, notre emploi du vocable « *terme* » n'est pas à prendre dans son acception officielle[23] mais plutôt au sens d' « *unité problématique pour le traducteur technique* ».

1.1.5 Contexte industriel

Si l'intérêt qualitatif des corpus comparables est avéré, ces derniers restent difficilement exploitables par les traducteurs. Par rapport aux corpus parallèles pour lesquels de nombreux outils existent, la recherche et la vérification manuelle de contextes informatifs et d'équivalences traductionnelles dans les corpus comparables est laborieuse. Ceci génère une perte de productivité et de motivation pour le traducteur.

Il existe très peu d'outils informatiques capables d'assister le traducteur dans son utilisation des corpus comparables. Nous ne pouvons citer que deux prototypes universitaires (Bennison et Bowker, 2000; Sharoff *et al.*, 2006) et - à notre connaissance - il n'existait, au moment où nous avons débuté notre thèse, aucun outil de TAO commercial capable de traiter les corpus comparables. Le transfert technologique des techniques d'extraction de lexiques bilingues à partir de corpus compararables a été notre première tâche lorsque nous avons commencé à travailler en tant qu'ingénieure de recherche pour la société LINGUA ET MACHINA[24]. Cette société, fondée par Emmanuel Planas sur la base de ses résultats de recherche (Planas, 1998; Planas et Furuse, 2000), édite le logiciel de mémoire de traduction SIMILIS (Planas, 2005) dont la particularité est de recourir à une analyse linguistique. Les textes sont étiquetés morpho-syntaxiquement et les phrases sont découpées en *chunks*. L'appariement entre segments de textes déjà traduits et segments de textes à traduire se fait également à un niveau linguistique (appariement sur les lemmes et catégories grammaticales) et non pas à un niveau graphique comme le font les autres logiciels mémoires de traduction.

LINGUA ET MACHINA édite également une application Web de gestion des contenus multilingues en entreprise appelée LIBELLEX. Cette plateforme intègre divers outils d'aide à la traduction (concordanciers bilingues, outils d'extraction et gestion de terminologies, mémoire de traduction, traduction automatique et outil de gestion de projets de traduction). La plateforme se distingue de SIMILIS par le fait d'être pensée non pas uniquement pour les traducteurs professionnels mais pour l'ensemble des collaborateurs de l'entreprise (figure 1.1).

La possibilité d'exploiter des corpus comparables représente un axe de Recherche et Développement majeur à LINGUA ET MACHINA dans la mesure où, les domaines de connaissance évoluant très vite, les entreprises clientes de LINGUA ET MACHINA doivent pouvoir rapidement créer des ressources de traduction, même dans des domaines pour lesquels il existe peu ou pas d'historique de traduction.

Une partie de notre travail de thèse a donc consisté à créer un prototype permettant l'acquisition de lexiques bilingues à partir de corpus comparables. Nous avons également développé une interface de consultation des lexiques extraits associant aux termes sources et cibles des fiches terminologiques constituées automatiquement. Généralement, l'acquisition

23. « *A term is a designation consisting of one or more words representing a general concept in a special language.* » (ISO, 2009, p.22)

24. www.lingua-et-machina-com

d'un lexique bilingue à partir de corpus comparables se fait en deux temps. Tout d'abord, les termes sources et cibles sont extraits de leurs corpus respectifs en utilisant les techniques d'extraction terminologique (Bourigault, 1994; Daille, 1994; Enguerard et Pantera, 1995). Puis, les termes extraits sont alignés à l'aide de techniques basés sur la similarité des contextes d'occurrence des termes. Nous décrivons ces techniques dans la section suivante.

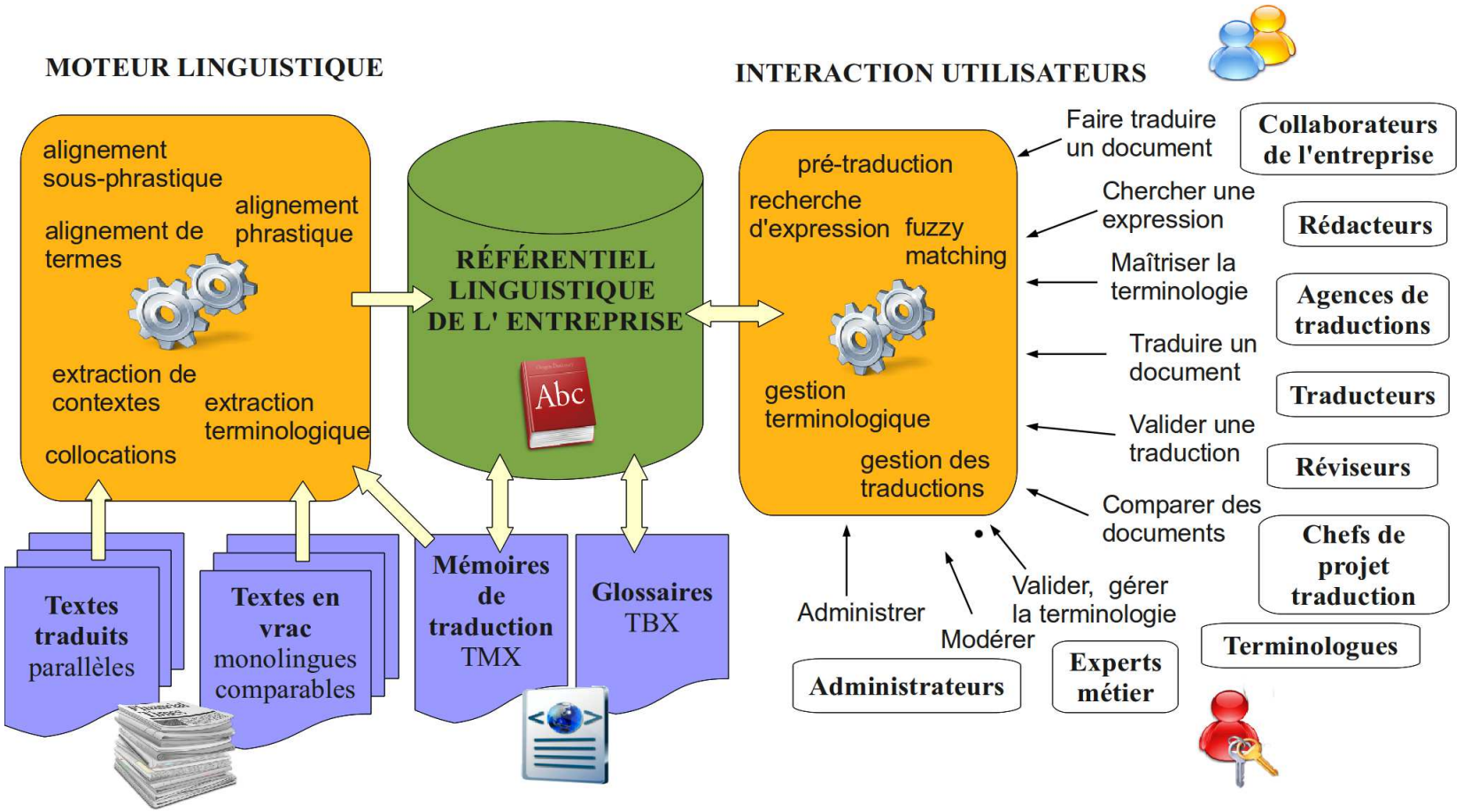


FIGURE 1.1 – Libellex : une plateforme multiservices pour la gestion des contenus multilingues

1.2 Techniques d'alignement de termes en corpus comparables

Des approches spécifiques ont été développées pour acquérir des lexiques bilingues à partir de corpus comparables. Il existe des méthodes basées sur les distributions des fréquences (Koehn et Knight, 2002) ou l'exploitation de relations sémantiques (Ji, 2009) mais nous ne les détaillons pas ici car soit leurs résultats sont peu probants, soit ils dépendent d'outils avancés d'extraction d'information, ce qui nous a conduit à les considérer comme difficiles à mettre en œuvre.

D'autres méthodes cherchent à extraire des segments parallèles de corpus comparables (Fung et Cheung, 2004; Rauf et Schwenk, 2009). Si cette approche est efficace pour créer des lexiques de langue générale, elle est difficilement applicable avec des corpus spécialisés car elle nécessite de grands corpus [25]. Or, en domaine spécialisé, les textes doivent non seulement appartenir au même domaine mais aussi être restreints à une thématique bien précise ce qui rend quasi-impossible la collecte d'un grand nombre de textes.

La méthode état-de-l'art la plus répandue pour aligner des termes en corpus comparables est appelée méthode *distributionnelle* ou méthode *d'alignement par similarité contextuelle*. Nous décrivons son principe dans la section 1.2.1.

1.2.1 Principe de l'approche distributionnelle

La sémantique distributionnelle, qui trouve son origine dans les travaux de Z. Harris, considère qu'il n'existe pas d'organisation ou de structuration du sens en soi et qu'il n'est pas possible d'assigner un sens aux mots *a priori*. Toutefois, elle estime qu'il est possible de caractériser un mot sémantiquement grâce à sa *distribution*, c'est-à-dire l'ensemble des mots avec lesquels il entretient des relations syntaxiques.

L'extraction de lexiques bilingues basée sur la sémantique distributionnelle fait l'hypothèse qu'il est possible de calculer des similarités distributionnelles entre mots de langues différentes et que des distributions similaires correspondent à des équivalences sémantiques, quelles que soient les langues en jeu. Cette hypothèse a été testée avec succès par Rapp (1995) et le premier modèle d'alignement a été proposé par Fung (1997).

Rapp (1995) démontre la pertinence de la sémantique distributionnelle pour l'alignement de termes en montrant qu'il existe une corrélation entre les motifs de co-occurrence de mots observés sur des corpus de langues différentes [26] : si les mots A et B co-occurrent de façon significative dans un corpus de langue $L1$, alors leurs traductions respectives A' et B' en langue $L2$ co-occurrent également de façon significative dans un corpus de langue $L2$. Par exemple, dans un corpus médical français-anglais, on peut s'attendre à ce que $dépistage$ et $radiographie$ co-occurrent de façon significative en français, tout comme leurs traductions respectives en anglais $screening$ et $radiography$.

Dans son expérience, Rapp représente les co-occurrences entre mots par une matrice où

25. Par exemple, Fung et Cheung (2004) utilisent un corpus chinois de 110 000 phrases et un corpus anglais de 290 000 phrases pour obtenir 2 500 paires de phrases alignées, avec une précision de 65,76 %.

26. Il faut noter que l'auteur, tout comme la plupart des travaux qui suivront, n'a pas directement recours à des analyseurs syntaxiques. D'une part, les contextes syntaxiques ne sont disponibles qui si le corpus utilisé pour l'alignement a été analysé syntaxiquement, ce qui implique le recours à des outils coûteux à développer et rarement disponibles. D'autre part, dès que les corpus atteignent des tailles suffisantes, les contextes syntaxiques peuvent être approximés grâce à une fenêtre de n mots entourant le mot à caractériser sémantiquement.

la valeur à l'intersection (i, j) indique la significativité [27] de la co-occurrence du mot i avec le mot j. L'expérience qu'il mène démarre avec deux matrices de ce type, l'une contenant les co-occurrences observées sur le corpus source (anglais), et l'autre contenant les co-occurrences observées sur le corpus cible (allemand). Au départ, les deux matrices sont alignées, c'est-à-dire que le mot i de la matrice en anglais est la traduction du mot i de la matrice en allemand. Puis, Rapp permute aléatoirement l'ordre des mots dans les matrices de façon à les désaligner. Il observe alors que la similarité [28] des matrices source et cible décroît lorsque le nombre de mots désalignés augmente.

Fung (1997) pousse plus loin l'expérimentation de Rapp et utilise un lexique bilingue qu'elle projette sur les corpus source et cible, ce qui permet d'obtenir des paires de traductions attestées dans les deux corpus. Elle calcule ensuite, pour chaque mot source et cible dont la traduction est inconnue, un $vecteur\ de\ contexte$. Le vecteur de contexte d'un mot m est une approximation de sa distribution : il donne, pour chacune des entrées e du lexique bilingue, le nombre de fois où m co-occurre avec l'entrée e au sein d'une fenêtre contextuelle donnée (par ex., trois mots à droite et trois mots à gauche). Comme les entrées sont attestées dans les corpus source et cible, il est possible de comparer les vecteurs de contextes indépendamment de leur langue. Plus deux vecteurs sont similaires (c'est-à-dire plus ils co-occurrent de façon significative avec des mots équivalents dans les deux langues), plus il est plausible que leurs têtes [29] aient un sens proche et soient des traductions.

Cette méthode d'alignement peut être résumée ainsi :

1. Construire les vecteurs de contexte des termes sources et cibles (figure 1.2) :
   - Le vecteur d'un terme t correspond à $\vec{t} = \{(m_1, cooc_1), ...(m_n, cooc_n)\}$ où chaque m_i est un mot co-occurrant avec t au sein d'une fenêtre contextuelle donnée (par exemple, 5 mots à droite et 5 mots à gauche de t) et $cooc_i$ est le nombre de fois où cette cooccurrence se produit.
2. Normaliser le nombre de co-occurrences à l'aide d'une mesure comme l'information mutuelle ou le taux de vraisemblance (cf. annexe A.1).
3. Traduire les vecteurs des termes sources en langue cible à l'aide d'un dictionnaire bilingue (figure 1.3).
4. Pour chaque terme source (figure 1.4) :
   - Comparer le vecteur de contexte traduit aux vecteurs de contexte des mots cibles à l'aide d'une mesure de similarité (cf. annexe A.2).
   - Ordonner les vecteurs des termes cibles par similarité décroissante.
   - Sélectionner les N vecteurs les plus similaires : les termes cibles associés à ces N vecteurs sont les traductions candidates du terme source.

Cette technique d'alignement correspond à la méthode état-de-l'art qui a ensuite été déclinée de diverses façons comme nous le verrons en section 1.2.3. Mais avant d'exposer ces variantes de la méthode distributionnelle, nous allons tout d'abord nous pencher sur les méthodologies d'évaluation de ces techniques d'alignement.

27. La mesure employée est l'information mutuelle décrite en annexe p. 192.
28. La similarité entre deux matrices correspond à la somme des différences entre les valeurs se trouvant à des positions identiques dans la matrice.
29. Nous reprenons la terminologie de Prochasson (2010, p. 15) et appelons par la suite le mot dont on calcule la distribution la ≪ *tête* ≫ du vecteur et chaque entrée du lexique bilingue présente dans le vecteur un ≪ *élément* ≫ du vecteur de contexte.

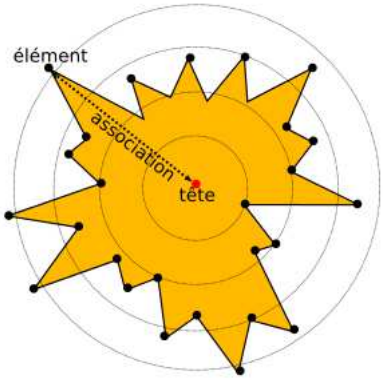


FIGURE 1.2 – Représentation d'un vecteur de contexte - emprunté à Prochasson (2010)

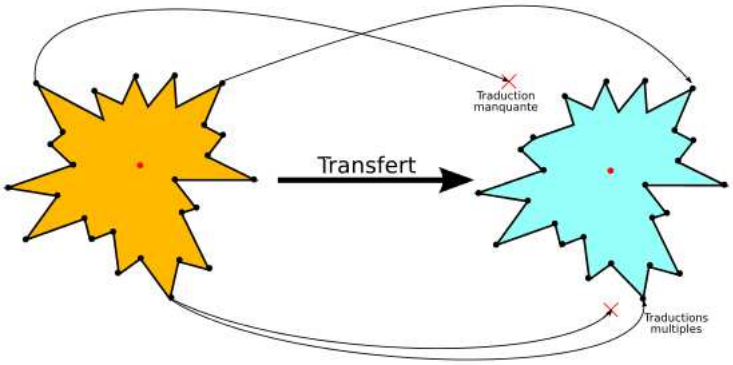


FIGURE 1.3 – Traduction d'un vecteur de contexte - emprunté à Prochasson (2010)

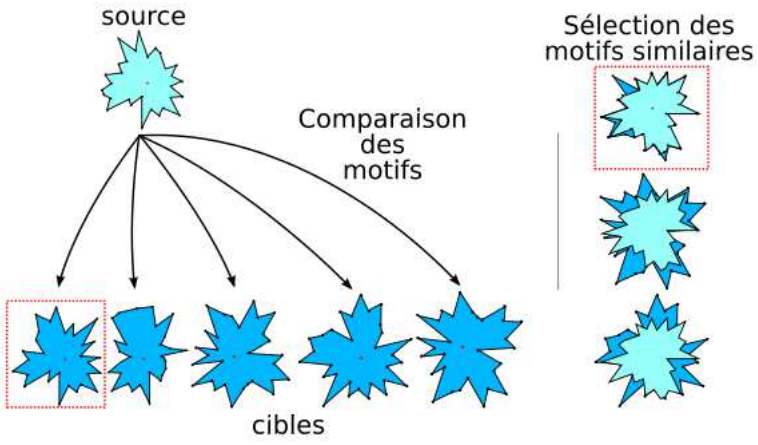


FIGURE 1.4 – Comparaison des vecteurs et sélection des vecteurs les plus similaires - emprunté à Prochasson (2010)

1.2.2 Évaluation des techniques d'alignement en corpus comparables

L'évaluation des techniques d'alignement se fait généralement par comparaison des sorties du système avec un lexique bilingue de référence. Les sorties d'un système d'extraction de lexiques à partir de corpus comparables correspondent à une liste de paires $(s, \{t_1, ...t_n\})$ où s est un terme source et $\{t_1, ...t_n\}$ l'ensemble ordonné de ses traductions candidates. Contrairement aux systèmes d'extraction de lexiques à partir de corpus parallèles, il est très difficile d'obtenir un lexique de qualité en se contentant de sélectionner la première traduction candidate. Les mesures habituellement employées pour les corpus parallèles comme par exemple l'*Alignement Error Rate* (Och et Ney, 2000) ne sont pas les plus pertinentes car on cherche plutôt à évaluer la capacité de l'algorithme à placer la traduction correcte le plus haut possible dans la liste des traductions candidates.

La littérature fait état de trois mesures d'évaluation : la précision au rang N (aussi appelée *TopN*), le MRR (*Mean Reciprocal Rank*) et la MAP (*Mean Average Precision*).

Précision au rang N ou TopN

C'est de loin la mesure la plus utilisée. Elle est dérivée de la mesure de précision utilisée en recherche d'information (nb. de documents pertinents / nb. de documents retournés). Elle représente la proportion de termes sources qui ont au moins une traduction correcte parmi leur N premières traductions candidates :

$$P_N = \frac{1}{|S|} \sum_{i=1}^{|S|} \alpha(T_{i^N}, R_i) \tag{1.1}$$

$$\alpha(T_{i^N}, R_i) = \begin{cases} 1 & \text{si } T_{i^N} \cap R_i \neq \emptyset \\ 0 & \text{sinon} \end{cases}$$

où :

- S est l'ensemble des termes sources ayant au moins une traduction candidate
- T_{i^N} est l'ensemble des N premières traductions candidates pour le terme source i
- R_i est l'ensemble des traductions de référence pour le terme source i

Il est aussi possible de calculer le rappel sur les N premières traductions candidates (R_N) qui correspond à l'équation 1.1 sauf que l'ensemble S est l'ensemble de *tous* les termes sources, pas uniquement les termes sources ayant reçu au moins une traduction candidate. L'augmentation de la précision faisant baisser le rappel, la $F1_N$, qui correspond à la moyenne harmonique de P_N et R_N, synthétise le compromis entre rappel et précision (Laroche et Langlais, 2010). Néanmoins, R_N et $F1_N$ sont peu employées. Les systèmes d'extraction de lexiques bilingues sont pour la plupart des systèmes d'*alignement* de termes : ils prennent en entrée un ensemble de termes sources et un ensemble de termes cibles puis calculent un score de traduction pour toutes les paires `(terme source, terme cible)`. Dans les faits, au moins une traduction est proposée pour chaque terme source.

MRR

La mesure MRR correspond à la moyenne de l'inverse des rangs des traductions correctes :

$$MRR = \frac{1}{|S|} \sum_{i=1}^{|S|} \frac{1}{rang_i} \tag{1.2}$$

où S est l'ensemble des termes sources avec au moins une traduction candidate et $rang_i$ le rang de la première traduction candidate correcte du terme source i (Yu et Tsujii, 2009).

MAP

Laroche et Langlais (2010) proposent d'utiliser la MAP, également une mesure de recherche d'information. Elle correspond à la moyenne de la précision obtenue sur l'ensemble des k documents (ou traductions candidates) retournés par le système avant chaque document pertinent (ou traduction de référence) :

$$MAP = \frac{1}{S} \sum_{i=1}^{|S|} \frac{1}{R_i} \sum_{j=1}^{R_i} Precision(T_{ij}) \tag{1.3}$$

où :
- S est l'ensemble des termes sources
- R_i est le nombre de traductions de référence pour le terme S_j
- T_{ij} est l'ensemble des traductions candidates données par le système pour le terme S_i avant la traduction de référence j.

1.2.3 Améliorations et variantes de l'approche distributionnelle

Plusieurs variantes et améliorations de l'approche distributionnelle ont été proposées. Ces dernières se concentrent sur la recherche d'une symétrie distributionnelle (Chiao, 2004; Sadat *et al.*, 2003), l'utilisation de contextes lexico-syntaxiques (par opposition à la fenêtre contextuelle) (Otero et Campos, 2005) et le recours à des points d'ancrage, comme c'est le cas pour l'alignement phrastique en corpus parallèle (Prochasson et Morin, 2009). D'autres variantes ont exploité des affinités sémantiques de second ordre combinées avec des classes sémantiques (Déjean et Gaussier, 2002). Enfin, les travaux de Morin *et al.* (2004) ont tenté d'aligner des unités polylexicales.

1.2.3.1 Favoriser la symétrie distributionnelle

Chiao (2004) s'appuie sur l'hypothèse de symétrie distributionnelle qui stipule que « *si deux mots sont proches dans une direction de traduction ainsi que dans l'autre (langue A $\leftrightarrow$ langue B) alors ils ont de plus fortes chances d'être traductions l'un de l'autre que s'ils ne sont proches que pour une seule* » (*op. cit.*, p. 13).

Chiao utilise donc ce qu'elle appelle une « *similarité croisée* » par opposition à la « *similarité classique* » (*op. cit.*, p. 89). Après avoir effectué deux processus d'alignement, l'un dans le sens source $\rightarrow$ cible, l'autre dans le sens cible $\rightarrow$ source, Chiao calcule, pour chaque paire de termes source et cible (M_S, M_C), la moyenne harmonique de rM_C, le rang de M_C parmi les traductions candidates de M_S et rM_S, le rang de M_S parmi les traductions candidates de M_C :

$$MH(rM_C, rM_S) = \frac{2 \times rM_C \times rM_S}{rM_C + rM_S} \tag{1.4}$$

Ses expériences montrent que la similarité croisée augmente le nombre de traductions trouvées, quel que soit le corpus (dans le meilleur des cas, la précision sur le Top1 passe de 28 % à 34 %).

Sadat *et al.* (2003) calculent également une similarité croisée $SIM_{S \leftrightarrow C}$ entre un mot source M_S et un mot cible M_C sur la base de :

$$SIM_{S \leftrightarrow C}(M_S, M_C) = SIM_{S \rightarrow C}(M_S, M_C) \times SIM_{C \rightarrow S}(M_C, M_S) \quad (1.5)$$

où $SIM_{S \rightarrow C}(M_S, M_C)$ est la similarité entre le vecteur de M_S traduit en langue cible et le vecteur de M_C et $SIM_{C \rightarrow S}(M_C, M_S)$ est la similarité entre le vecteur de M_C traduit en langue source et le vecteur de M_S.

Les auteurs appliquent également un filtrage morphologique aux traductions : un nom ne peut être traduit que par un nom, un verbe ne peut être traduit que par un verbe, etc. Les alignements ainsi obtenus sont évalués à travers leur apport dans un système de recherche d'information interlingue : l'utilisation du lexique acquis par similarité croisée augmente significativement la précision moyenne[30] de la recherche d'information de 27,1 % (de 0,1417 à 0,1801) par rapport au cas où la recherche d'information utilise uniquement le lexique acquis sans la similarité croisée.

1.2.3.2 Utiliser des contextes syntaxiques

Otero et Campos (2005) emploient des contextes syntaxiques exprimés sous la forme de patrons lexico-syntaxiques acquis sur un corpus parallèle. Ces patrons sont de la forme $<$ `lemma [POS]` $>$. Par exemple, le patron $<$ `import of [NOM]` $>$ s'apparie avec tout nom apparaissant à droite de $import\ of$.

Les patrons syntaxiques bilingues sont acquis en trois temps à partir d'un corpus parallèle anglais-espagnol :

1. Acquisition des patrons syntaxiques anglais sur la partie source du corpus, par exemple : <`import of [NOM]`>, <`aid [VERBE]`>, <`[NOM] against fraud`>
2. Acquisition des patrons syntaxiques espagnols sur la partie cible du corpus, par exemple : <`importación de [NOM]`>, <`ayuda [VERBE]`>, <`[NOM] contra fraude`>
3. Alignement des patrons anglais et espagnols :

   <`import of [NOM]`> → <`importación de [NOM]`>

   <`aid [VERBE]`> → <`ayuda [VERBE]`>

   <`[NOM] against fraud`> → <`[NOM] contra fraude`>

   Les patrons sont alignés en utilisant le coefficient de Dice :

$$Dice(patron\ source, patron\ cible) = \frac{2|S \cap C|}{|S| + |C|} \quad (1.6)$$

où S correspond au nombre de phrases où le patron source apparaît, C correspond au nombre de phrases où le patron cible apparaît et $|S \cap C|$ correspond au nombre de fois où le patron source et le patron cible apparaissent dans les mêmes phrases alignées. Seuls les couples de patrons ayant le meilleur coefficient sont gardés (le seuil est déterminé empiriquement).

Ces patrons bilingues sont utilisés à la place du lexique bilingue : le vecteur de contexte d'un mot m contient, pour chaque patron syntaxique bilingue p, une mesure d'association entre m et p. Par exemple, si m est un nom, son vecteur de contexte indiquera son degré d'association avec les patrons <`import of [NOM]`> et <`[NOM] against fraud`>. Le degré d'association entre la

30. Moyenne des précisions obtenues pour un taux de rappel variant de 0 à 1.

tête du vecteur m et un patron syntaxique p est calculé à partir du nombre de fois où m instancie p [31], du nombre de patrons instanciés par m et nombre de mots instanciant p.

Otero et Campos obtiennent un précision de 89 % sur le Top1 et 96 % sur le Top5. Ces très bons résultats s'expliquent par la nature de leurs données : le lexique d'évaluation est composé de mots dont la fréquence est supérieure à 100 et le corpus comparable correspond à des parties non alignées appartenant à un même corpus parallèle.

1.2.3.3 Privilégier les éléments de confiance

Prochasson et Morin (2009) utilisent des points d'ancrage c'est-à-dire des mots servant d'éléments de confiance car identifiables automatiquement, peu ambigus et appartenant à la thématique du corpus comparable. Les auteurs proposent de leur donner plus de poids qu'aux autres éléments du vecteur de contexte car leurs propriétés en font des éléments hautement discriminants. Travaillant du japonais vers le français et l'anglais, ils utilisent comme points d'ancrage les translittérations et les composés savants. Au départ, la mesure d'association entre tête et éléments du vecteur est le taux de vraisemblance. Cette mesure est recalculée de façon à favoriser les points d'ancrage. Pour cela, la somme des taux de vraisemblance d'un même vecteur est re-répartie entre les co-occurrents de façon à ce que ceux qui sont des points d'ancrage soient renforcés et ceux qui n'en sont pas soient minimisés :

$$TV(M,m) = \begin{cases} TVI(M,m) + \beta & \text{si } m \in PA \\ TVI(M,m) - décalage_M & \text{si } m \notin PA \end{cases} \tag{1.7}$$

$$décalage_M = \frac{|PA|_M}{|\neg PA|_M} \times \beta \tag{1.8}$$

où $TVI(M,m)$ est le taux de vraisemblance initial entre la tête du vecteur M et son co-occurrent m, PA est l'ensemble des points d'ancrage, PA_M les co-occurrents de M qui sont des points d'ancrage, $\neg PA_M$ les co-occurrents de M qui ne sont pas des points d'ancrage, β est un coefficient variant entre 1 et 20 (dans les expérimentations, les meilleurs résultats ont été obtenus avec $\beta = 8$).

Par rapport à la méthode état-de-l'art, l'utilisation de points d'ancrage permet d'augmenter la précision de +18 % (de 17 % à 20 %) sur le Top1 pour les traductions anglais-japonais et de +10 % (de 20 % à 22 %) sur le Top1 pour les traductions français-japonais.

1.2.3.4 Améliorer la représentation de l'information sémantique

Hazem et Morin (2012) constatent que la façon dont est représentée l'information dans l'approche état-de-l'art n'est pas optimale car elle contient à la fois de l'information redondante et peut être lacunaire par ailleurs. Les auteurs proposent d'améliorer la représentation des vecteurs de contextes en leur appliquant une transformation par analyse en composantes indépendantes (ACI). Cette transformation permet de produire un nouvel espace de représentation dans lequel les informations sont aussi indépendantes que possible. L'approche se déroule en quatre temps :

1. Réduction des dimensions de la matrice, en lui appliquant une transformation par analyse en composante principale (ACP) ;

31. Par exemple dans $import\ of\ sugar$, le nom $sugar$ instancie le patron <`import of [NOM]`>

2. Transformation par ACI de la matrice en prenant en compte des informations de nature globale (contextes des entrées du dictionnaire bilingue) de façon à obtenir un espace de représentation appelé GICA puis calcul des distance entre termes sources et cibles dans ce nouvel espace ;
3. Transformation par ACI de la matrice en prenant en compte des informations de nature locale (contextes des termes cibles) de façon à obtenir un espace de représentation appelé LICA puis calcul des distances entre termes sources et cibles dans ce nouvel espace ;
4. Calcul des distances entre termes sources et cibles par combinaison linéaire des distances LICA et GICA

Les résultats obtenus montrent que l'approche GLICA donne de meilleurs résultats que l'approche état-de-l'art à partir du Top6 lorsque l'on utilise la meilleure combinaison de paramètres [32]. L'approche état-de-l'art obtient une précision de 73,77 % sur le Top 20 et l'approche GLICA obtient 75,40 % sur le Top 20. Les approches ont été testées sur deux corpus : un corpus spécialisé de petite taille et un corpus journalistique de grande taille.

1.2.3.5 Utilisation d'affinités sémantiques de second ordre

La méthode état-de-l'art établit la correspondance entre distribution d'un mot source et distribution d'un mot cible par traduction directe : on "projette" chacun des co-occurrents du mot source dans le corpus cible via le lexique bilingue, puis on tente de trouver un mot cible ayant une distribution similaire à cette "projection". Cette méthode est dépendante de la couverture du lexique bilingue : seuls les éléments présents dans le lexique et dans les deux corpus seront présents dans les vecteurs de contexte.

Pour pallier ce problème de couverture, Déjean et Gaussier (2002) proposent une méthode qui exploite les similarités distributionnelles entre termes à aligner et entrées du lexique bilingue. Ils considèrent comme traductions potentielles deux termes dont les proximités sémantiques avec les entrées du lexique bilingue sont similaires. La méthode d'alignement se décompose en 5 étapes :

1. Construire les vecteurs de contexte des termes sources et cibles à aligner ;
2. Construire les vecteurs de contexte des mots sources et cibles présents dans le lexique bilingue ;
3. Construire, pour chaque terme source, resp. cible, son *vecteur de similairité* : ce vecteur indique, pour chacune des entrées e du lexique bilingue, la similarité entre le vecteur de contexte du terme et le vecteur de contexte de l'entrée e. La taille du vecteur de similarité peut être paramétrée, i.e. on peut ne retenir que les n entrées les plus similaires ;
4. Pour chaque paire `(terme source, terme cible)`, calculer la similarité entre leurs vecteurs de similarité respectifs ;
5. Pour chaque terme source, sélectionner les N termes cibles dont le vecteur de similarité est le plus similaire au vecteur de similarité du terme source.

32. Pour l'approche état-de-l'art, les meilleurs résultats sont obtenus avec le taux de vraisemblance comme mesure de normalisation des vecteurs et le jaccard comme mesure de similarité. Pour l'approche GLICA, les meilleurs résultats sont obtenus avec l'information mutuelle comme mesure de similarité. La mesure de similarité est la distance euclidienne normalisée. Ces mesures sont données dans l'annexe A, page 191.

Cette méthode permet donc de traduire n'importe quel mot du corpus, même si aucun élément de son vecteur ne peut être traduit. Déjean et Gaussier (2002) nomment leur méthode « *traduction par similarité interlingue* » (*op. cit.*, p. 13) et l'opposent à la méthode par « *traduction directe* » (*op. cit.*, p. 13) de Fung (1997). En effet, ce qui est projeté en langue cible, c'est le degré de similarité avec les entrées du lexique et le mot à traduire, et non pas directement le contexte du mot à traduire.

Les résultats obtenus sont peu probants. Au mieux, la méthode interlingue permet d'obtenir 51 traductions correctes sur le Top20 alors que la méthode état-de-l'art en obtient 57.

1.2.3.6 Améliorer la ressource bilingue à l'aide de classes sémantiques

Déjean et Gaussier (2002) expérimentent l'exploitation des classes sémantiques d'un thésaurus en combinaison avec la méthode interlingue (1.2.3.5). Au lieu d'utiliser une ressource bilingue classique, les auteurs utilisent un thésaurus. Le thésaurus sert à inclure de nouvelles entrées dans les vecteurs de similarité utilisés dans la méthode interlingue. Pour un terme à aligner t associé au vecteur de similarité vs, l'inclusion de nouvelles entrées se fait de la façon suivante :

1. E est un ensemble de départ vide
2. Sélectionner les n entrées du thésaurus les plus proches de t, ces entrées forment l'ensemble E_0
3. Pour toutes les paires d'entrées (e_1, e_2) de E_0 :
   - ajouter à E toutes les entrées du thésaurus se trouvant sur le chemin minimal entre e_1 et e_2
   - ajouter e_1 et e_2 à E
4. Ajouter à vs toutes les entrées présentes dans E

Au final, cette technique permet d'obtenir 63 traductions correctes sur le Top20 contre 57 pour la méthode état-de-l'art.

1.2.3.7 Traduction d'unités polylexicales

Les travaux cités jusqu'ici se soucient uniquement de traduire des unités monolexicales (i.e. composées d'un seul mot). Morin *et al.* (2004) proposent une adaptation de l'approche par similarité interlingue aux unités polylexicales. Cette dernière méthode nous intéresse particulièrement puisque, dans le cadre de la traduction assistée par ordinateur, nous allons avoir besoin de traduire de telles unités.

Morin *et al.* (2004) proposent de construire le vecteur de contextes des unités polylexicales comme l'union des vecteurs de contexte de chacun des mots qui la compose. La méthode d'alignement employée exploite également des affinités sémantiques de second ordre :

1. Construire les vecteurs de contexte des unités monolexicales sources
2. Construire les vecteurs de contexte des unités polylexicales sources comme l'union des vecteurs de contexte de chacun des mots qui les composent
3. Construire les vecteurs de contexte des unités monolexicales cibles
4. Construire les vecteurs de contexte des unités polylexicales cibles comme l'union des vecteurs de contexte de chacun des mots qui les composent

5. Pour chaque unité source à traduire (figure 1.5) :
   - Sélectionner les n entrées du lexique bilingue les plus proches de l'unité source
   - Sélectionner les vecteurs de contextes cibles des n entrées bilingues
   - Calculer le barycentre de ces n vecteurs cibles : on obtient un vecteur de contexte moyen en langue cible
   - Comparer ce vecteur moyen aux vecteurs de contexte des unités cibles
   - Sélectionner les N vecteurs les plus similaires : les têtes de ces N vecteurs sont les traductions candidates de l'unité source.

La méthode donne de bons résultats pour les termes polylexicaux dont la traduction est aussi un terme polylexical (88 % sur le Top20). Les résultats sont plus mitigés pour les termes polylexicaux dont la traduction est soit un terme monolexical soit un terme polylexical (55 % sur le Top20). À titre de comparaison, la précision obtenue pour les termes monolexicaux dont la traduction est aussi un terme monolexical est de 51 % sur le Top20.

Nous avons vu dans cette sous-section différentes techniques d'alignement (méthode état-de-l'art, méthode interlingue) ainsi que plusieurs variantes de ces méthodes (exploitation de classes sémantiques, de points d'ancrage, utilisation de contextes lexico-syntaxiques...). Le tableau 1.1 p. 30 donne une synthèse des résultats de ces travaux. Ce tableau précise entre autres le type de termes à traduire (monolexicaux, notés *UML* ou polylexicaux, notés *UPL*) ainsi que la taille et la nature des corpus employés : textes généralistes (LG), textes spécialisés scientifiques (SC), de vulgarisation (VG) ou techniques (TECH).

Nous observons qu'au-delà des techniques d'alignement, les données employées ainsi que les divers paramétrages influencent également la qualité des résultats. Cet impact est analysé dans la sous-section suivante.

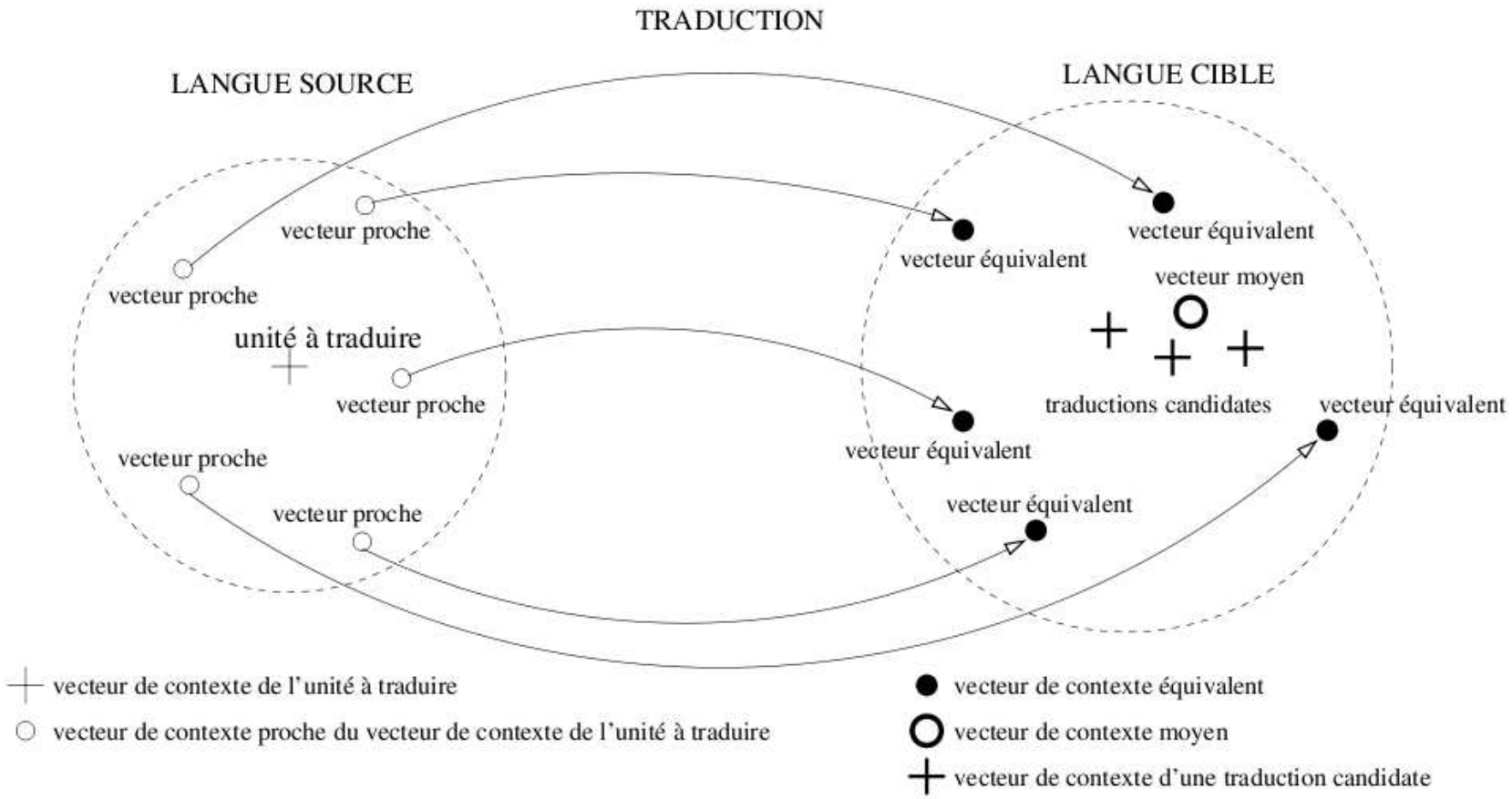


FIGURE 1.5 – Adaptation de l'approche interlingue pour l'alignement de termes polylexicaux - emprunté à Morin *et al.* (2004)

RÉFÉRENCE	LANGUES	ÉLÉMENTS À TRADUIRE			DOMAINE	TAILLE	RESSOURCES BILINGUES		MÉTHODE	PRÉCISION		
		NB.	TYPE	NB OCC.		CORPUS	TAILLE	TYPE		TOP1	TOP10	TOP20
Rapp (1999)	DE → EN	100	UML		presse	298M	16k	LG	directe	,65	,89	
Chiao et Zweigenbaum (2002)	FR → EN	95	UML	≥ 100	médical SC	1,2M	18k	médical	directe	,13	,61	,94
Déjean et Gaussier (2002)	DE → EN	1800	UML		médical SC	200k	46k	LG	directe		,44	,57
							15k	thésaurus médical	interlingue		,43	,51
		180	UML	élevé	sc. sociales SC	8M	46k	LG	directe		,35	,42
							10k	thésaurus sc. sociales	interlingue		,79	,84
Morin *et al.* (2004)	FR → EN	100	UML	≥ 5	environnement TECH	4,9 M	22 k	LG	interlingue		,41	,51
		100	UPL traduit par UML ou UPL	≥ 5							,45	,55
		100	UPL traduit par UPL	≥ 5							,87	,88
Morin *et al.* (2007)	FR → JA	100	UML	≥ 2	médical SC + VG	1,5M	173k	LG, médical	directe		,51	,6
		100	UML et UPL	≥ 2	médical SC	659k					,3	,42
Prochasson (2010)	FR → EN	122	UML	≥ 5	médical SC	507k	173k	LG, médical	directe	,21	,47	,57
		648	UML	≥ 15						.13	,34	.41
Hazem et Morin (2012)	FR → EN	122	UML	≥ 5	médical SC	530k	244k	LG	ICA	,34	,64	,76
		500	UML		presse	5M				,16	,32	.40

TABLE 1.1 – Résultats de l'état de l'art - alignement par similarité contextuelle

1.2.4 Influence des donnés et du paramétrage sur la qualité des résultats

L'article de Laroche et Langlais (2010) donnent déjà un très bon aperçu de l'impact des données et du paramétrage sur la qualité du lexique extrait. Nous complétons ici leurs observations avec une analyse des résultats obtenus par les approches que nous avons décrites en section 1.2.3. Nous commençons par décrire l'impact des données (section 1.2.4.1) puis l'impact du paramétrage (section 1.2.4.2).

1.2.4.1 Données

Les facteurs qui influencent le plus la qualité des alignements sont liés à la nature des données :

Fréquence des éléments à traduire Un élément est d'autant mieux traduit qu'il est fréquent : son contexte étant calculé à partir d'un plus grand nombre d'occurrences, il est d'autant plus représentatif et donne une meilleure caractérisation sémantique de l'élément à traduire. Ceci est particulièrement bien démontré dans les expériences de Prochasson (2010, p. 65) : les mots les moins fréquents (maximum 25 occurrences) obtiennent environ 7 % sur le Top20 alors que les mots les plus fréquents (au-delà de 800 occurrences) obtiennent une précision de 100 % sur le Top20.

Spécialisation des éléments à traduire Chiao (2004, ch.4, p.77) indique que les éléments spécialisés sont mieux traduits que les éléments de langue générale, quelle que soit leur fréquence. Un résultat similaire est obtenu par Hazem et Morin (2012) : les alignements obtenus sont de meilleure qualité pour le corpus spécialisé alors même que le corpus de presse est de taille beaucoup plus grande. Ceci peut s'expliquer par le fait que les termes sont généralement peu ambigus sémantiquement alors que la polysémie ou les nuances de sens sont fréquentes pour les mots courants : il en résulte un vecteur de contexte plus "flou" et moins discriminant.

Taille des corpus Lorsque les corpus comparables sont volumineux, les termes à traduire ont généralement un grand nombre d'occurrences. Ceci permet donc de construire des vecteurs de contexte plus représentatifs. Mais la taille à elle seule ne suffit pas, les corpus doivent aussi être suffisamment comparables.

Comparabilité des corpus Li et Gaussier (2010) ont défini une mesure de comparabilité des corpus. Cette mesure indique l'espérance de rencontrer la traduction d'un mot source dans le corpus cible (et inversement) [33]. Elle est basée sur la projection d'un dictionnaire bilingue dans le corpus (cf. annexe A.3).

S'appuyant sur cette mesure, Li et Gaussier (2010) démontrent l'impact de la comparabilité des corpus sur la précision des alignements. Li et Gaussier partent d'un corpus original noté C duquel ils extraient deux corpus hautement comparables notés C^1 (comparabilité de 0,882) et C^2 (comparabilité de 0,916). Les lexiques extraits de C^1 et C^2 sont de meilleure qualité que ceux extraits de C : Li et Gaussier gagnent de 5,3 % à 9,5 % de précision sur le Top20.

Spécialisation du lexique bilingue Dans leurs expériences, Laroche et Langlais (2010) ont comparé les résultats obtenus en fonction du degré de spécialisation des entrées du

33. ≪ *For the comparable corpus C, if we consider the translation process from the English part C_e to the French part C_f , a comparability measure M_{ef} can be defined on the basis of the expectation of finding, for each English word w_e in the vocabulary C_e^v of C_e , its translation in the vocabulary C_f^v of C_f.* ≫ (*op. cit.*, p. 645)

lexique bilingue utilisé pour traduire les vecteurs de contextes. Pour un lexique bilingue de 5 000 entrées, les résultats sont légèrement meilleurs lorsque le lexique est composé en partie de lexies spécialisées que lorsqu'il est composé uniquement d'entrées appartenant à la langue générale (La F1-mesure sur le Top1 passe de 38,9 à 39,4 et la MAP passe de 0,471 à 0,473). Il en va de même pour Prochasson (2010) qui décide de renforcer les éléments spécialisés (points d'ancrages). Dans une expérience personnelle, nous avons également observé que la présence d'éléments spécialisés dans le lexique bilingue améliore les résultats (cf. section 1.3.1.4).

Chiao (2004, ch.4, p.85) obtient des résultats contradictoires. Elle indique que l'ajout de mots de langue générale au lexique bilingue améliore les résultats (de 59,4 % à 100 % sur le Top20). Ceci est particulièrement net lorsque les éléments à traduire sont des termes et non des mots de langue générale. Cependant, le premier lexique est composé de 4 963 entrées appartenant uniquement au domaine médical et le lexique "amélioré" contient 6 210 entrées appartenant au domaine médical et à la langue générale. Il est difficile de dire si l'amélioration des résultats est due uniquement à l'ajout de vocabulaire général dans le lexique bilingue ou à la simple augmentation du nombre d'entrées.

1.2.4.2 Paramétrage

Les paramètres en jeu dans l'approche distributionnelle sont :

– Taille et nature du contexte où sont collectés les co-occurrents : phrase, paragraphe, fenêtre de n mots autour de la tête du vecteur, contexte syntaxique.
– Calcul de la significativité des co-occurrences : plusieurs mesures d'association sont possibles comme le taux de vraisemblance, l'information mutuelle ou le TF-IDF. Ces mesures sont détaillées dans l'annexe A.1.
– Calcul de la similarité entre vecteurs, par ex. : mesure cosine, jaccard, distance euclidienne. Ces mesures sont décrites dans l'annexe A.2.

Ces paramètres sont complexes à manipuler. Prochasson (2010, pp. 64-69) montre que la combinaison optimale des paramètres dépend du corpus utilisé et des langues en jeu et qu'il n'est pas possible de la déterminer *a priori*.

De plus, Laroche et Langlais (2010) indiquent que le choix du contexte peut dépendre aussi de l'application visée. Dans leurs expériences, si le contexte correspond au paragraphe, on obtient un très bon rappel sur le Top20, ce qui peut convenir à un outil d'aide à la création semi-supervisée de ressources linguistiques. Par contre, si le contexte correspond à la phrase, on obtient une meilleure précision, ce qui est idéal pour la construction automatique de lexiques bilingues.

Les expériences d'Hazem et Morin (2012) montrent que sur un corpus de presse de grande taille, les meilleurs résultats sont obtenus avec le taux de vraisemblance alors que sur un corpus spécialisé de petite taille, les meilleurs résultats sont obtenus avec l'information mutuelle. Seule la taille du contexte peut être anticipée : les éléments peu fréquents sont mieux traduits lorsque leur vecteur de contexte est calculé sur une fenêtre large et les éléments fréquents sont mieux traduits lorsque leur vecteur de contexte est calculé sur une fenêtre courte (Prochasson, 2010, p. 72).

Une solution à ce problème de paramétrage pourrait être de procéder à un apprentissage du meilleur jeu de paramètres avant l'extraction du lexique. Plusieurs configurations seraient testées en utilisant les paires de traduction présentes dans le lexique bilingue comme lexique d'évaluation. La meilleure configuration obtenue serait ensuite appliquée pour l'alignement des termes.

1.2.5 Limites de l'approche distributionnelle

Si les outils d'alignement en corpus parallèle produisent des paires de traductions correctes dans plus de 80 % des cas (Daille *et al.*, 1994; Macken *et al.*, 2008; Š. Vintar, 2010), c'est loin d'être le cas pour l'alignement à partir de corpus comparables. Nous avons vu dans la section précédente que les résultats obtenus en corpus comparables varient entre 30 % et 89 % sur le $Top10$ et entre 40 % et 94 % sur le $Top20$ selon les paires de langues, le volume et la qualité des données, la nature et la fréquence des éléments à traduire.

Ce contraste entre résultats obtenus en corpus comparables et résultats obtenus en corpus parallèles s'explique pour deux raisons :

Espace de recherche Dans un corpus parallèle, l'espace de recherche est progressivement réduit : on commence par repérer des ancrages (cognats, chiffres), puis on aligne les phrases, puis les traductions des termes sont recherchées au sein de paires de phrases alignées. Dans un corpus comparable, les traductions des termes sont recherchées dans tout le corpus.

Présence de la traduction Dans un corpus parallèle, à moins d'une omission du traducteur, le terme source a toujours une traduction. Dans un corpus comparable, non seulement un terme peut ne pas avoir de traduction, mais il est aussi très difficile de déterminer si cette traduction est potentiellement présente.

À ceci s'ajoute des limites inhérentes à la méthode distributionnelle :

Homogénéité sémantique des vecteurs Si l'élément à traduire est polysémique ou présente des nuances de sens, son vecteur de contexte sera moins homogène sémantiquement, puisque l'élément en question est employé dans des contextes plus variés.

Fréquence des termes L'élément à traduire et sa traduction doivent être suffisamment fréquents : plus les vecteurs sont construits à partir d'un grand nombre de cooccurrences, plus ils sont représentatifs de la distribution du terme.

Les choses se compliquent plus encore lorsque l'alignement est fait en corpus spécialisé, dans le but d'acquérir des listes terminologiques bilingues :

Extraction terminologique préalable Il y a une dépendance à l'extracteur de termes : le terme cible peut se trouver dans le corpus cible mais ne pas avoir été extrait par l'extracteur.

Pertinence du dictionnaire bilingue Le lexique bilingue utilisé pour le transfert peut comporter des traductions qui ne sont pas pertinentes dans le corpus et contribuer à biaiser la projection du vecteur source en langue cible.

Taille des corpus Les corpus spécialisés, parce qu'ils correspondent à une thématique bien définie, sont souvent de petite taille : leur volume est plutôt de l'ordre de centaines de milliers de mots (Prochasson, 2010) que de millions de mots (Rapp, 1999). En conséquence, les termes ont moins d'occurrences et leurs vecteurs sont moins représentatifs.

Termes polylexicaux On va également chercher à aligner des unités polylexicales. Or, comme l'indiquent Morin *et al.* (2007), les termes complexes ont des fréquences plus faibles que les termes simples, ce qui rend leurs vecteurs de contextes moins représentatifs. Si le vecteur du terme complexe est un vecteur composé des vecteurs de mots lexicaux qui composent le terme, alors cela fait baisser l'homogénéité sémantique du vecteur.

Dans cette section, nous avons présenté un état-de-l'art des techniques d'alignement à partir de corpus comparables. Dans la section suivante, nous décrivons la façon dont nous avons prototypé un outil de TAO qui s'appuie sur la méthode distributionnelle pour extraire des lexiques bilingues de corpus comparables.

1.3 Prototypage d'un outil de TAO destiné aux corpus comparables

Dans le contexte industriel de LINGUA ET MACHINA, l'extraction en corpus comparables sera destinée à amorcer la création de ressources linguistiques pour des domaines émergents ou des domaines pour lesquels l'entreprise aura peu d'historique de traduction. Les corpus fournis devraient être des petits corpus spécialisés (moins de 2 millions de mots). Les précisions obtenues seraient, au mieux, entre 34 % sur le Top1 et 76 % sur le top20. Nous pouvons d'ores et déjà anticiper le fait que les traducteurs ne pourront pas se satisfaire d'une simple liste d'alignements terme source → traductions candidates. Pour pallier ces résultats incertains, il sera nécessaire d'accompagner ces alignements d'informations diverses présentées sous la forme d'une fiche terminologique et qui permettront au traducteur de déterminer quelle traduction candidate est la bonne traduction.

Le prototype développé est schématisé dans la figure 1.9 page 40. Il est capable, à partir de textes en langue source et cible, d'extraire des termes et de les aligner grâce à une méthode basée sur l'approche distributionnelle (1.3.1). Dans un second temps, il collecte dans les textes du corpus et sur le Web des informations qui seront ensuite présentées au traducteur sous la forme de fiches terminologiques (1.3.3). Une interface de consultation des lexiques extraits a également été développée (1.3.2).

1.3.1 Implantation d'une méthode d'acquisition de lexiques bilingues

1.3.1.1 Choix d'implantation et données de test

Le but de ce travail était de créer un premier prototype simple qui pourrait ensuite être utilisé pour observer la façon dont les traducteurs appréhendent les lexiques extraits de corpus comparables et évaluer l'apport de ces lexiques à la traduction spécialisée (chapitre 2). Nous avons choisi d'implanter la méthode état-de-l'art pour sa simplicité de mise en œuvre. En effet, les diverses variantes proposées nécessitent soit des ressources particulières (corpus parallèle pour l'apprentissage de patrons lexico-syntaxiques, thésaurus) ; soit le développement d'outils de pré-traitement (extracteur de translittérations et de paires de composés savants) ; soit ce sont des méthodes couteûses en temps, ce qui est toujours problématique en contexte industriel (similarité croisée, similarité interlingue, méthode GLICA). De toutes les approches présentées, nous retenons uniquement l'approche de Morin *et al.* (2004) pour l'alignement des termes polylexicaux.

Comme cela a été montré par Prochasson (2010), il n'est pas possible, en l'état actuel des recherches, de décider *a priori* quelle sera la meilleure combinaison de paramètres à employer. Dans un contexte industriel, nous appliquerons toujours les mêmes paramètres, quelle que soit la taille ou thématique des corpus. Nous avons arbitrairement choisi d'employer le jaccard pondéré comme mesure de similarité et le taux de vraisemblance comme mesure de normalisation des vecteurs de contexte. En revanche, nous avons mené quelques expériences

concernant les interactions entre taille de la fenêtre contextuelle et la fréquence des termes à traduire. En effet, la fréquence des termes à traduire est un élément connu et le système peut adapter ce paramètre de façon automatique. Nous avons également évalué l'apport de ressources spécialisées qui peuvent être un moyen simple et efficace pour améliorer la qualité des lexiques.

Ces tests ont été menés sur un petit corpus spécialisé anglais-français (environ 400 000 mots par langue) portant sur le cancer du sein. Ce corpus est décrit plus en détail dans les sections 2.2.2.1 et 5.1. Pour valider nos expériences, nous avons utilisé deux lexiques de référence :

Lexique spécialisé 177 paires de termes monolexicaux anglais-français collectées par Prochasson (2010, p. 31) à partir de l'UMLS [34] et du Grand dictionnaire terminologique [35].

Lexique généraliste 1 842 paires de mots anglais-français extraits de notre dictionnaire bilingue.

Il y a un recouvrement de 45 entrées entre les deux lexiques. Conformément aux autres travaux de recherche, nous nous sommes assurés que chaque terme à traduire apparaissait au moins 5 fois dans le corpus. La traduction est faite de l'anglais vers le français.

La méthode d'acquisition de lexiques bilingues implantée se décompose en 4 étapes :
- Extraction des termes à aligner (1.3.1.2)
- Collecte des vecteurs de contexte (1.3.1.3)
- Traduction des vecteurs de contexte (1.3.1.4)
- Alignement des termes (1.3.1.5)

1.3.1.2 Extraction des termes à aligner

Les termes à aligner sont extraits des corpus source et cible. Ces termes sont soit des termes polylexicaux extraits par l'extracteur terminologique [36] intégré à SIMILIS (logiciel de mémoire de traduction de LINGUA ET MACHINA), soit des termes monolexicaux (i.e. de simples mots) de catégorie grammaticale NOM, ADJECTIF, ADVERBE, VERBE et ayant un nombre d'occurrences supérieur à 5. Le seuil minimal de 5 occurrences a été choisi pour deux raisons : (i) cela réduit le nombre de termes à aligner et donc le temps de traitement ; (ii) c'est le nombre d'occurrences minimal choisi dans les travaux de recherches présentés en section 1.2 ; nous supposons qu'en-dessous de 5 occurrences, le vecteur de contexte n'est pas significatif. Le prétraitement du corpus (segmentation en mots, lemmatisation, étiquetage morpho-syntaxique) est fait à l'aide de l'analyseur linguistique XELDA [37].

1.3.1.3 Collecte des vecteurs de contexte

Vecteurs de contexte des termes monolexicaux

La taille du contexte a été choisie après plusieurs essais sur notre corpus. Contrairement à Prochasson (2010), nous n'avons pas observé d'influence de la fréquence des termes à traduire sur la taille de fenêtre contextuelle idéale (figure 1.6). Dans notre prototype, la taille du contexte

34. méta-thésaurus médical, www.nlm.nih.gov/research/umls
35. www.granddictionnaire.com
36. Cet outil, bien qu'appelé ≪ *extracteur terminologique* ≫ au sein de LINGUA ET MACHINA n'extrait pas des termes à proprement parler (i.e. désignation d'un concept propre à un domaine) mais des groupes nominaux et verbaux.
37. www.temis.com

sera donc de 3 mots lexicaux à gauche et 3 mots lexicaux à droite du terme à traduire, quelle que soit sa fréquence.

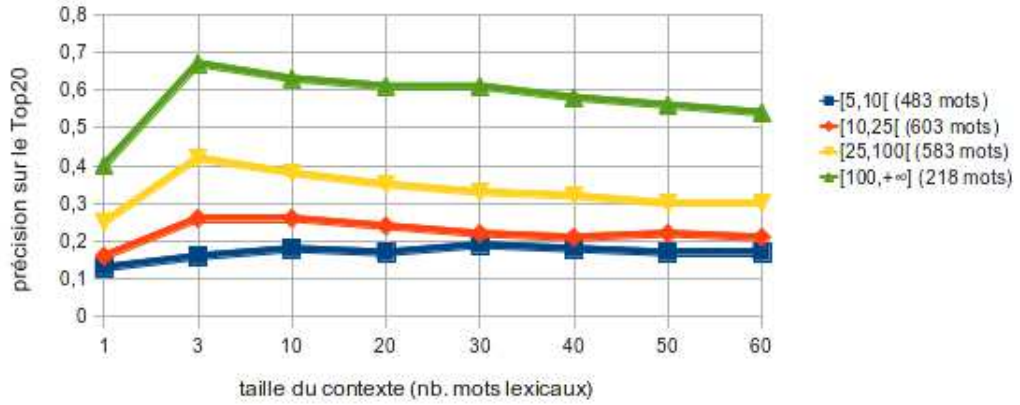


FIGURE 1.6 – Influence de la fréquence des termes à traduire sur la taille de fenêtre contextuelle optimale

Les unités à traduire correspondent aux entrées terminologiques et aux entrées généralistes. Chaque courbe correspond à une gamme de fréquence, entre parenthèses est indiqué le nombre d'entrées dans cette gamme de fréquence.

Le nombre de co-occurrences est normalisé à l'aide du taux de vraisemblance (Dunning, 1993). Son calcul est détaillé en annexe p. 192.

Vecteurs de contexte des termes polylexicaux

À l'instar de Morin *et al.* (2004), nous calculons le vecteur de contexte d'un terme polylexical à partir des vecteurs de contextes de chacun des mots lexicaux qui le compose. Morin *et al.* calculent l'union de ces vecteurs ; pour notre part, nous calculons un vecteur moyen[38] :

- Le terme $cancer\ du\ sein$ possède deux mots lexicaux : $cancer$ et $sein$
- Leurs vecteurs de contextes sont :
  - cancer $= \{(cancer \leftrightarrow 50), (sein \leftrightarrow 30), (traitement \leftrightarrow 25)\}$
  - sein $= \{(sein \leftrightarrow 60), (cancer \leftrightarrow 30), (ablation \leftrightarrow 20)\}$
- Le vecteur de contexte de $cancer\ du\ sein$ est donc :
  - cancer du sein $= \{(sein \leftrightarrow 45), (cancer \leftrightarrow 40), (traitement \leftrightarrow 12.5), (ablation \leftrightarrow 10)\}$

1.3.1.4 Traduction des vecteurs de contextes sources

Les vecteurs de contextes sources sont traduits en langue cible grâce à plusieurs dictionnaires. Nous utilisons tout d'abord le dictionnaire bilingue intégré à notre analyseur linguistique (XELDA). Ce dictionnaire comporte 37 655 entrées (un mot anglais est traduit en moyenne par 1,58 mots français). La taille du dictionnaire a été augmentée à l'aide des liens de traductions entre les entrées de Wikitionary et de Wikipédia. Ceci permet de traduire 18 %

38. Ceci fait gagner environ 2 points de précision sur le Top20 mais nous ne savons pas si cela est propre à notre corpus ou indépendant des données.

de mots en plus dans les vecteurs de contexte. La figure 1.7 montre l'apport de ces deux ressources.

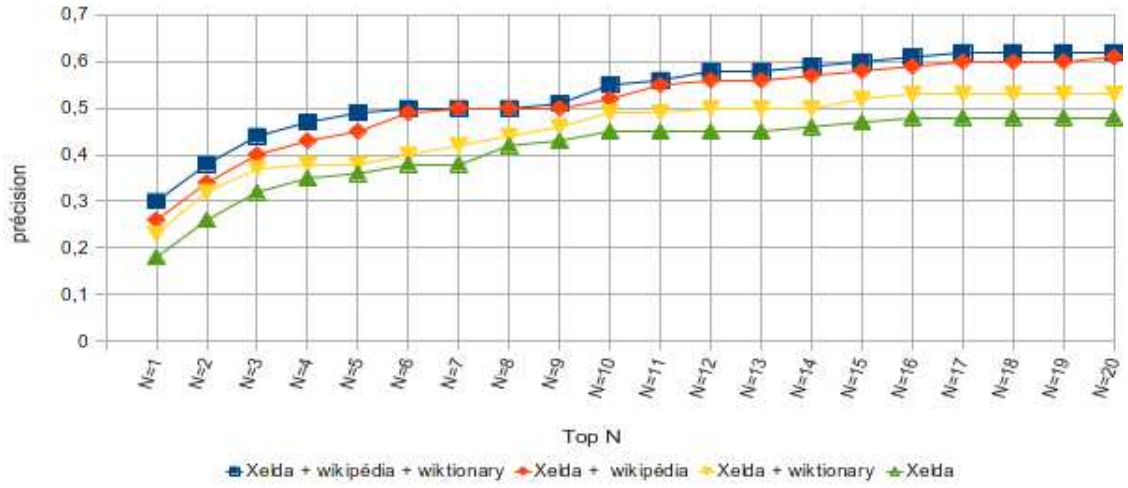


FIGURE 1.7 – Influence du dictionnaire bilingue
(entrées terminologiques)

Lorsqu'un mot de contexte a plusieurs traductions possibles, son taux de vraisemblance est ventilé sur toutes ses traductions, en fonction de la fréquence de chaque traduction dans le corpus cible. Par exemple :

- Le vecteur de $patient$ contient l'association $(related \leftrightarrow 60)$
- $related$ peut être traduit par :
  - $parent$: 10 occurrences dans le corpus cible
  - $proche$: 5 occurrences dans le corpus cible
- Le vecteur de $patient$ traduit en langue cible contiendra donc les associations :
  - $(parent \leftrightarrow 40)$
  - $(proche \leftrightarrow 20)$

Lorsque plusieurs mots de contexte correspondent à une même traduction, les taux de vraisemblance s'ajoutent. Par exemple :

- Le vecteur de $patient$ contient l'association $(rebuilding \leftrightarrow 10)$ et l'association $(reconstruction \leftrightarrow 20)$
  - $rebuilding$ est traduit par $reconstruction$
  - $reconstruction$ est aussi traduit par $reconstruction$
- le vecteur de $patient$ traduit en langue cible contiendra donc l'association :
  - $(reconstruction \leftrightarrow \underline{30})$

1.3.1.5 Alignement des termes

La mesure de similarité utilisée pour comparer les vecteurs source et cible est le jaccard pondéré (Morin *et al.*, 2004) qui est détaillé en annexe p. 193. L'alignement se fait de l'anglais vers le français et nous retenons au maximum les 20 meilleures traductions candidates.

Les expériences de Morin *et al.* sur l'alignement entre termes monolexicaux et polylexicaux ayant donné des résultats mitigés, nous avons séparé ces deux types d'unités. Les termes

monolexicaux ne peuvent être alignés qu'avec d'autres termes monolexicaux et les termes polylexicaux ne peuvent être alignés qu'avec d'autres termes polylexicaux.

Concernant les termes monolexicaux, nous avons rajouté un filtre sur les catégories grammaticales (à l'instar de Sadat *et al.*) : les noms ne peuvent être alignés qu'avec des noms, les adjectifs avec d'autres adjectifs, etc.

La figure 1.8 montre les différences de résultats obtenus sur les deux lexiques. On voit que les termes du lexique spécialisé sont mieux traduits que ceux du lexique généraliste. Ceci s'explique pour deux raisons : (i) les entrées spécialisées sont plus fréquentes en moyenne (166 occurrences contre 54 pour les entrées généralistes) ce qui rend leur vecteur de contexte plus représentatif ; (ii) en tant que vocabulaire spécialisé, ces termes sont probablement moins soumis à la polysémie ce qui rend leurs vecteurs de contextes plus homogènes.

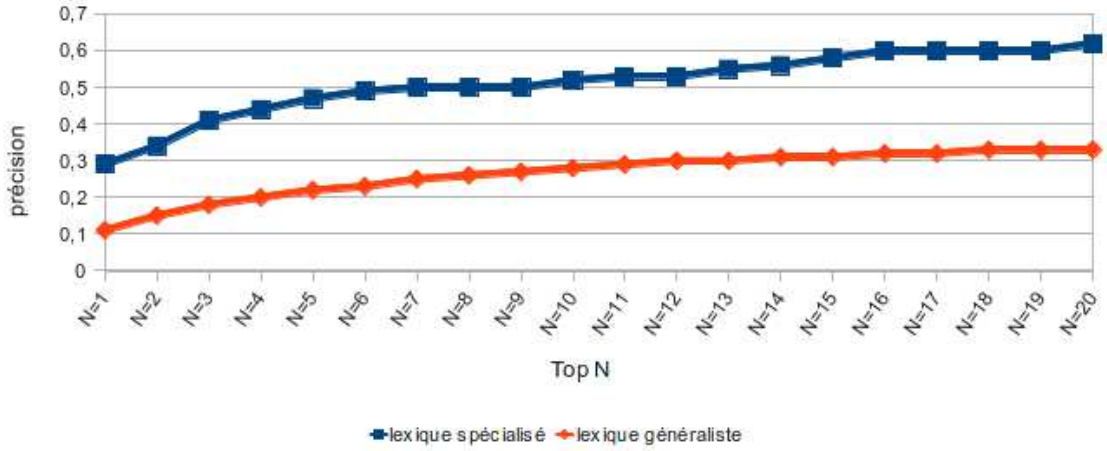


FIGURE 1.8 – Précision au rang N selon le type de termes à traduire

1.3.2 Extraction de fiches terminologiques

Comme évoqué dans la section 1.1.3, une simple liste d'alignements n'est pas suffisante pour les traducteurs : ces derniers doivent pouvoir accéder à des informations qui recontextualisent le terme et lui permettent d'en comprendre le sens.

C'est pourquoi nous avons développé un module d'extraction de fiches terminologiques qui collecte, pour chaque terme, les informations suivantes :

Forme du terme vedette Il s'agit du lemme pour les termes monolexicaux et de la forme fléchie la plus fréquente pour les termes polylexicaux.

Catégorie grammaticale Nous affichons la partie du discours attribuée par XELDA (termes monolexicaux) ou un type grammatical attribué par SIMILIS (termes polylexicaux).

Fréquence Le nombre d'occurrences ou la fréquence (nombre d'occurrences divisé par le nombre de mots dans le corpus) ne serait pas très parlant pour les traducteurs. Nous avons préféré utiliser trois classes calculées à partir de la distribution des occurrences des mots lexicaux :

– usage fréquent (le nombre d'occurrences du terme est supérieur au 90ème percentile) ;

– usage irrégulier (le nombre d'occurrences se situe entre le 51ème percentile et le 90ème percentile inclus) ;
– usage rare (le nombre d'occurrences est inférieur ou égal au 50ème percentile).

Définition Lorsqu'elle existe, nous affichons un lien vers l'entrée de Wikipédia ou de Wiktionary.

Collocations Par collocation, nous entendons un mot qui apparait de façon significativement fréquente dans le contexte gauche ou droit du terme (la mesure statistique employée est le taux de vraisemblance). Par exemple, $lymph\ node$ est associée à $axillary\ lymph\ node$. Pour choisir les collocations, nous ordonnons toutes les collocations trouvées dans le corpus par taux de vraisemblance décroissant et retenons celles qui appartiennent au premier quart supérieur.

Contextes Il s'agit de toutes les phrases dans lesquelles le terme apparaît. Le terme est mis en gras dans la phrase. Un lien vers le document d'où est extraite la phrase est fourni. Les contextes ne sont pas ordonnés.

Variantes Il s'agit uniquement de variantes orthographiques (ex. : traits-d'union, équivalence entre les caractères $œ$ et oe ; alternance $\text{-}or$ / $\text{-}our$ en anglais).

Termes proches Il s'agit de termes ayant au moins un mot lexical en commun avec le terme-vedette, par exemple $tumeur$ est associée à $tumeur\ bénigne,\ croissance\ de\ tumeurs$, etc.

1.3.3 Interface de consultation des lexiques extraits

Les lexiques extraits enrichis des fiches terminologiques sont consultables dans une interface dédiée. Des copies d'écran sont fournies en annexe p. 239 et le prototype est librement consultable à l'adresse http://80.82.238.151/Metricc/InterfaceValidation/ [39]. Cette interface de consultation propose des facilités pour la recherche de termes. Le traducteur dispose d'un champ de recherche qui lui permet d'explorer plus aisément le lexique et des recherches "floues" sont possibles. Par exemple, la requête « `lymph%` » ramènera tous les termes débutant par $lymph\text{-}$. Si aucun terme ne correspond à la requête, une recherche est effectuée directement dans les textes du corpus.

39. Le nom d'utilisateur est "test". Laisser le champ mot de passe vide.

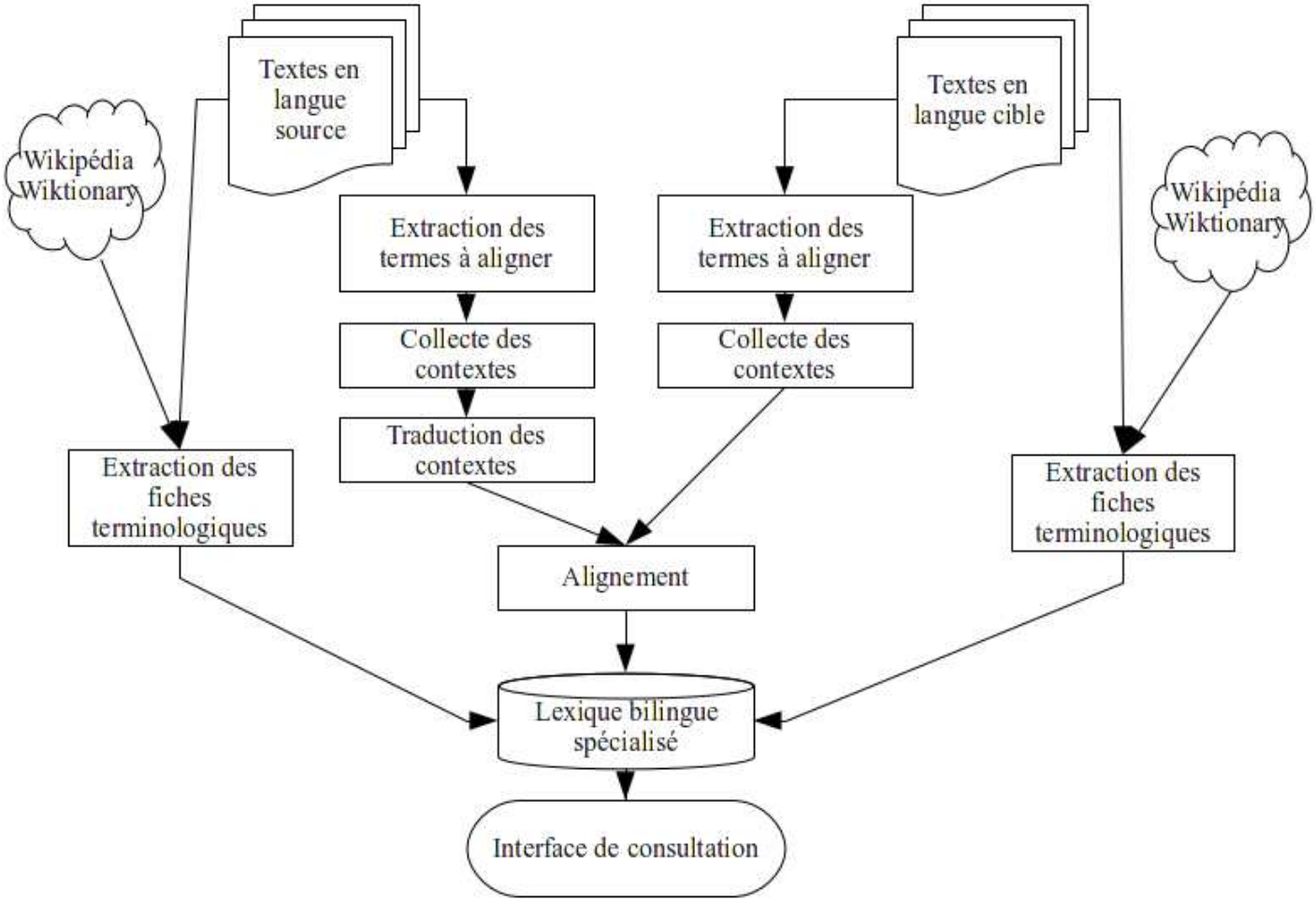


FIGURE 1.9 – Implantation d'une méthode d'acquisition de lexiques bilingues et d'un outil de consultation des lexiques extraits

1.4 Synthèse

Ce chapitre a débuté par un bref rappel historique sur les débuts de la traduction automatique. Nous avons vu que l'*aide à la traduction*, domaine dans lequel nous avons effectué nos recherches, a été impulsée suite aux résultats décevants des premiers outils de traduction automatique. L'aide à la traduction poursuit un objectif moins ambitieux mais plus réaliste : le but est d'outiller le traducteur pour améliorer sa productivité et non de le remplacer.

Jusqu'il y a peu, les logiciels d'aide à la traduction ont toujours nécessité l'existence d'un historique de traduction pour fonctionner. Cet état de fait pose problème lorsque ce dernier n'existe pas (langues peu dotées, domaines émergents). De plus, les recherches en traductologie montrent que les traducteurs préfèrent, pour des raisons qualitatives, accéder à des corpus de textes multilingues non produits par une traduction (corpus comparables). Ces corpus comparables leur permettent de se documenter sur les usages linguistiques et la terminologie employés dans le domaine technique sur lequel ils travaillent. Malgré ce besoin évident, il existe actuellement très peu d'outils capables d'aider le traducteur à explorer les corpus comparables et à y identifier des paires de traductions.

La première partie de notre travail de thèse a donc consisté à prototyper un tel outil. Après un état de l'art des techniques d'alignement de termes à partir de corpus comparables, nous avons opté pour l'implantation de la méthode état-de-l'art qui ne nécessite ni pré-traitements ni ressources linguistiques particulières et est moins coûteuse en temps que des méthodes plus élaborées.

Une fois les termes alignés, notre prototype constitue automatiquement des fiches terminologiques qui contiennent diverses informations susceptibles d'aider le traducteur. Une interface de consultation a été développée pour faciliter l'accès à la ressource.

Si les études en traductologie confirment l'intéret des corpus comparables pour vérifier des hypothèses de traduction ou pour mieux appréhender un domaine technique, nous ne savons pas dans quelle mesure ces corpus et les lexiques qui en ont été extraits peuvent être utilisés comme unique source d'aide à la traduction (cas où aucun historique de traduction n'existe). C'est la piste de recherche que nous proposons d'explorer dans le chapitre suivant.

Chapitre 2

Évaluation applicative des lexiques issus de corpus comparables

Sommaire

Introduction

Le but de ce chapitre est de proposer et d'expérimenter un protocole d'évaluation applicative des lexiques bilingues issus de corpus comparables et destinés à la traduction spécialisée. Pour cela, nous allons nous appuyer sur le prototype développé à LINGUA ET MACHINA que nous avons décrit dans la section 1.3 [1].

À notre connaissance, les lexiques issus de corpus comparables ont été évalués en termes applicatifs uniquement dans le cadre de la recherche d'information cross-lingue et de la traduction automatique.

En recherche d'information cross-lingue, Li *et al.* (2011) augmentent le dictionnaire bilingue utilisé par le moteur de recherche à l'aide de traductions extraites de corpus comparables.

1. Le travail présenté s'inscrit dans un des lots de travail du projet ANR METRICC auquel nous avons participé. Notre participation à ce lot de travail a concerné uniquement la conception du protocole ainsi qu'une première expérimentation du protocole visant à rôder le processus d'évaluation. L'évaluation proprement dite a été menée par le Dr. Planas de l'Université Catholique de l'Ouest avec un plus grand nombre de participants.

Ils démontrent que la combinaison des ressources généralistes et du lexique issu du corpus comparable améliorent significativement les résultats du système (jusqu'à + 0,016 points de MAP).

En traduction automatique, Carpuat *et al.* (2012) utilisent les corpus comparables pour adapter les systèmes de traduction à un domaine de spécialité. L'inclusion des traductions issues du corpus comparable dans le système de traduction fait gagner de 2 à 3 points de BLEU en fonction des corpus.

En ce qui concerne la traduction (humaine) spécialisée, nous allons également adopter une approche contrastive : partant d'une situation de base (le traducteur dispose uniquement de ressources bilingues généralistes), nous allons observer si la mise à disposition de lexiques issus de corpus comparables en plus des ressources généralistes permet effectivement d'améliorer la qualité finale des traductions. Nous emploierons également une deuxième base de comparaison, qui elle, correspondra à la situation habituelle du traducteur, dans laquelle il a à sa disposition toutes sortes de ressources spécialisées en plus des dictionnaires généralistes.

Une fois les traductions produites, se posera alors la question de l'évaluation de leur qualité. Si la recherche d'information crosslingue et la traduction automatique disposent de mesures de référence, ce n'est pas le cas pour la traduction humaine.

C'est pourquoi nous nous sommes penchés dans la section 2.1 sur les méthodologies d'évaluation utilisées à la fois en traduction automatique (section 2.1.1) et en traductologie (2.1.2). Bien que nous souhaitions, au final, évaluer des traductions humaines, il nous a semblé intéressant de s'enquérir des techniques employées en traduction automatique, d'autant plus qu'elles sont nettement plus opérationnelles que celles rencontrées en traductologie.

Dans la section 2.2, nous exposons le protocole d'évaluation que nous avons mis au point et décrivons sa mise en œuvre ainsi que les résultats obtenus. La section 2.3 dresse un bilan de notre expérience et propose de nouvelles pistes de recherche.

2.1 Méthodologies d'évaluation de la qualité des traductions

2.1.1 L'évaluation en traduction automatique

L'évaluation en traduction automatique (TA) remplit deux objectifs. D'une part, il s'agit d'analyser, lors du développement d'un système de TA, les impacts d'une modification du système sur la qualité des traductions. D'autre part, l'évaluation permet de comparer les systèmes entre eux, généralement lors de campagnes d'évaluation de grande envergure. À ces deux objectifs correspondent deux techniques d'évaluation.

Pour une évaluation lors du développement du système, les mesures utilisées sont des mesures calculables automatiquement à partir de traductions de référence, on parle alors d'*évaluation automatique* ou d'*évaluation objective*. Ces mesures, simples et peu coûteuses à mettre en œuvre, restent néanmoins perçues comme les substituts pratiques d'une évaluation bien plus coûteuse mais jugée meilleure : l'évaluation humaine.

L'*évaluation humaine* ou *évaluation subjective* est celle utilisée dans les campagnes d'évaluation du *Statistical Workshop on Machine Translation* de l'ACL dont les résultats des dernières éditions sont donnés par Koehn et Monz (2006) et Callison-Burch *et al.* (2007, 2008, 2009, 2010). Elle consiste à demander à des juges de noter la qualité des traductions. On imagine facilement le coût en termes de temps, d'organisation et de formation des juges, sans compter que les résultats sont difficilement reproductibles. Toutefois, le consensus actuel est

en faveur de l'évaluation humaine, jugée comme plus à même de rendre compte de la qualité d'une traduction.

Dans les parties suivantes, nous rendons compte des techniques d'évaluation automatique (section 2.1.1.1) et des techniques d'évaluation par des humains (section 2.1.1.2).

2.1.1.1 Mesures pour l'évaluation automatique

L'évaluation automatique mesure la qualité d'une traduction de façon indirecte : on n'évalue pas la qualité de la traduction elle-même mais sa ressemblance avec une traduction de référence produite par un traducteur professionnel. À défaut de pouvoir manipuler et comparer des paramètres linguistiques tels que la conservation du sens ou la fluidité du texte, les mesures d'évaluation emploient des indices de surface comme les mots ou suites de mots communs entre traduction évaluée et traduction de référence.

La mesure la plus connue et certainement la plus utilisée est BLEU de Papineni *et al.* (2002). Elle s'appuie sur les critères suivants :

- le nombre de n-grammes de mots communs à la traduction à évaluer et à la traduction de référence, pour n allant de 1 à 4 ;
- les différences de taille (en nombre de mots) entre traduction à évaluer et à la traduction de référence ;
- les possibilités de variation dans la traduction : un même texte pouvant être traduit de plusieurs façons différentes, le score BLEU peut être calculé avec plusieurs traductions de référence, de façon à autoriser plus de variation dans les formulations.

À la suite de BLEU, d'autres métriques ont été proposées dans le but d'améliorer la justesse de l'évaluation des systèmes de TA. Parmi les mesures concurrentes à BLEU, on trouve :

NIST (Doddington, 2002) : équivalente à BLEU, si ce n'est que les n-grammes sont pondérés en fonction de leur fréquence (les n-grammes les plus fréquents étant jugés moins informatifs) et que la précision globale est calculée en utilisant la moyenne arithmétique au lieu de géométrique.

Une adaptation de la F-mesure (Turian *et al.*, 2003) : Cette mesure a été conçue dans le but d'être facilement "interprétable" : elle est empruntée à la recherche d'information. Rappel et précision sont dans ce cas calculés sur le nombre de n-grammes communs à la traduction à évaluer et à la traduction de référence.

Meteor (Banerjee et Lavie, 2005) : associe précision et rappel calculés sur des unigrammes de mots à une mesure prenant en compte l'ordre des mots. En plus des mots identiques, Meteor considère également les mots semblables tels que les variantes morphologiques ou les synonymes. Un des buts de cette mesure est de permettre une évaluation au niveau de la phrase, alors que les autres mesures ne fonctionnent bien que lorsque l'on évalue tout un corpus de traductions.

TER (Snover *et al.*, 2006) : calcule le nombre d'opérations d'édition nécessaires pour parvenir de la traduction évaluée à la traduction de référence.

Ces mesures d'évaluation peuvent elles-mêmes être méta-évaluées en calculant leur corrélation avec des jugements humains. Les métriques sont évaluées sur un corpus de traductions - elles sont dans ce cas plutôt fiables - ou des phrases. D'après Callison-Burch *et al.* (2009), l'évaluation automatique de traductions de phrases reste un problème ouvert : les meilleures métriques sont cohérentes avec les jugements humains dans 54 % des cas, alors que la probabilité d'un accord aléatoire entre métrique automatique et jugement humain est de 0,5.

	Adéquation[a]	Fluidité[b]
5	tout le sens	anglais sans fautes
4	majeure partie du sens	bon anglais
3	une partie du sens	anglais non-natif
2	peu de sens	mauvais anglais
1	aucun sens	incompréhensible

TABLE 2.1 – Échelles d'évaluation de l'adéquation et de la fluidité utilisées par Koehn et Monz (2006)

[a] Échelle originelle : *all meaning, most meaning, much meaning, little meaning, none*.
[b] Échelle originelle : *flawless English, good English, non-native English, disfluent English, incomprehensible*.

Il semble aussi difficile d'identifier une technique d'évaluation automatique qui donnerait des résultats plus fiables qu'une autre. Par exemple, dans l'édition 2009 du *Workshop on Statistical Machine Translation* (Callison-Burch *et al.*, 2009), les mesures les mieux corrélées aux jugements humains sont plutôt des mesures combinant plusieurs mesures ou des mesures basées sur des correspondances entre structures sémantiques et syntaxiques. Dans l'édition 2010 du même workshop (Callison-Burch *et al.*, 2010), les meilleures mesures sont celles qui emploient des informations de surface telles que des n-grammes de lettres. Or, les jeux de données utilisés dans l'édition de 2009 et de 2010 sont quasi-similaires.

La stabilité du comportement de ces mesures "objectives" face aux données est aussi questionnable : les résultats de Callison-Burch *et al.* (2009, 2010) affichent d'importantes variations dans les performances d'une même mesure selon le couple de langue, le sens de traduction ou le niveau de granularité de l'évaluation considérée.

Les mesures d'évaluation objectives ont par ailleurs été critiquées par Blanchon et Boitet (2007) qui expliquent que ces dernières sont d'autant moins corrélées aux jugements humains que la qualité de la traduction augmente. Ils décrivent également une expérience consistant à faire évaluer des traductions automatiques post-éditées par des humains. Ces traductions sont jugées de qualité moindre que des traductions produites par des systèmes automatiques, et ce, sur la base de mesures telles que BLEU, NIST, etc. Les auteurs s'appuient sur cette expérience pour rappeler que ces mesures ne sont pas directement liées à la qualité des traductions mais qu'elles évaluent seulement la ressemblance avec une traduction de référence, référence qui est, de plus, considérée comme discutable, tout particulièrement en traduction.

2.1.1.2 Évaluation humaine de la TA

L'évaluation humaine consiste à présenter des traductions de phrases à des humains qui doivent alors juger de leur qualité. Cette méthodologie a évolué au cours des années. En 2006, Koehn et Monz demandent à des juges de donner deux notes aux traductions sur une échelle de 1 à 5 (cf. tableau 2.1) : l'une concerne l'adéquation entre traduction et texte d'origine (conservation du sens) et l'autre concerne la fluidité (bonne formation grammaticale). L'annotation des traductions se fait via une interface. Chaque juge peut voir le texte d'origine et annote cinq traductions à la fois, de façon à lui permettre de contraster les phrases et obtenir

	accord inter-annotateur	accord intra-annotateur
fluidité	0,25	0,54
adéquation	0,23	0,47
classement des phrases	0,37	0,62
classement des constituants	0,54	0,74

TABLE 2.2 – Accord intra- et inter- annotateur lors du Workshop on Statistical Machine Translation de 2007 - (Callison-Burch *et al.*, 2007)

	temps moyen par élément (secs.)
fluidité et adéquation	26
classement des phrases	20
classement des constituants	11

TABLE 2.3 – Temps d'annotation lors du Workshop on Statistical Machine Translation de 2007 - (Callison-Burch *et al.*, 2007)

un meilleur jugement.

En 2007, Callison-Burch *et al.* testent deux autres méthodes :

Classement des phrases Les juges doivent ordonner les phrases de la moins bien à la mieux traduite (avec la possibilité d'égalités).

Classement de constituants syntaxiques Même principe que le classement des phrases, sauf qu'il s'applique à des traductions de syntagmes.

Ces deux méthodes ont été rajoutées pour restreindre les possibilités d'interprétation car il s'est avéré que les échelles d'adéquation et de fluidité laissent beaucoup trop de place à la subjectivité. Par exemple, il est difficile de cerner la valeur de *majeure partie du sens* (« *much meaning* ») dans l'échelle d'adéquation. De plus, les juges ont du mal à noter séparément l'adéquation et la fluidité. À l'inverse, le classement, qui ramène l'évaluation à une simple comparaison, est plus simple à appréhender.

Les deux méthodes ont été comparées en mesurant le degré d'accord inter- et intra-annotateurs. La mesure utilisée est le $Kappa$ de Carletta (1996) (cf. annexe A.6). Comme indiqué dans le tableau 2.2, la méthode de classement obtient un accord intra- et inter-annotateur plus élevé. De plus, elle permet une annotation plus rapide (tableau 2.3). Le classement des constituants syntaxiques est lui même plus fiable et plus rapide que le classement des phrases.

Dans l'édition 2008 du workshop, Callison-Burch *et al.* abandonnent la méthode d'évaluation basée sur l'adéquation et la fluidité. À la place, ils proposent une méthode plus simple, dans laquelle on présente aux juges des traductions de constituants syntaxiques et on leur demande d'indiquer si la traduction est acceptable ou pas. Les juges ont aussi la possibilité d'indiquer qu'ils ne sont "pas sûrs". Cette méthode a obtenu le plus haut taux d'accord : 0,64 et 0,86 - respectivement inter- et intra- annotateur. Finalement, dans les éditions 2009 et 2010, seule la

méthode consistant à classer les traductions a été gardée.

Toute la difficulté de l'évaluation humaine touche à sa subjectivité et à son manque de reproductibilité, puisque, comme le montre l'accord inter-annotateur, une même traduction n'est pas toujours jugée de la même façon par les juges, ce qui peut faire douter de la fiabilité de ces jugements. La solution consiste alors à juger la traduction sur la base d'un grand nombre de jugements, ce qui permet de neutraliser les différences individuelles. Blanchon et Boitet (2007) remarquent que les juges ont tendance à devenir plus sévères sur la durée, ils indiquent aussi que le fait de former les juges augmente le taux d'accord. La préparation en question consiste à fournir aux juges une fiche d'instruction et à effectuer une première évaluation à blanc. Les divergences sont ensuite discutées afin de normaliser la notation.

2.1.2 L'évaluation en traductologie

En traductologie, la question de l'évaluation est en elle-même un champ de recherche. Williams (2004) y réfère par le vocable *Appréciation de la Qualité des Traductions* (AQT) [2]. L'AQT trouve ses origines dans la critique de la traduction, activité qui consiste à commenter la qualité littéraire du texte traduit, avec ou sans référence au texte original. La discipline se développe dans les années 70 où la traductologie souhaite se doter de modèles avec un double objectif : donner à l'industrie de la traduction des moyens de contrôler la qualité de ses produits et permettre aux écoles de traduction d'évaluer leurs étudiants. L'évaluation en traductologie est différente de l'évaluation en TA à plusieurs niveaux :

- le niveau d'exigence est supérieur : on évalue des traductions faites par des professionnels, et non pas la ressemblance avec une traduction humaine ;
- la TA évalue les traductions en relation à d'autres, le but étant de classer des traductions de façon à classer les systèmes qui les ont produites ; la traductologie évalue les traductions en elles-mêmes, il ne s'agit pas de comparer les traducteurs professionnels entre eux ;
- la TA utilise une traduction professionnelle comme référence, la traductologie n'a pas de référence de qualité, le juge lui-même est la référence.

On trouve un panorama de l'AQT dans les articles de Williams (2004) et Secară (2005). Larose (1998) propose une réflexion théorique sur la méthodologie de l'évaluation des traductions. Williams (2004) distingue deux types de modèles : les modèles quantitatifs et les modèles non-quantitatifs. Les modèles quantitatifs (section 2.1.2.1) sont plutôt pragmatiques, ils doivent permettre de donner un score de qualité à toute traduction. Ces modèles produisent des grilles d'évaluation utilisées dans l'industrie de la traduction ou dans l'enseignement. Les modèles non quantitatifs (section 2.1.2.2) constituent plutôt des approches théoriques du problème de l'évaluation et se concentrent surtout sur la définition de ce qu'est une "bonne" traduction.

2.1.2.1 Modèles quantitatifs

La plupart des modèles quantitatifs ont été conçus par et pour des organismes qui cherchaient un moyen de maîtriser la qualité de leurs traductions. Le premier modèle d'AQT a été crée par le Bureau de la traduction du Canada en 1976. Ce modèle, appelé Sical (Système canadien d'appréciation de la qualité linguistique) est décrit par Williams (2001) et Secară (2005). Il sépare erreurs de langue (intelligibilité, grammaticalité, idiomaticité) et erreurs de

2. *Translation Quality Assessment* (TQA).

		Nombre maximal de défauts dans une tranche de 4000 mots	
Cote	**Qualité**	**Défauts graves**	**Défauts mineurs**
A	supérieure	0	0 à 6
B	acceptable	0	7 à 12
C	à revoir	1	13 à 18
D	innacceptable	1 et +	18 et +

TABLE 2.4 – Grille d'évaluation du modèle Sical - (Larose, 1998; Williams, 2004)

transfert (conservation du sens). Chaque erreur est jugée comme grave ou mineure, la gravité étant déterminée sur la base des conséquences supposées de l'erreur (par exemple, pour la traduction d'un manuel d'utilisation : erreur pouvant engendrer une utilisation dangereuse). La qualité globale de la traduction est estimée sur le nombre et le type d'erreurs rencontrées dans un passage de 4000 mots sélectionné aléatoirement (voir tableau 2.4).

La grille Sical a donné lieu à d'autres variantes par la suite. De même, diverses grilles d'évaluation ont été proposées par des organismes comme l'ATA (American Translators Association), la SAE (Society of Automotive Engineers) et la LISA (Localization Industry Standards Association) ou l'agence de traduction ITR.

Toutes ces grilles d'évaluation suivent le même schéma : elles consistent en une typologie d'erreurs de traduction, chaque type d'erreur étant associé à un coût représentant sa gravité. Certaines, comme le SEPT - décrit dans Larose (1998) - vont jusqu'à dénombrer 675 types d'erreurs. Larose (1998) remarque à juste titre que tous les modèles séparent erreurs de transfert (sens ou contenu) et erreurs de langue (forme ou expression), avec une prédominance du sens sur la forme. On retrouve le même principe dans les premières versions de l'évaluation humaine de la TA où l'on demande aux juges de noter séparément adéquation (erreurs de transfert) et fluidité (erreurs de langue).

En comparaison au domaine de la TA, on peut être surpris par l'absence de processus de validation ou de comparaison des différents modèles proposés. Bien que généralement conscients de la subjectivité des jugements humains, rien n'est fait pour tenter de la quantifier. On pourrait tout à fait envisager de comparer ces modèles sur la base d'un accord inter-annotateur. Il en est de même pour le coût en temps. Mise à part pour le Sical, qui est supposé prendre 1h pour évaluer un passage de 4 000 mots, aucun auteur n'indique le temps que prend une évaluation en suivant telle ou telle grille.

Ces modèles quantitatifs, à visée opérationnelle, sont assez critiqués par les tenants des modèles non-quantitatifs, comme nous le verrons dans la partie suivante.

2.1.2.2 Modèles non-quantitatifs

Une des principales critiques des tenants des modèles non-quantitatifs envers les grilles d'évaluation utilisées dans l'industrie est le niveau d'analyse de ces grilles. En effet, la plupart des modèles quantitatifs restent au niveau des mots et de la phrase et se préoccupent rarement du niveau discursif. Les grilles d'évaluation sont monolithiques, supposées valables pour toutes les traductions, sans prendre en compte la fonction du texte, la situation de communication dans

laquelle il a été produit ou les attentes du commanditaire de la traduction.

Williams (2004) par exemple, suggère de passer d'une approche micro-textuelle (celle des modèles quantitatifs, basée sur la phrase) à une approche macro-textuelle qui repose sur l'analyse et la comparaison de la structure argumentale des textes source et cible. Williams découpe chaque texte en six modules d'argumentation, qui sont indépendants du genre, type, fonction ou domaine du texte traduit. Dans ce cadre théorique, une bonne traduction est une traduction qui reprend chacun des modules présents dans le texte source et reproduit fidèlement leur contenu et relations. L'auteur considère l'absence d'un des modules comme une erreur majeure, mais ne donne pas plus de détails.

Reiss (1971) propose une approche fonctionnelle de la traduction. Elle affirme que les critères d'évaluation doivent dépendre de la fonction du texte. Pour cela, elle spécifie quatre types de textes :

Textes centrés sur le contenu - ≪ **content-focused** ≫ Ce sont des textes dénotatifs, référentiels, qui privilégient la description de faits : articles de presse, travaux scientifiques, notices. Le traducteur adapte totalement la forme du texte à la langue cible, il amène le texte au lecteur, en respectant en priorité le sens du texte source.

Textes centrés sur la forme - ≪ **form-focused** ≫ Ce sont les textes ayant une fonction poétique, par exemple les textes littéraires, artistiques. Le traducteur amène le lecteur au texte, en respectant en priorité la forme du texte source, le traducteur jouit d'une plus grande liberté au niveau du transfert du sens.

Textes incitatifs - ≪ **appeal-focused** ≫ Ce sont les textes conatifs, destinés à provoquer une réaction chez leur lecteur : publicité, propagande. Dans ce cas, la traduction devient une adaptation libre : son but premier est de conserver l'effet du texte sur le lecteur, il n'y a pas d'obligation de respect ni de la forme ni du sens.

Textes audio-médiaux ≪ **audio-medial** ≫ Ce sont les textes qui ne sont pas transmis par le support écrit : pièces de théâtres, discours. Le traducteur doit adapter le texte à son environnement et à la manière dont il sera prononcé : mouvement des lèvres dans le sous-titrage, rythme dans les chansons. Cette dernière catégorie est assez bancale car elle se situe à un niveau de classification supérieur aux trois autres (oral vs. écrit) : un texte peut être incitatif et audio-médial (publicité radio vs. publicité sur affiche), centré sur la forme et audio-médial (pièce de théâtre vs. œuvre littéraire), etc.

À l'instar des méthodes d'évaluation présentées jusqu'ici, Reiss distingue sens et forme et donne des critères sur lesquels évaluer les traductions. Par contre, elle ne donne pas de grille d'évaluation à proprement parler. Elle différencie les éléments linguistiques comme les aspects sémantiques, lexicaux, grammaticaux, stylistiques des éléments extra-linguistiques (situation de communication, sujet, époque, lieu, audience, locuteur, enjeux affectifs). Chaque critère a une influence plus ou moins grande sur la qualité globale de la traduction, en fonction du type de texte traduit. Par exemple, dans le cas des textes orientés vers le contenu, l'équivalence totale entre éléments sémantiques du texte source et cible est obligatoire, alors que le non-respect de l'équivalence stylistique est tolérable, voire recommandé si cela permet, en adaptant le texte source à la langue cible, un meilleur transfert du sens.

2.1.3 Discussion

Nous avons vu dans cette section diverses méthodes d'évaluation des traductions : évaluation automatique et humaine en TA ; approches quantitatives et non-quantitatives en

traductologie. La traductologie n'offre pas de solutions adaptées à notre objectif : soit les approches restent trop théoriques soit les modèles proposés sont trop complexes à mettre en œuvre.

En TA, l'évaluation automatique fait peu de sens. Bien que l'on perde en reproductibilité, nous pensons, à l'instar de Blanchon et Boitet (2007), que l'évaluation automatique est une évaluation très indirecte de la qualité des traductions, qui évalue plus la ressemblance à la référence que la réelle qualité des traductions, telle qu'elle peut être perçue par un humain.

Nous nous inspirerons donc des méthodologies d'évaluation humaine de la TA en essayant de gérer du mieux possible la subjectivité inhérente à cette méthode. Une des solutions semble résider dans l'entraînement et la formation des évaluateurs. De même, le recours à plusieurs juges est un moyen de lisser les préférences individuelles. L'utilisation de mesures d'accord inter-annotateurs de style Kappa permettra de quantifier la fiabilité des jugements et de s'assurer qu'il existe un accord inter-juges suffisant. Idéalement, il faut aussi multiplier le nombre de traducteurs et de domaines de façon à collecter plus de données et augmenter la représentativité de l'évaluation.

Enfin, comme nous travaillons en domaine spécialisé, l'évaluation devra tout particulièrement porter sur la qualité de la traduction des termes, des lexies spécialisées et de tout autre unité lexicale dont l'usage dévie de la langue courante. La mise en œuvre concrète du protocole est décrite dans la section suivante.

2.2 Conception et expérimentation d'un protocole d'évaluation applicative

2.2.1 Reflexions méthodologiques

La mise au point du protocole d'évaluation à soulevé plusieurs questions :

Critères et objet de l'évaluation Quels critères choisir pour déterminer la qualité d'une traduction ? Faut-il évaluer la qualité du texte traduit dans son entier ou seulement certains aspects ?

Expertise dans le domaine de spécialité Sachant que nous n'avons pas d'expert du domaine à notre disposition, à partir de quelle référence évaluer les traductions produites ?

Situations contrastées À quelles autres ressources doit être comparé le lexique issu de corpus comparables ?

2.2.1.1 Critères et objet de l'évaluation

La qualité d'une traduction est difficile à évaluer comme l'ont mis en exergue les travaux évoqués en section 2.1. Si on retrouve universellement les deux critères du sens et de la forme, il est difficile d'affiner plus avant la question. Dans le monde de la traduction, aucun barème ou mode d'évaluation ne fait consensus. Et pour cause : en compilant les diverses grilles d'évaluation et travaux non-quantitatifs, on se rend compte que la qualité globale d'une traduction dépend de l'interaction complexe de nombreux paramètres linguistiques (orthographe, lexique, sémantique, style, structure argumentale) comme extra-linguistiques (lieu, époque, audience...). De plus, leur interaction et le poids de chaque paramètre seraient

réglés par la fonction du texte et les attentes du commanditaire de la traduction. Pour reprendre les mots de Larose (1998, p. 2), on se trouve face à un ≪ *fol magma de variables variables* ≫.

Comment, alors, mesurer l'impact des lexiques extraits des corpus comparables sur la qualité des traductions ? S'attend-on à ce qu'ils influent directement sur la qualité globale des traductions ou à ce qu'ils agissent uniquement sur quelques paramètres, qui à leur tour, influencent la qualité de la traduction ? Quels sont les paramètres les plus importants dans le cas d'une traduction spécialisée ?

Pour répondre à ces questions, nous poserons qu'un lexique bilingue spécialisé a pour but d'aider le traducteur lorsqu'il/elle bute sur un terme ou une expression propre au domaine de spécialité du texte. Deux cas de figure sont possibles :

Problème de décodage Il se peut que le sens du terme ou de l'expression soit opaque : le lexique, étant enrichi d'informations extraites du corpus, donne accès à un concordancier, donne des liens vers des entrées semblables et éventuellement fournit une définition. Toutes ses informations se conjuguent pour permettre au traducteur de cerner le sens du terme.

Problème d'encodage Il se peut que le traducteur comprenne le terme mais ne sache pas comment le traduire, i.e il ne connaît pas son équivalent en langue cible : le lexique propose alors des traductions candidates et chaque traduction candidate est assortie d'informations contextuelles permettant de faire le bon choix de traduction. Dans le cas où le traducteur a une intuition de traduction qui n'apparaît pas parmi les traductions candidates, le logiciel de gestion terminologique lui permet de chercher cette traduction potentielle dans le corpus duquel a été extrait le lexique.

Les lexiques spécialisés bilingues sont donc supposés agir sur les deux méta-critères de qualité que sont le transfert du sens (décodage) et la production d'une forme adéquate (encodage). Nombre de paramètres de qualité sont cités par les travaux de traductologie, pourtant, les ressources spécialisées ne sont censées agir que sur quelques uns, par exemple l'orthographe, le respect des normes terminologiques, l'idiomaticité, l'interprétation correcte du terme source. On ne peut donc pas juger la valeur ajoutée de nos lexiques spécialisés sur la base de la qualité globale de la traduction, puisque cette qualité dépend d'autres paramètres sur lesquels nos alignements ont peu ou pas d'influence : grammaticalité, omissions / insertions, cohérence, respect de la structure argumentale, localisation, choix du registre...

Nous essaierons donc de mesurer uniquement la capacité de nos lexiques bilingues à aider le traducteur à traduire des termes ou expressions spécialisées sur lesquels il/elle bute. Pour cela, nous demanderons aux traducteurs de noter les expressions qu'ils ont eu du mal à traduire ainsi que la traduction qu'ils ont finalement retenue. L'évaluation portera sur l'exactitude de la traduction retenue.

Nous aurons donc très probablement à évaluer des traductions de syntagmes, de locutions ou d'unités monolexicales, comme l'on fait Callison-Burch *et al.* (2008). En plus de permettre de mieux cibler l'évaluation, le recours à des segments inférieurs à la phrase aura également pour effet de réduire le temps d'annotation et de faciliter la tâche des juges, comme l'ont montré Callison-Burch *et al.* (2008).

Comme la traduction est faite en dehors de toute finalité professionnelle, on ne cherche pas à mettre au point une grille d'évaluation qui, dans le style de l'AQT, associerait des coûts différents aux fautes d'orthographe, au manque d'idiomaticité, etc. On se contente d'utiliser les critères généraux du sens (adéquation) et de la forme (fluidité). Pour cela, nous nous conformons aux recommandations de Reiss, qui recommande de donner la priorité au sens

plutôt qu'à la forme lorsqu'on évalue des traductions de textes centrés sur le contenu[3]. Nous utiliserons trois catégories pour juger de la qualité des traductions (résumées dans le tableau 2.5) :

EXACT Le terme choisi est le terme de référence ou l'expression consacrée en usage dans le domaine, ex : $distributional\ semantics \rightarrow sémantique\ distributionnelle$

ACCEPTABLE Il ne s'agit pas du terme ou de l'expression de référence et la formulation peut-être maladroite mais le traducteur est quand même parvenu à donner une équivalence sémantique et le sens est conservé, ex : $distributional\ semantics \rightarrow sémantique\ distributionnaliste$

FAUX La traduction est incorrecte : le traducteur n'a pas compris le terme et/ou il n'est pas parvenu à donner une équivalence sémantique, ex : $distributional\ semantics \rightarrow sémantique\ distribuée$

	transfert du sens	**respect de la forme**
EXACT	$+$	$+$
ACCEPTABLE	$+$	$-$
FAUX	$-$	$-$

TABLE 2.5 – Critères pour juger la qualité des traductions

2.2.1.2 Expertise sur le domaine de spécialité

Le fait de travailler avec des textes spécialisés rajoute un obstacle supplémentaire à l'évaluation. En plus de maîtriser les langues source et cible, le juge doit aussi être expert dans les domaines de spécialité des textes à traduire. En l'absence d'expert disponible, la solution retenue consiste à employer des textes spécialisés qui existent en langue source et en langue cible et qui ont été produits par un expert du domaine. La version cible des textes constituera notre traduction de référence sur laquelle évaluer les traductions produites. Les résumés d'articles scientifiques constituent une ressource parfaite pour cet usage. Le fait que l'auteur soit un expert du domaine assure sa légitimité en termes de choix terminologiques. Les articles étant nécessairement révisés avant publication, cela garantit que des fautes de langue éventuelles ont été corrigées. Enfin, la référence n'est pas une traduction mais bien une deuxième version du texte, produite par une même personne dans une autre langue.

En plus d'une traduction de référence, les juges pourront aussi s'aider d'une base terminologique qui peut leur permettre de valider les cas où le traducteur n'a pas employé le terme attendu mais une variante ou une expression de sens équivalent. Les traductions à juger sont toujours montrées en contexte : le juge a accès aux phrases source et cible qui contiennent le terme ainsi qu'aux documents d'origine.

2.2.1.3 Situations contrastées

La mise au jour de la valeur ajoutée des lexiques se fait par contraste : nous comparons le résultat de la traduction d'un même texte source traduit à l'aide de ressources linguistiques

3. Les textes spécialisés correspondent à des textes centrés sur le contenu dans la théorie de Reiss (1971).

différentes. Ces situations différentes dans lesquelles les traducteurs traduisent chacun un même texte à l'aide de ressources différentes sont appelées *situations de traduction*. Nous avons déterminé trois situations de traduction :

Situation minimale Dans cette situation, les traductions sont faites à l'aide de ressources minimales, une sorte de "kit de survie" du traducteur, c'est-à-dire un dictionnaire bilingue généraliste, un dictionnaire monolingue généraliste en langue source et un dictionnaire monolingue généraliste en langue cible. Dans ce cas, on considère que les traductions seront également de qualité minimale : il ne faut pas descendre sous ce seuil de qualité.

Situation maximale Dans cette situation, les traductions sont faites grâce à un maximum de ressources, on considère alors qu'il est impossible d'obtenir de meilleures traductions. Dans ce cas le traducteur dispose des ressources bilingues et monolingues généralistes ainsi que de diverses ressources terminologiques.

Situation cible Cette situation est la situation évaluée, elle correspond au cas où les traductions sont faites à l'aide de la ressource que l'on souhaite évaluer. Dans ce cas, les traducteurs disposent du "kit de survie" (ressources généralistes) et d'un lexique bilingue extrait d'un corpus comparable spécialisé.

Avec ce protocole basé sur différentes situations de traductions, il faut éviter un *effet d'apprentissage* qui apparaît lorsqu'un même traducteur traduit des textes issus d'un même domaine dans plusieurs situations de traduction. En effet, lorsqu'un traducteur traduit un texte à l'aide d'une ressource donnée, il garde forcément en mémoire une partie des traductions des termes sur lesquelles il a buté. Si ce même traducteur doit ensuite retraduire le texte en question (ou un texte du même domaine de spécialité) dans une autre situation, il réutilisera forcément les traductions apprises lorsqu'il a traduit le texte pour la première fois. La deuxième situation de traduction est alors favorablement avantagée. Il faut donc faire en sorte qu'un traducteur ne traduise jamais des textes issus d'un même domaine dans des situations de traduction différentes.

Les choix méthodologiques étant argumentés, nous décrivons dans la section suivante notre expérimentation du protocole.

2.2.2 Expérimentation du protocole

Cette partie décrit une première expérimentation de la méthode d'évaluation. Les données utilisées sont décrites en section 2.2.2.1, le déroulement de l'évaluation est exposé en section 2.2.2.2

2.2.2.1 Données

Nous avons mené l'évaluation sur un sens de traduction (de l'anglais vers le français) et deux thématiques : cancer du sein (domaine médical) et sciences de l'eau (domaine de l'environnement).

	CANCER DU SEIN	SCIENCES DE L'EAU
textes scientifiques	3 résumés d'articles 508 mots portail *Elsevier*	3 résumés d'articles 499 mots revue *Sciences de l'eau*
textes de vulgarisation	1 page web 613 mots site *Société canadienne du cancer du sein*[a]	1 page web 425 mots site *Lenntech* sur le traitement des eaux[b]

[a] http ://www.cbcf.org/
[b] http ://www.lenntech.com/

TABLE 2.6 – Taille, origine, thématique et degré de spécialisation des textes à traduire

Corpus comparables et lexiques extraits

Le corpus portant sur la thématique CANCER DU SEIN comporte environ 400 000 mots par langue, il a été constitué manuellement à partir de publications scientifiques collectées sur le portail *Elsevier*[4] et d'articles de sites Internet de vulgarisation à destination des patientes et de leur proches. Il y a une répartition équivalente entre textes scientifiques et vulgarisés. Les textes proviennent de sources françaises.

Le corpus portant sur la thématique SCIENCES DE L'EAU comporte deux millions de mots par langue. Il a été constitué automatiquement en aspirant les sites de la revue francophone *Sciences de l'eau*[5] et de la revue anglophone *Water Science Technology*[6]. Pour la partie francophone, nous avons pu obtenir les articles scientifiques entiers au format PDF (la conversion a été faite à l'aide de l'utilitaire Unix `pdf2txt` suivi de quelques heuristiques filtrant les entêtes et pieds de pages). Pour la partie anglophone, nous avons pu récupérer uniquement le résumé des articles au format HTML qui a ensuite été converti au format texte.

Nous avons mesuré la comparabilité des corpus en utilisant la mesure de Li et Gaussier (2010). La comparabilité du corpus CANCER DU SEIN est de 0,74 ; celle du corpus SCIENCES DE L'EAU est de 0,77. Des extraits des corpus sont visibles dans l'annexe B.1.

Les lexiques ont été extraits en suivant la méthode décrite dans le chapitre 1, section 1.3. Ils peuvent être consultés en ligne[7].

Textes à traduire

Nous avons sélectionné huit textes anglais pour lesquels il existe une traduction en français. Les textes à traduire sont les textes anglais ; la version française sera utilisée pour évaluer le travail des traducteurs. Les textes sont répartis équitablement entre les thématiques et le degré de spécialisation comme le montre le tableau 2.6. Aucun de ces textes n'apparaît dans les corpus comparables utilisés pour extraire les lexiques. Des extraits des textes sont consultables en annexe B.2. Il faut noter

4. http ://www.elsevier.com/
5. http ://www.rse.inrs.ca/
6. http ://www.iwaponline.com/wst/
7. http ://80.82.238.151/Metricc/InterfaceValidation/ ; le nom d'utilisateur est *test*, laisser le champ "mot de passe"

Les textes ont été choisis intuitivement, en respectant un critère unique : que leur sujet corresponde à la thématique du lexique. Les textes scientifiques sont des publications scientifiques et proviennent de la même source que les textes du corpus d'acquisition. Les textes vulgarisés proviennent, pour la thématique CANCER DU SEIN, d'un site de prévention du cancer du sein édité par le gouvernement canadien (nous avons choisi un texte expliquant les avantages et les risques du dépistage). Pour la thématique SCIENCES DE L'EAU, les textes proviennent du site d'une entreprise commercialisant des solutions de traitement des eaux (le site comporte des pages de vulgarisation expliquant aux clients le principe d'adsorption par charbon actif).

Ressources employées dans les situations de traduction

Situation minimale Les textes sont traduits sans aucune ressource spécialisée. Le traducteur a uniquement accès à trois ressources généralistes en ligne :
- Le Larousse bilingue français/anglais [8] et anglais/français [9]
- Le Larousse monolingue français [10]
- Le Cambridge monolingue anglais [11]

Situation cible En plus des ressources généralistes de la situation minimale, le traducteur a accès aux lexiques extraits des corpus comparables spécialisés qu'il consulte grâce à l'interface présentée dans la section 1.3.3. En plus des équivalences traductionnelles, le traducteur a donc accès aux fiches terminologiques. Il peut également vérifier une hypothèse de traduction en la cherchant dans le corpus.

Situation maximale En plus des ressources généralistes de la situation minimale, le traducteur a un accès total à Internet où il peut consulter les différentes ressources spécialisées, concordanciers, forums de traductions, etc. Il peut aussi utiliser les moteurs de recherche pour contextualiser le terme à traduire ou vérifier une intuition. Cependant, on lui interdit les sites dont sont extraits les textes à traduire et les corpus d'acquisition ainsi que le site de la base de données terminologique TERMIUM [12] qui est utilisée plus tard lors de l'évaluation des traductions.

Traducteurs et juges

Disposant de peu de moyens humains (3 personnes) pour expérimenter le protocole, nous avons dû faire quelques entorses méthodologiques : il y a eu des collisions entre les rôles d'organisateur/traducteur et traducteur/juge. Le traducteur 1, auteure de la thèse, a aussi organisé l'évaluation. Sa seule expérience en traduction spécialisée consiste en des exercices de traduction de textes journalistiques de niveau L3 LLCE Anglais. Les traducteurs 2 et 3 étaient des étudiants de dernière année d'école de traduction. Ils ont aussi jugé et classé les traductions (l'anonymisation empêchant les juges de savoir qui ou dans quelle situation avait été produites les traductions). La langue maternelle des trois personnes est le français. Aucun des traducteurs n'est familier avec la thématique des sciences de l'eau ou avec celle du cancer du sein.

---

vide. Utiliser le menu déroulant "Glossaire" pour choisir la thématique.

8. http ://www.larousse.com/en/dictionaries/french-english
9. http ://www.larousse.com/en/dictionaries/english-french
10. http ://www.larousse.com/en/dictionaries/french/
11. http ://dictionary.cambridge.org/
12. http ://www.termiumplus.gc.ca/

Situation	textes CANCER DU SEIN	textes SCIENCES DE L'EAU
minimale	traducteur 1	traducteur 1
cible	traducteur 2	traducteur 3
maximale	traducteur 3	traducteur 2

TABLE 2.7 – Répartition des textes et situations de traduction entre traducteurs

2.2.2.2 Déroulement de l'évaluation

L'évaluation s'est déroulée en deux phases : phase de traduction et phase d'évaluation de la qualité des traductions.

Phase de traduction

Nous avons utilisé la personne non spécialiste de la traduction pour traduire uniquement dans la situation minimale, qui est censée produire les moins bonnes traductions. Les deux autres traducteurs ont traduit alternativement dans les situations maximale et cible. Cette alternance permet d'éviter qu'un même traducteur traduise des textes d'une même thématique dans différentes situations de traductions (effet d'apprentissage évoqué dans la section 2.2.1.3).

Chaque traducteur à reçu les textes à traduire accompagnés de l'instruction suivante :

> *Traduisez chaque texte selon la situation de traduction spécifiée. Indiquez le temps que vous avez mis pour traduire chaque texte. Une fois la traduction finie, listez les termes ou expressions qui vous ont posé problème. Indiquez quelles ressources vous avez utilisées pour trouver la traduction et notez la traduction finalement retenue.*

Chaque situation est décrite précisément au traducteur comme en section 2.2.2.1. La traduction se fait de langue seconde vers la langue maternelle du traducteur, dans notre cas, de l'anglais vers le français. Une fois les textes traduits, on collecte tous les termes relevés comme problématiques et la traduction retenue par le traducteur.

Pour l'évaluation, on ne garde que les termes problématiques communs à au moins deux situations de traduction (82 % étaient communs aux trois situations), ce qui a donné un jeu de 148 termes problématiques (87 pour la thématique CANCER DU SEIN ; 61 pour la thématique SCIENCES DE L'EAU).

Phase d'évaluation de la qualité des traductions

Deux juges notent la qualité des traductions des termes. Ils sont aidés par une traduction de référence, qui correspond au terme trouvé dans la version cible du texte. Terme source et traduction de référence sont contextualisés, c'est-à-dire présentés dans leur phrase d'origine. Les juges ont aussi accès aux documents d'origine source et cible. Ils peuvent recourir, en plus de la traduction de référence, à la base de données terminologique TERMIUM [13]. Les traductions sont anonymisées et mélangées aléatoirement, de façon à ce que le juge ne puisse pas savoir

13. http ://www.termiumplus.gc.ca/

dans quelle situation ont été traduits les termes. Le tout est fourni dans un fichier tableur, où chaque groupe de traductions est présenté comme dans l'extrait 2.1.

```
  #6  mammogram
VG-3  Research has shown that women who have regular mammograms
      are more likely to survive breast cancer.
  #7  mammographie
VG-3  La recherche indique que les femmes qui passent régulièrement des
      mammographies sont plus susceptibles de survivre au cancer.
```

ID	traduction	rang	exact	acceptable
8	mammogramme	2	0	1
9	mammographie	1	1	0
10	mammographie	1	1	0

Extrait 2.1 – Exemple de traductions annotées

Les juges effectuent deux tâches d'évaluation :

Tâche de classement Les juges ordonnent les traductions de la meilleure à la moins bonne (les égalités sont autorisées).

Tâche de jugement : Les juges notent séparément la qualité de chaque traduction selon les critères définis plus haut (EXACT, ACCEPTABLE, FAUX).

Afin d'homogénéiser au maximum l'évaluation, des instructions d'annotation détaillées et quelques exemples d'annotations sur des cas difficiles ont été fournis aux juges. Compte-tenu du petit nombre de données (seulement 148 groupes de termes problématiques), nous n'avons pas procédé à une première évaluation "à blanc" qui aurait permis d'améliorer encore plus l'homogénéité de l'évaluation (Blanchon et Boitet, 2007).

2.2.3 Résultats obtenus

Les lexiques extraits ont été évalués du point de vue de leur utilisabilité (section 2.2.3.1) et du point de vue de la qualité des traductions qu'ils permettent de produire (section 2.2.3.2).

2.2.3.1 Utilisabilité des lexiques

Temps de traduction

Il est estimé qu'un traducteur professionnel peut traduire entre 2 000 et 2 500 mots par journée de huit heures en agence de traduction (250 à 313 mots/heure) (SFT, 2009). Les résultats obtenus font état de temps de traduction plutôt rapides. Dans nos expériences, la situation minimale est celle qui demande le moins de temps de traduction (503 mots/heure). Comme c'est *a priori* la situation la plus difficile, on aurait pu penser que c'est celle qui aurait demandé le plus temps (difficulté à décider d'une traduction...). Une explication serait que, dans cette situation, le traducteur a moins de ressources à parcourir. Concernant les situations avec ressources spécialisées, le temps passé est équivalent aux meilleurs temps d'une situation professionnelle : 322 mots/heure et 310 mots/heure pour les situations cible et maximale

respectivement, ce qui est aussi en partie étonnant car les étudiants ont une moyenne de temps traduction estimée aux alentours de 200 mots/heure.

Utilisation des ressources

Les traducteurs ont noté, pour chaque terme problématique, les ressources auxquelles ils avaient eu recours pour traduire le terme :

Ressources généralistes Les dictionnaires de langue générale bilingues et monolingues.

Ressources spécialisées Les lexiques issus des corpus comparables pour la situation cible, Internet pour la situation maximale.

Intuition Le recours à des heuristiques intuitives, comme la reprise de la graphie du terme modulo une adaptation à la morphologie de la langue cible, par exemple, la traduction de *sensitivity* par *sensitivité*.

Une traduction a pu être produite à l'aide de plusieurs sources, par exemple, en traduisant un terme mot à mot à l'aide de la ressource générale, puis en cherchant une attestation de la traduction candidate dans la ressource spécialisée.

Le figure 2.1 montre le nombre de fois où chaque ressource a été utilisée en fonction des situations de traduction. Dans les situations cible et maximale, les traducteurs ont très peu eu recours aux ressources généralistes : plus la ressource spécifique est large, plus elle a joué un rôle dans la production de la traduction, et moins l'intuition ou les ressources généralistes ont été utilisées. Dans la situation cible, les traducteurs ont autant eu recours à l'intuition que dans la situation minimale, ce qui semble indiquer qu'ils n'ont pas été entièrement satisfaits de cette ressource.

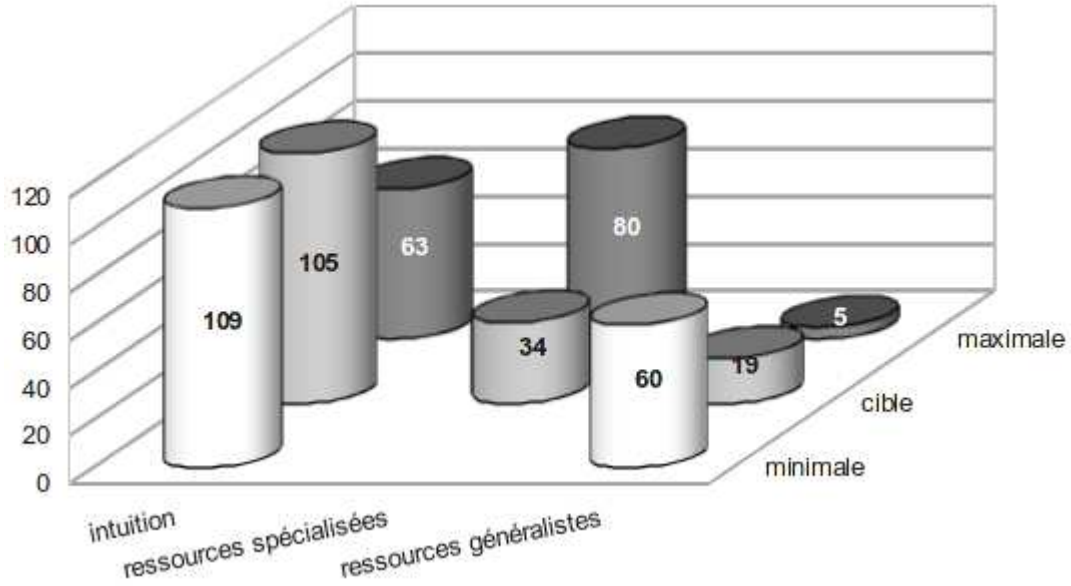


FIGURE 2.1 – Nb. de fois où chaque ressource a été utilisée en fonction des situations de traduction

Impressions de traducteurs sur les lexiques issus de corpus comparables

La traduction dans la situation cible n'a pas été aisée. Les textes spécialisés se sont révélés très durs à traduire et, bien que les retours soient bons sur l'interface, le contenu du lexique a fortement été critiqué. On peut même dire qu'il y a eu, dans un premier temps, un fort rejet. Voici le verbatim d'un courriel envoyé par l'un des traducteurs :

> « *En gros, 75 % des mots techniques ne figurent pas dans le glossaire, et sur les 25 % restants, 99 % ont entre 10 et 20 traductions candidates, mais aucune de validée. Du coup, dans le meilleur des cas on est "à peu près sûr", mais jamais totalement. Et dans le pire des cas (très fréquemment, malheureusement) on y va "à l'instinct".* »

Après discussion, il s'avère que les traducteurs s'attendaient à trouver directement la traduction d'un terme en le tapant dans le champ de recherche. Ils n'étaient pas suffisamment préparés à l'utilisation d'un lexique proposant plusieurs traductions candidates. Nous avons vu avec eux comment ils pouvaient mettre en place des stratégies pour tirer le meilleur parti du lexique bilingue, par exemple nous leur avons suggéré d'exploiter les informations présentes dans les fiches terminologiques (termes proches, concordancier, collocations) et d'utiliser la recherche plein-texte pour valider des idées de traductions. Cette réaction devant le lexique peut être en partie expliquée par le fait que les traducteurs étaient des étudiants et manquaient d'expérience.

2.2.3.2 Qualité des traductions produites

Accord inter-annotateur

La mesure d'accord inter-annotateur utilisée est le Kappa de Carletta (1996). Cette mesure prend en compte l'accord observé $P(A)$ et la probabilité d'un accord aléatoire $P(E)$:

$$Kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

Concernant la tâche de jugement, nous obtenons un accord faible (0,36) mais meilleur que Callison-Burch *et al.* (2007) (0,23 et 0,25) qui jugent des phrases entières alors que nous jugeons des groupes syntaxiques. Concernant la tâche de classement, l'accord est fort (0, 65). Ces résultats sont cohérents avec ceux de Callison-Burch *et al.* (2007) : il est plus facile d'annoter des groupes syntaxiques que des phrases entières et il est plus facile de classer des traductions que de les juger dans l'absolu. Enfin, nous notons que l'accord est meilleur pour la thématique SCIENCES DE L'EAU (0,42) que pour la thématique CANCER DU SEIN (0,25).

Tâche de jugement

Les résultats de la tâche de jugement sont présentés dans la figure 2.2. Nous avons évalué séparément les traductions de la thématique CANCER DU SEIN et les traductions de la thématique SCIENCES DE L'EAU.

Pour la thématique CANCER DU SEIN, la proportion de traductions jugées fausses est quasi-équivalente dans les trois situations. Le lexique issu du corpus comparable (situation cible) a permis d'augmenter le nombre de traductions jugées exactes par rapport à la situation

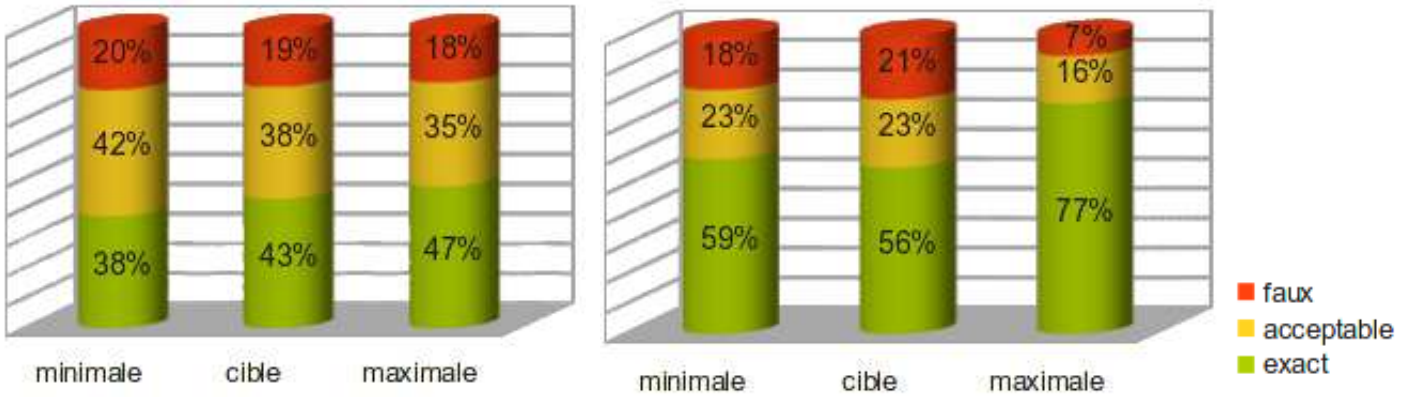
minimale : 43 % contre 38 %. La situation maximale est celle qui a permis de produire les meilleures traductions (47 % de traductions exactes).

Concernant la thématique SCIENCES DE L'EAU, les traductions produites dans la situation cible (lexique issu des corpus comparables + dictionnaires généralistes) sont plus souvent fausses que celles traduites dans la situation minimale (dictionnaires généralistes uniquement) : 25 % vs. 10 %. Ceci n'est pas normal car les deux situations partagent un socle commun de ressources généralistes. Les traductions produites dans la situation cible auraient dû être au moins aussi bonnes que celles produites dans la situation minimale. Une explication possible est que, comme montré dans la figure 2.1, les traducteurs qui ont traduit dans la situation cible se sont surtout servi des ressources généralistes et de leur intuition et ont peu utilisé le lexique spécialisé, peut-être à cause de leur première réaction de rejet.

Tâche de classement

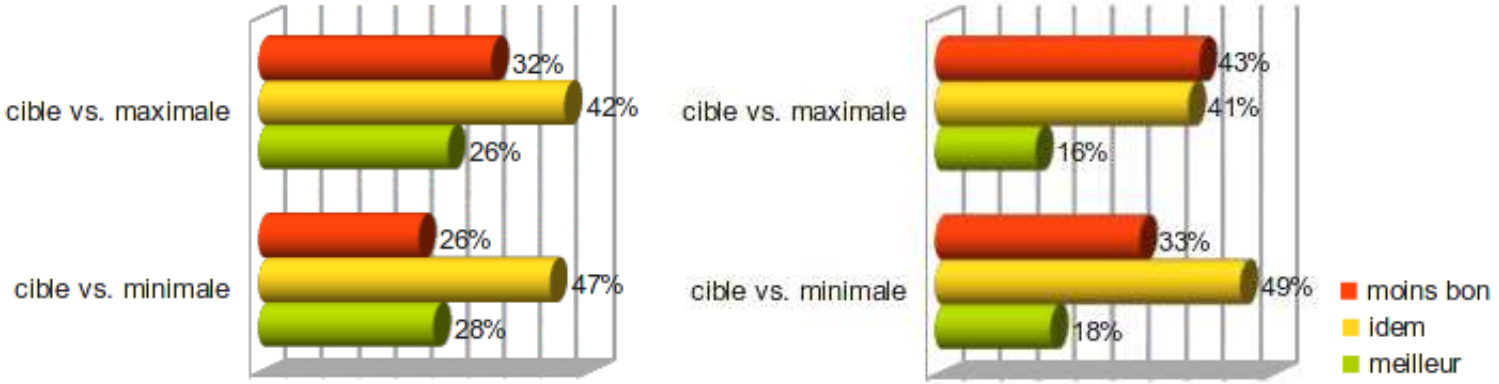
Les résultats de la tâche de classement sont présentés dans la figure 2.3. On retrouve des résultats similaires à ceux de la tâche de jugement :

- Lorsque les traductions d'un même terme sont comparées entre elles, celles produites dans la situation maximale (dictionnaires généralistes + Internet) sont toujours les meilleures, quelle que soit la thématique, ceci est plus marqué avec la thématique SCIENCES DE L'EAU.
- Les traductions faites dans la situation cible sont meilleures que celles produites dans la situation minimale uniquement pour la thématique CANCER DU SEIN et pas pour la thématique SCIENCES DE L'EAU.



(a) CANCER DU SEIN (b) SCIENCES DE L'EAU

FIGURE 2.2 – Résultats de la tâche de jugement



(a) CANCER DU SEIN (b) SCIENCES DE L'EAU

FIGURE 2.3 – Résultats de la tâche de classement

2.2.3.3 Couverture des lexiques

Nous observons que, quel que soit le mode d'évaluation, le lexique spécialisé ne semble présenter un intérêt par rapport à la situation minimale qu'avec les données CANCER DU SEIN. Nous expliquons ceci par le fait que le vocabulaire des textes SCIENCES DE L'EAU était très peu couvert par le lexique extrait à partir des corpus comparables.

Les textes à traduire ont été sélectionnés manuellement, de façon intuitive, avec pour seul critère le fait que leur sujet soit le même que celui de la thématique du corpus comparable. Or, la thématique SCIENCES DE L'EAU est beaucoup trop générale [14], ce qui a pu nous conduire à sélectionner des textes contenant un vocabulaire différent de celui du corpus comparable. Nous pouvons aussi noter que les textes du corpus CANCER DU SEIN proviennent uniquement des sources françaises alors qu'une partie des textes à traduire provient de sources canadiennes. Ceci a pu participer à réduire la couverture des lexiques. Toutefois, ce problème n'a pas été relevé par les traducteurs.

Nous avons calculé, pour chaque thématique, la proportion du vocabulaire des textes à traduire T qui se trouve effectivement dans le lexique L :

$$couverture(L,T) = \frac{|L| \cap |T|}{|T|}$$

Les résultats sont donnés dans le tableau 2.8. Dans le meilleur des cas, même si le lexique contenait 100 % de traductions correctes, le traducteur n'y trouverait des traductions utiles que pour 67 % des mots du textes à traduire pour la thématique CANCER DU SEIN et 14 % uniquement pour la thématique SCIENCES DE L'EAU. Le lexique CANCER DU SEIN, bien qu'acquis sur un corpus plus petit et légèrement moins comparable que le corpus SCIENCES DE L'EAU, couvre plus de vocabulaire que le lexique SCIENCES DE L'EAU.

	CANCER DU SEIN	SCIENCES DE L'EAU
textes à traduire (EN)	0,94	0,14
traductions de référence (FR)	0,67	0,78

TABLE 2.8 – Couverture des textes à traduire (et leurs traductions) par les lexiques extraits

2.2.3.4 Reproduction du protocole à plus grande échelle

Le travail présenté jusqu'ici s'inscrit dans un lot de travail du projet ANR METRICC auquel nous avons participé. Notre participation avait pour but de proposer un protocole d'évaluation et d'en faire une première expérimentation à petite échelle afin de rôder le processus. La

14. D'après le site de la revue *Sciences de l'eau*, il s'agit d'un champ pluridisciplinaire empruntant à sept grands domaines :
- l'hydrologie, l'hydrogéologie, la gestion des ressources en eaux ;
- la qualité physicochimique des eaux souterraines et des eaux de surface ;
- l'hydrobiologie, la microbiologie, la toxicologie et l'écotoxicologie ;
- la structure et le fonctionnement des écosystèmes aquatiques ;
- la qualité et le traitement de l'eau potable ;
- l'épuration des eaux résiduaires ;
- les aspects socioéconomiques et juridiques de la gestion de l'eau.

reproduction à grande échelle du protocole a été assurée par le Dr. Emmanuel Planas (Planas, 2011) qui enseigne les outils d'aide à la traduction aux étudiants-traducteurs de l'Université Catholique de l'Ouest. Nous avons souhaité faire état des résultats obtenus par le Dr. Planas car ils s'inscrivent dans la suite logique de notre travail et contribuent à éclairer la question de l'apport des corpus comparables à la traduction spécialisée.

Le protocole a été reproduit avec les données de la thématique CANCER DU SEIN sur un groupe de 20 étudiants traducteurs. Chaque traducteur a traduit un texte dans une seule situation de traduction. Les groupes d'étudiants affectés à chaque situation de traduction sont de niveaux équivalents. Notons qu'il s'agit d'étudiants de première année de Master alors que dans notre expérimentation, nous avions une personne ayant fait de la traduction en Licence et deux étudiant-e-s en dernière année de Master. Pour l'évaluation, seule la tâche de jugement à été effectuée. Un entraînement préalable a été mené sur les textes de la thématique SCIENCES DE L'EAU, ce qui garantit une certaine homogénéité dans l'annotation.

La figure 2.4 permet de comparer les résultats obtenus par Planas avec les nôtres. Le premier aspect frappant est que, globalement, les traducteurs de Planas sont moins compétents que les nôtres (18 % à 20 % de termes faux pour nous ; 26 % à 44 % pour Planas). Cette différence est normale puisque que les traducteurs de Planas sont des étudiants de M1 et non de M2 comme dans notre cas. Planas note d'ailleurs que les erreurs de traduction faites dans la situation maximale sont surtout dues au manque de culture scientifique et au manque d'expérience des traducteurs. Dans plusieurs cas, la traduction est erronée car le traducteur a fait une confiance aveugle à une source Internet non fiable.

On retrouve le même rapport entre les situations : la situation maximale est la meilleure, la situation minimale est la moins bonne. La différence entre situation cible et situation minimale est beaucoup plus marquée pour Planas : -7 % de traductions fausses lorsque les traducteurs utilisent le lexique issu de corpus comparables (-1 % dans notre cas).

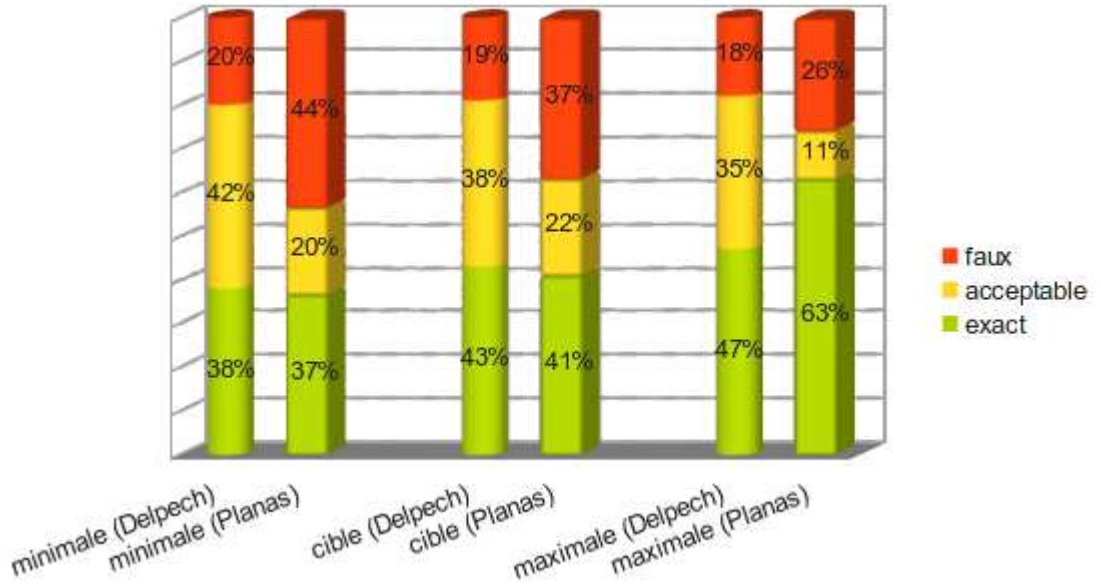


FIGURE 2.4 – Comparaison de la qualité des traductions obtenus avec les données CANCER DU SEIN : expérimentation de Planas (2011) vs. notre expérimentation

2.3 Discussion

Il est assez difficile de tirer des conclusions générales sur l'utilité des lexiques issus de corpus comparable à partir de notre expérience car d'une part, une partie de nos données a été mal construite (SCIENCES DE L'EAU) et d'autre part, nous n'avons pas démontré que les différences de qualité observées étaient significatives. L'expérience de Planas (2011), effectuée uniquement avec les données CANCER DU SEIN, fait état d'une différence plus marquée entre traductions faites uniquement avec les ressources généralistes et traductions faites avec les ressources généralistes et le lexique issu du corpus comparable. Toutefois, l'expérience a été menée avec des étudiants traducteurs et non des professionnels. Or, on sait que les étudiants travaillent différemment des professionnels (Carl *et al.*, 2011). Au mieux, nous pouvons dire que les lexiques issus de corpus comparables sont une aide pour les traducteurs inexpérimentés. Une perspective de recherche sera donc de reproduire l'expérience avec des données mieux construites et des traducteurs professionnels.

Au-delà de la mesure de l'utilité des corpus comparables, ces expériences ont révélé plusieurs obstacles à l'exploitation des corpus comparables pour la traduction spécialisée.

Tout d'abord, nous avons observé l'importance de travailler avec des thématiques de granularité fine et la nécessité de collecter un corpus en forte adéquation avec les textes à traduire.

Ensuite, nous nous sommes rendu compte que la forme des lexiques extraits est gênante pour les traducteurs, voire provoque un rejet. La présentation d'une liste de 20 traductions candidates est perturbante (≪ *99 % ont entre 10 et 20 traductions candidates, mais aucune n'est validée* ≫). Après discussion avec les traducteurs, il semble qu'une liste de 3 ou 4 traductions serait le maximum acceptable.

Enfin, la simple énumération des traductions candidates, bien qu'associée à un concordancier, n'est pas suffisante : les traducteurs manquent d'indices leur permettant de valider les traductions (≪ *on est "à peu près sûr", mais jamais totalement* [...] on y va "à l'instinct" ≫).

Forts de ces constats, nous dégageons cinq axes de recherche pour améliorer l'exploitation des corpus comparables :

Formation et recueil des besoins Les lexiques issus de corpus comparables représentent un nouveau type de ressource, moins précis que les ressources auxquelles sont accoutumés les traducteurs. Bien que reconnus comme utiles dans les études de traductologie, il semble que ce type de ressource ne soit pas entré dans les habitudes des traducteurs ni même ne fasse partie de leur formation. Il serait donc bon d'établir un dialogue entre développeurs d'outils de TAO, ergonomes de la traduction et formateurs afin que les outils d'analyse des corpus comparables correspondent effectivement aux besoins des traducteurs et que ces outils soient enseignés dans les formations.

Collecte du corpus Il est nécessaire d'améliorer la comparabilité du corpus, que ce soit à l'aide de métriques basées sur le vocabulaire partagé entre les deux corpus (Li *et al.*, 2011; Li et Gaussier, 2010) ou par une prise en compte du type discours lors de la constitution du corpus (Morin *et al.*, 2007; Goeuriot, 2009). L'adéquation entre les textes à traduire et les textes du corpus doit également être optimale. Des recherches récentes ont donné lieu à des *crawlers* thématiques (Talvensaari *et al.*, 2008; de Groc, 2011) et à des systèmes d'acquisition de corpus comparables à partir du Web (Kilgarriff *et al.*, 2011).

Contextualisation des termes Il est essentiel de donner aux traducteurs les moyens de s'assurer de leurs choix de traduction. De simples concordanciers monolingues et

décorrélés l'un de l'autre ne peuvent suffire. Traductions candidates et terme source doivent être présentés dans des contextes pertinents pour la traduction. Cet axe de recherche est actuellement exploré dans le cadre du projet ANR CRISTAL [15] qui s'intéresse aux contextes riches en connaissances pour la traduction (Meyer, 2001).

Outillage Les outils des traducteurs doivent être adaptés afin de leur permettre d'exploiter correctement les lexiques issus de corpus comparables. Très peu d'outils de TAO ont été conçus pour traiter ce type de ressource (Sharoff *et al.*, 2006; Brown de Colstoun *et al.*, 2011) et les fonctionnalités de recherche en corpus sont encore limitées. De plus, très peu de format d'échanges de lexiques bilingues (TBX, XLIFF...) permettent d'échanger des lexiques issus de corpus comparables (Delpech et Daille, 2010).

Équivalences traductionnelles Il s'agit là d'améliorer directement la qualité des lexiques extraits. Nous avons vu que les traducteurs n'apprécient pas d'avoir une longue liste de traductions candidates, surtout si dans bon nombre de cas, aucune ne s'avère utile. Afin d'améliorer l'utilisabilité des lexiques par les traducteurs techniques, il conviendra de favoriser la précision des lexiques, quitte à produire moins de traductions candidates. Dans tous les cas, on s'attachera à faire apparaître la traduction correcte parmi les premières candidates.

Nous avons choisi de poursuivre nos recherches dans le cadre de la recherche d'équivalences traductionnelles et notamment, nous nous sommes donné pour objectif d'essayer d'obtenir des lexiques plus précis, même si ces derniers sont de taille plus modeste. La méthode distributionnelle est une méthode employée depuis les années 90 et, malgré les nombreuses recherches visant à l'améliorer, nous avons vu dans le chapitre 1 que les résultats se situent entre 13 % et 65 % de précision sur le Top1 (tableau 1.1). Il faut prendre en compte au moins 20 traductions candidates pour espérer obtenir entre 40 % et 94 % de précision selon les domaines et les couples de langues. De plus, l'approche distributionnelle nécessite de grands volumes de textes pour obtenir les meilleurs scores. Or, le recours aux corpus comparables a lieu lorsque, justement, les données se font rares.

Ces raisons font que nous nous sommes orientés vers d'autres méthodes d'acquisition de lexiques qui ne sont pas spécifiques aux corpus comparables. Il s'agit de méthodes visant non pas à extraire préalablement des termes puis à les aligner mais plutôt à *générer* automatiquement les traductions d'un terme source. Ces méthodes ont produit récemment de bons résultats sur les corpus comparables (Morin et Daille, 2010; Weller *et al.*, 2011; Harastani *et al.*, 2012). Nous en faisons un état-de-l'art dans le chapitre suivant avant de proposer une amélioration de ces méthodes dans le chapitre 4.

15. Contextes RIches en connaissanceS pour la TrAduction terminoLogique, www.projet-cristal.org

Chapitre 3

Génération automatique de traductions de termes

Sommaire

Introduction

Dans ce chapitre, nous présentons des méthodes de *génération* de traduction. Contrairement à des approches cherchant à *aligner* des termes prédéfinis sur la base d'une similarité, la génération consiste à produire une traduction à partir de connaissances sur les équivalences traductionnelles d'une langue à une autre et sur les réalisations possibles en langue cible.

Une approche possible est de se baser sur la sémantique compositionnelle : l'unité à traduire est décomposée en sous-unités porteuses de sens puis ces sous-unitées sont traduites et recomposées de façon à former un terme en langue cible (section 3.1). Une seconde famille d'approches consiste à s'appuyer sur des connaissances empiriques (section 3.2). Ces approches envisagent l'unité à traduire plus comme une chaîne de caractères à réécrire que comme une unité linguistique. La traduction est effectuée grâce à des connaissances apprises automatiquement à partir d'exemples de paires de traductions.

3.1 Approches compositionnelles

Dans cette section, nous décrivons des approches expertes basées sur le principe de compositionnalité. Dans ce paradigme, les recherches se préoccupent avant tout de gérer des phénomènes de variation morphologique, lexicale, morphosyntaxique et de fertilité. Ces approches sont utilisées pour traduire des unités polylexicales, qui sont découpées en mots (3.1.2), ainsi que des unités monolexicales morphologiquement complexes qui sont découpées en morphèmes (3.1.3). Dans ce dernier cas, la difficulté est de pouvoir traiter les différentes constructions morphologiques, en plus des cas de variation et de fertilité. Une fois les traductions générées, il convient de s'assurer qu'elles sont possibles en langue cible. Pour cela, diverses méthodes de sélection et de filtrage ont été mises au point (3.1.3).

3.1.1 Principe de la traduction compositionnelle

Le principe de compositionnalité peut être énoncé ainsi (Keenan et Faltz, 1985, pp. 24-25) :

≪ *Le sens du tout est fonction du sens de ses constituants* ≫ [1]

Ce principe sous-tend qu'il est possible de comprendre le sens d'une expression inconnue, pour peu que le sens de ses composants soit connu et qu'il soit possible d'en d'interpréter sa structure. En se basant sur le principe de compositionnalité pour générer des traductions, on suppose qu'il est possible de traduire une unité lexicale inconnue, pour peu que l'on sache interpréter sa structure, traduire chacun de ses composants et les recombiner en langue cible.

Le processus de traduction compositionnelle peut-être formalisé ainsi :

$$\begin{aligned}
\mathcal{CT}(\text{“ab”}) &= \mathcal{S}(\mathcal{R}(\mathcal{T}(\mathcal{D}(\text{“ab”})))) \\
&= \mathcal{S}(\mathcal{R}(\mathcal{T}(\{\mathsf{a}, \mathsf{b}\}))) \\
&= \mathcal{S}(\mathcal{R}(\{\mathcal{T}(\mathsf{a}) \times \mathcal{T}(\mathsf{b})\})) \\
&= \mathcal{S}(\mathcal{R}(\{\mathtt{A}, \mathtt{B}\})) \\
&= \mathcal{S}(\{\mathtt{A}, \mathtt{B}\}, \{\mathtt{B}, \mathtt{A}\}) \\
&= \text{“}\mathtt{BA}\text{”}
\end{aligned}$$

où “ab” est un terme source composé de *a* et *b*, “BA” est un terme cible composé de B et A et il existe une ressource bilingue liant *a* à A et *b* à B.

En pratique, la traduction compositionnelle ($\mathcal{CT}$) revient donc à :

1. Décomposer le terme source en composants “atomiques” ($\mathcal{D}$)
2. Traduire ces composants en langue cible ($\mathcal{T}$)
3. Recomposer les constituants traduits de façon à former des traductions candidates ($\mathcal{R}$)
4. Filtrer les traductions candidates à l'aide d'une fonction de sélection ($\mathcal{S}$) de façon à ne retenir que les traductions correctes

La première implémentation apparaît à la fin des années 90 avec la publication de Grefenstette (1999). Le but de cette publication est avant tout de démontrer l'intérêt du Web comme ressource pour le Traitement Automatique des Langues. Toutefois, c'est aussi une démonstration de la pertinence de l'approche compositionnelle pour générer la traduction d'expressions complexes. Dans son expérience, G. Grefenstette utilise un lexique de référence

1. ≪ *The meaning of the whole is a function of the meaning of the parts* ≫

construit à partir d'une ressource bilingue. Il sélectionne pour l'expérimentation 724 composés nominaux allemands et 1 140 composés nominaux espagnols destinés à être traduits en anglais. Pour chaque composé nominal à traduire l'auteur s'assure que :

- il est possible de traduire le composé à partir des traductions de ses composants (mots)
- la ressource bilingue contient les traductions des composants
- il est possible de construire plus d'une traduction candidate
- la traduction correcte du composé nominal est connue

En traduisant les composés allemands et espagnols de façon compositionnelle, G. Grefenstette obtient 3556 traductions candidates pour l'allemand et 6186 pour l'espagnol. Les traductions candidates sont requêtées sur le moteur AltaVista, qui donne pour chaque requête, son nombre d'occurrences dans les pages indexées par le moteur. La traduction candidate retenue est celle ayant le plus grand nombre d'occurrences. En suivant cette méthodologie, G. Grefenstette obtient respectivement 87 % et 86 % de traductions correctes pour l'allemand et l'espagnol.

Cette version de la traduction compositionnelle est élémentaire : on se contente simplement d'effectuer une traduction mot à mot, un peu à la manière des premiers traducteurs automatiques. Cette approche par traduction directe suppose un fort parallélisme entre langue source et langue cible et ne prend pas en compte divers phénomènes qui peuvent subvenir lors de la traduction.

Ces phénomènes sont bien connus et sont fréquemment listés dans la littérature :

Variation morpho-syntaxique Termes source et cible ont des structures morphosyntaxiques différentes, il y notamment un changement au niveau des catégories grammaticales. Par exemple, un nom peut être traduit par un adjectif :
- *thérapie génique* (NOM ADJECTIF) → *gene therapy* (NOM NOM)
- *anti-cancer* (NOM) → *anti-cancéreux* (ADJECTIF)

Variation lexicale Les langues source et cible utilisent des mots sémantiquement proches mais qui ne sont pas des traductions exactes l'un de l'autre :
- *traduction automatique* → *machine translation*
- *mixed departmentalization* → *structuration mixte*

Variation terminologique Un terme source peut être traduit par un terme cible ayant plusieurs variantes, plusieurs traductions sont donc possibles :
- *mixed departmentalization* → *départementalisation mixte, structuration mixte*
- *oophorectomy* → *ablation des ovaires, ovariectomie*

Fertilité Les termes source et cible sont de longueur différente[2] :
- *isothermal snowpack* → *manteau neigeux isotherme*
- *oophorectomy* → *ablation des ovaires*

Dans la suite, nous montrons comment ces phénomènes de variation ont été traités dans les travaux exploitant le principe de compositionnalité.

3.1.2 Traduction compositionnelle d'unités polylexicales

La traduction compositionnelle d'unités polylexicales passe par un découpage en mots. Parmi les approches présentées, deux s'inscrivent dans la lignée de Grefenstette (1999) et proposent des solutions pour gérer la variation lexicale et/ou morphologique (Robitaille *et al.*,

2. Telle que définit par Robitaille *et al.* (2006), nous redéfinissons ce concept p. 104.

2006; Morin et Daille, 2010). L'apport de Léon (2008) concerne l'exploitation de documents mixtes pour la recherche de traductions candidates. L'approche de Š. Vintar (2010) court-circuite la phase de recombinaison des composants en passant par un "sac d'équivalents".

3.1.2.1 Variation lexicale et décomposition multiple

Robitaille *et al.* (2006) améliorent la traduction compositionnelle de Grefenstette (1999) sur plusieurs points :

Décomposition multiple Pour un terme composé de n mots lexicaux, Robitaille *et al.* (2006) produisent 2^{n-1} décompositions possibles, y compris le terme lui-même. Par exemple, le terme $système\ à\ base\ de\ connaissances$ peut être décomposé de quatre façons : $système\ /\ base\ /\ connaissance$, $système\ à\ base\ /\ connaissance$, $système\ /\ base\ de\ connaissance$ et $système\ à\ base\ de\ connaissance$. Cette décomposition multiple permet, lors de la phase de traduction, de trouver directement des traductions pour des sous-parties du terme à traduire.

Gestion de la variation lexicale La traduction d'un mot peut être : (i) sa traduction donnée par le lexique bilingue ; (ii) son synonyme ou un mot sémantiquement proche donné par un thésaurus.

Constitution d'un corpus thématique "à la volée" Ceci permet de filtrer les traductions candidates sur un plus grand nombre de termes cibles.

La méthode proposée part d'un jeu de paires de traductions (*seeds* ou *graines*) qui sont utilisées pour requêter un moteur de recherche et construire un petit corpus comparable duquel sont extraits des termes en langue source et en langue cible appartenant à la même thématique. Les termes sources sont traduits en langue cible via la méthode compositionnelle améliorée : génération des traductions partielles et prise en compte des synonymes des traductions. Les traductions candidates sont filtrées sur la liste de termes cibles. L'utilisation de mots sémantiquement proches des traductions a pour effet d'augmenter le rappel et de baisser la précision : la précision passe de 92 % à 46 % et le rappel passe de 53 % à 65 %. Notant la forte précision obtenue sur les alignements effectués sans recours aux synonymes (92 %), les auteurs proposent d'utiliser ces paires de traductions comme nouvelles graines pour amorcer la collecte d'un nouveau corpus. De cette manière, Robitaille *et al.* (2006) augmentent itérativement le nombre de termes traduits tout en limitant la baisse de la précision.

Cette méthode permet de gérer les problèmes de variation lexicale et dans une certaine mesure, celui de la variation terminologique. Cependant, si un des mots du terme n'est pas traduit par le lexique bilingue, le terme source ne sera pas traduit dans son entier. De plus, cette méthode ne permet pas de gérer les cas où terme source et terme cible n'ont pas la même structure morpho-syntaxique. Pour remédier à cela, Morin et Daille (2010) suggèrent d'exploiter les parentés morphologiques.

3.1.2.2 Parentés morphologiques

Lors de la traduction compositionnelle d'une unité polylexicale, il est possible que l'un de ses composants n'apparaisse pas dans le lexique bilingue. Dans ce cas, soit la traduction échoue, soit on ne peut générer qu'une traduction partielle. La solution de Morin et Daille (2010) est d'utiliser la traduction d'un mot appartenant à la même famille morphologique que le composant manquant. Par exemple, si au moment de traduire $bilan\ énergétique$, la traduction

de l'adjectif *énergétique* n'est pas trouvée dans le lexique, c'est la traduction du nom dont il est dérivé (*energy*) qui est utilisée pour générer les candidats *energy balance* et *balance energy*.

Morin et Daille (2010) testent cette approche sur les adjectifs relationnels [3]. Leur expérience consiste donc à traduire des termes comportant un adjectif relationnel, et dans le cas où la traduction de l'adjectif n'est pas présente dans le lexique bilingue, celle-ci est remplacée par la traduction du nom dont il est dérivé. L'expérimentation porte sur 1 578 termes français de structure NOM ADJECTIF à traduire en japonais. Sur les 1 578 termes, 829 contiennent un adjectif relationnel (notés NOM ADJECTIFREL). Les traductions générées sont filtrées sur une liste de termes japonais. Termes français et termes japonais ont été extraits d'un corpus comparable.

Avec la méthode compositionnelle simple, les auteurs obtiennent 69 % de précision sur les termes NOM ADJECTIF et 63 % sur les termes NOM ADJECTIFREL mais très peu de termes NOM ADJECTIFREL sont traduits : 8 sur 829. Avec la méthode basée sur la parenté morphologique, les auteurs traduisent 128 termes NOM ADJECTIFREL avec une précision de 88 %. Ils comparent l'approche compositionnelle basée sur la morphologie avec l'approche distributionnelle. Cette dernière donne une précision de 15 % sur le Top10 et 20 % sur le Top20, ce qui est loin des résultats obtenus avec l'approche compositionnelle. Ceci s'explique par la faible fréquence des termes (80 % ont un nombre d'occurrences inférieur à 20) et justifie pleinement le recours à l'approche compositionnelle pour traduire des termes complexes.

3.1.2.3 Extraire les traductions à partir de documents mixtes

Léon (2008) a pour objectif de traduire des unités polylexicales de structure NOM ADJECTIF ou NOM *de* NOM du français vers l'anglais. La traduction de chacun des éléments est obtenue via un lexique bilingue, la traduction candidate est recomposée en suivant des patrons de traduction :

FR : NOM_1 ADJECTIF_1 → EN : ADJECTIF_1 NOM_1
FR : NOM_1 *de* NOM_2 → EN : NOM_2 NOM_1
FR : NOM_1 *de* NOM_2 → EN : NOM_1 *of* NOM_2

En appliquant la méthode compositionnelle simple, Léon (2008) peut générer des traductions pour 72 % des unités à traduire. Les unités non traduites à ce stade sont soit des unités dont un des éléments n'apparaît pas dans le dictionnaire, soit des unités au sens non-compositionnel (ex. : *caisse claire* → *snare drum 'tambour piège'*) ou présentant une divergence lexicale. La solution déployée par Léon (2008) consiste à obtenir des extraits de textes en langue cible fortement susceptibles de contenir la traduction de l'unité à traduire. Ces extraits de textes sont issus de documents bilingues, obtenus en demandant à un moteur de recherche de rechercher l'unité source dans des documents en langue cible. La page de résultats fournie par le moteur de recherche contient donc des *snippets* [4] mixtes (dans les deux langues).

Léon commence par rechercher dans les snippets des cognats [5] du mot qui n'a pu être traduit via le dictionnaire bilingue. Les traductions des lexies/termes sont générées en utilisant le(s) cognat(s) comme traduction. De cette façon, l'auteure obtient 8 % de traductions

3. Les adjectifs relationnels sont des adjectifs dérivés de noms qui illustrent une relation entre le nom qu'ils modifient et le nom dont ils sont dérivés. Ils constituent l'équivalent syntaxique d'un complément de nom ou d'une relative qui expliciterait cette relation : *la race chevaline* → *la race des chevaux, une boucherie chevaline* → *une boucherie où l'on vend de la viande de cheval*. Ils s'opposent aux adjectifs qualificatifs qui indiquent une caractéristique du nom qu'ils modifient, par ex. : *un livre bleu* (Riegel *et al.*, 2005).

4. Fragments de textes affichés sous le titre de chaque document web dans la page de résultat retournée par le moteur.

5. Pour Léon (2008), deux cognats sont deux mots ayant les mêmes quatre premières lettres.

supplémentaires. Dans le cas où aucun cognat satisfaisant n'a été trouvé, Léon recherche les bigrammes qui sont les plus fréquents dans les *snippets*. Ces bigrammes sont considérés comme des traductions candidates et permettent de traduire 2 % de traductions en plus. Afin de sélectionner la meilleure traduction, divers filtres sont appliqués aux traductions candidates (les filtres sont détaillés dans la section 3.1.4 sur le filtrage des traductions).

3.1.2.4 Sac d'équivalents

La majorité des techniques de traduction compositionnelle utilise des patrons syntaxiques pour recomposer les lexies cibles. Ceci a pour effet d'augmenter la tâche de l'expert qui doit décrire toutes les équivalences structurelles entre langues source et cible. Une solution plus économique consiste à générer toutes les permutations possibles des mots traduits puis à les filtrer grâce à un corpus en langue cible comme le font Robitaille *et al.* (2006). Dans le même esprit, Š. Vintar (2010) propose de passer un « *sac d'équivalents* »[6]. Š. Vintar commence par acquérir sur un corpus parallèle des probabilités de traductions de mots. Pour traduire un terme, celui-ci est découpé en mots et pour chacun de ces mots, sa traduction ainsi que la probabilité qui lui est associée sont ajoutées à une liste d'équivalents (le fameux "sac").

Par exemple, si le terme à traduire est $destruction\ of\ anti\text{-}personnel\ mines$, son sac d'équivalents en slovène est :

composants du terme source	sac d'équivalents	
`destruction`	`uničevanje`	`0,86`
	`uničenje`	`0,14`
`antipersonnel`	`protipehoten`	`1,00`
`mine`	`mina`	`1,00`

Ce sac d'équivalents est ensuite comparé aux termes cibles. Le score de traduction d'un terme cible est la moyenne des probabilités de traduction des mots du sac d'équivalents qui se trouvent aussi dans le terme cible. Si on reprend l'exemple de la traduction du terme anglais $destruction\ of\ anti\text{-}personnel\ mines$ en slovène, la traduction candidate $uničevannje\ protipehotnih\ min$ a un score de traduction de 0,95 et la traduction candidate $uničenje\ protipehotnih\ min$ a un score de traduction de 0,71. Cette technique permet d'obtenir une précision variant entre 64 % et 97 % selon les corpus (en sélectionnant les 300 meilleurs alignements) mais ces résultats sont à relativiser car termes sources et termes cibles ont été extraits d'un même corpus parallèle.

Le désavantage de l'approche est qu'elle nécessite un lexique probabiliste préalablement acquis sur un corpus parallèle. Elle ne serait donc applicable que dans le cas où on aurait un corpus parallèle suffisamment grand pour obtenir un lexique probabiliste mais encore trop petit pour obtenir un nombre satisfaisant de paires de termes. Le corpus comparable serait utilisé pour augmenter le nombre de paires de traductions.

Comme pour les travaux de Robitaille *et al.* (2006), l'approche de Š. Vintar permet de gérer les divergences lexicales et syntaxiques. Elle permet aussi de gérer la variation terminologique puisqu'un même mot source peut avoir plusieurs traductions. La technique du sac d'équivalents fait abstraction de l'ordre des mots et d'éventuels mots grammaticaux : les termes $weapon\ confiscation$ et $confiscation\ of\ weapon$ s'apparient tous deux avec le sac d'équivalents {`weapon`, `confiscation`}. Enfin, le recours au sac de mots allège le processus de traduction car il

6. « *bag-of-equivalents* » (*op. cit.*, p. 152)

évite d'avoir à générer toutes les permutations des traductions de chacun des mots du terme complexe [7].

3.1.2.5 Hybridation avec la méthode distributionnelle

La méthode compositionnelle de base utilisée par Morin et Daille (2012) consiste à identifier les mots lexicaux du terme complexe, traduire chacun de ces mots lexicaux en langue cible (un même mot peut avoir plusieurs traductions possibles), générer toutes les combinaisons des mots traduits et sélectionner les combinaisons qui se trouvent parmi les termes complexes préalablement extraits du corpus cible. Ces traductions sont ensuite ordonnées par fréquence. En cet état, si un des composants ne peut être traduit, la traduction du tout échoue.

Morin et Daille proposent de s'appuyer sur la méthode distributionnelle pour pouvoir traiter les cas où la traduction échoue parce qu'un des éléments n'a pu être traduit. Pour un terme complexe composé de deux composants $c1$ et $c2$ (ex : $antécédent\ familial$) :

- Si le composant c_1 ($antécédent$) ne peut pas être traduit :
  - on collecte son vecteur de contexte dans le corpus source :
    $\overrightarrow{\text{antécédent}} = \{(famille, 332), (familial, 73), (cancer, 68)...\}$
  - son vecteur est traduit en langue cible à l'aide d'un dictionnaire bilingue :
    $\overrightarrow{t(\text{antécédent})} = \{(family, 332), (familial, 73), (cancer, 68)...\}$
- Si le composant c_2 ($familial$) peut être traduit par le dictonnaire bilingue, on collecte les vecteurs de contextes de chacune de ces traductions. Par exemple, $familial$ peut se traduire en anglais par $familial$ et $family$, on collecte donc deux vecteurs de contextes cibles :
  - $\overrightarrow{\text{familial}} = \{(risk, 37), (cancer, 68)...\}$
  - $\overrightarrow{\text{family}} = \{(history, 372), (mutation, 50), (cancer, 24)...\}$

À l'issue de cette étape, le terme source $antécédent\ familial$ peut être représenté de deux façons en langue cible :

1. $\overrightarrow{t(\text{antécédent})}$ + $\overrightarrow{\text{familial}}$
2. $\overrightarrow{t(\text{antécédent})}$ + $\overrightarrow{\text{family}}$

Ces deux représentations sont comparées à celles des termes cibles préalablement extraits du corpus. Par exemple, le terme anglais $family\ history$ est associé à deux vecteurs : $\overrightarrow{\text{family}}$ + $\overrightarrow{\text{history}}$ où :

- $\overrightarrow{\text{family}} = \{(history, 372), (mutation, 50), (cancer, 24)...\}$
- $\overrightarrow{\text{history}} = \{(family, 37), (cancer, 68)...\}$

Pour obtenir le score de traduction entre $antécédent\ familial$ et $family\ history$, on calcule, pour chaque appariemment possible entre composants du terme source et composants du terme cible, la moyenne géométrique des similarités des vecteurs des composants appariés :

1. $\sqrt{sim(\overrightarrow{t(\text{antécédent})}, \overrightarrow{\text{family}}) \times sim(\overrightarrow{\text{familial}}, \overrightarrow{\text{history}})}$
2. $\sqrt{sim(\overrightarrow{t(\text{antécédent})}, \overrightarrow{\text{history}}) \times sim(\overrightarrow{\text{familial}}, \overrightarrow{\text{family}})}$

7. Le nombre de traductions générées est de $\prod_{i=1}^{n} l_i p!$ où l_i est le nombre de traductions possibles pour chacun des mots lexicaux composant le terme complexe et p le nombre de mots lexicaux composant le terme complexe (Morin et Daille, 2010).

3. $\sqrt{sim(\overrightarrow{t(\text{antécédent})}, \overrightarrow{\text{family}}) \times sim(\overrightarrow{\text{family}}, \overrightarrow{\text{history}})}$
4. $\sqrt{sim(\overrightarrow{t(\text{antécédent})}, \overrightarrow{\text{history}}) \times sim(\overrightarrow{\text{family}}, \overrightarrow{\text{family}})}$

On obtient ainsi 4 scores différents pour l'alignement $antécédent\ familial \leftrightarrow family\ history$.

Ces scores sont calculés pour chaque paire `(terme source, terme cible)` du corpus. Les traductions candidates sont finalement ordonnées par score décroissant.

Les résultats obtenus montrent que la méthode hybride permet d'augmenter le nombre de termes sources traduits, bien que cela passe par une baisse de la précision. Pour la traduction du français vers l'anglais, la méthode compositionnelle de base ne peut traduire que 140 termes sur 836 (16,7 %) et la précision est de 73,2 % sur le Top1 et 79,1 % sur le Top5. Avec la méthode hybride, le système génère des traductions pour 514 termes sources (61,1 %) et la précision est de 42,1 % sur le Top1 et de 55,4 % sur le Top5.

3.1.3 Traduction compositionnelle d'unités monolexicales

La traduction compositionnelle d'unités monolexicales ne permet de traduire que des unités morphologiquement complexes. Par conséquent, ces approches produisent peu de traductions mais celles-ci sont généralement de très bonne qualité. Trois processus de constructions morphologiques ont été abordés : la préfixation, la composition savante (à base de racines d'origine grecques ou latines) et la composition populaire (à partir de bases lexicales "autochtones").

3.1.3.1 Traduction de mots construits par préfixation

Les publications de Cartoni (2005, 2009a) décrivent un système de traduction automatique des néologismes construits par préfixation. Le système s'appuie sur des « *Règles de Construction des Lexèmes (RCL)* » (*op. cit.*, p. p. 5). Une RCL décrit un processus de dérivation ainsi que le changement de sens qui en découle. Dans la partie ITALIEN du tableau 3.1, on voit l'illustration d'une RCL décrivant la dérivation d'un verbe X de sens *'X'* en un verbe riX de sens *'réitérativité (X')'*. Concrètement, cette RCL peut décrire la dérivation de *costruire 'construire'* en *ricostruire 'réitérer l'action de construire'*. Une RCL bilingue établit une équivalence entre deux RCL de langues différentes comme illustré dans le tableau 3.1.

	ITALIEN		FRANÇAIS	
	INPUT	OUTPUT	INPUT	OUTPUT
G	X	riX	Y	reY
SX	cat :v	cat :v	cat :v	cat :v
S	X'(...)	réitérativité (X'(...))	Y'(...)	réitérativité (Y'(...))

TABLE 3.1 – Exemple de RCL bilingue - adapté de Cartoni (2009a, p.5)

Dans Cartoni (2005), le processus de traduction est appliqué à des néologismes de type verbes préfixés et noms déverbaux préfixés et se déroule en trois étapes :

1. Recherche de mots préfixés dans un corpus qui s'apparient au versant italien d'une RCL bilingue et dont la base est présente dans le dictionnaire bilingue.
2. Traduction de la base via le dictionnaire bilingue.

3. Construction de la traduction via le versant français de la RCL.

La première étape produit du bruit : 15 % des verbes et 2 % des noms sont de faux mots construits[8]. Les données issues de la première étape sont nettoyées manuellement et seuls les mots vraiment construits sont traduits. Les traductions sont considérées correctes à 97 % pour les verbes et 91 % pour les noms. Les autres traductions sont jugées incertaines, aucune traduction n'est jugée fausse.

Cartoni (2009a) adapte la méthode des RCL à la traduction des adjectifs relationnels préfixés. Malgré les similarités morphosyntaxiques entre l'italien et le français, la production d'adjectifs relationnels en italien est plus libre et plus productive qu'en français. Il existe donc de nombreux adjectifs relationnels italiens qui ne peuvent pas être traduits par un adjectif relationnel en français, par ex. : *gattesco 'relatif au chat'*, *creditizione 'relatif au crédit'*. Afin de gérer cela, Cartoni propose que les adjectifs relationnels partagent leur RCL avec le nom dont ils sont dérivés. Ainsi, un adjectif italien peut être traduit par un nom français, pour peu que la correspondance entre adjectif et nom soit établie, tout comme le font Morin et Daille (2010). En plus d'un lexique bilingue, cette méthode de traduction s'appuie aussi sur deux ressources monolingues (français et italien) permettant de faire le lien entre adjectifs relationnels et noms dont ils sont dérivés.

La traduction débute par la déconstruction de l'adjectif préfixé en `préfixe + adjectif`, puis le système recherche la traduction de l'adjectif dans le dictionnaire bilingue et tente de reconstruire l'adjectif en langue cible :

1. *anticostituzionale* est découpé en *anti-* et *costituzionale*
2. *costituzionale* est traduit par *constitutionnel*
3. L'adjectif *anticonstitutionel* est reconstruit à partir de *anti-* et *constitutionnel*

Si l'adjectif italien ne peut être directement traduit en français, le système suit la procédure suivante :

1. Retrouver le nom italien dont l'adjectif est dérivé : *costituzionale* → *costituzion*
2. Traduire ce nom en français : *costituzion* → *constitution*
3. Préfixer le nom : *constitution* → *anticonstitution*

À l'issue de cette procédure *anticonstituzionale* est traduit par *anticonstitution*. Éventuellement, s'il est possible de retrouver, à partir du nom français, l'adjectif relationnel correspondant, on peut traduire *anticonstituzionale* par *anticonstitutionel* :

1. Remplacer le nom par l'adjectif : *constitution* → *constitutionnel*
2. Préfixer l'adjectif : *constitutionnel* → *anticonstitutionel*

Sur 1 783 adjectifs relationnels préfixés, la méthode a permis d'en traduire 88 % par un adjectif relationnel français et 12 % par un nom. L'auteur n'indique aucune évaluation de la qualité linguistique des néologismes générés.

3.1.3.2 Traduction de mot construits par composition savante

Un composé savant est un mot formé par la combinaison de racines d'origine grecque ou latine. Ces racines sont des morphèmes liés, c'est-à-dire qu'ils ne peuvent pas être employés

8. Par exemple, *débuter* ne résulte pas de la préfixation du verbe *buter* et n'a pas le sens de *'action inverse de buter'*.

de manière autonome. Par exemple, le terme *vagotomie* est formé de la racine *vago*, du latin *vagus*, et de *tomie*, du grec *tomê*. Aucune des racines ne peut fonctionner de façon autonome : *vago* et *tomie* ne sont pas des mots du français.

Harastani *et al.* (2012) distinguent deux formes d'éléments néoclassiques :

ICF Les formes initiales comme *cardio-, patho-* qui apparaissent en début de composé.

FCF Les formes finales comme *-logy, -cide* qui apparaissent en fin de composé.

Les composés néoclassiques traduits sont de la forme : `ICF+(FCF|mot)`, c'est-à-dire une ou plusieurs ICF suivies d'une FCF ou d'un mot, par exemple : *histo/patho/logy*, *cardio/vasculaire*, *photo/bio/reactor*. Le processus de génération du lexique bilingue se déroule ainsi :

1. Extraction des composés néoclassiques sources et cibles du corpus comparable, c'est-à-dire tout adjectif ou nom qui contient une ICF ou une FCF.
   – ex. : *neurology* est extrait du corpus source et *neurologie* est extrait du corpus cible
2. Identification de la structure des composés sources, qui doivent correspondre à la forme `ICF+(FCF|mot)` [9].
   – *neurology* est découpé en *neuro+logy*
3. Traduction du composé : les ICF et FCF sont traduits via une liste d'éléments néoclassiques alignés et les mots identifiés sont traduits via un dictionnaire bilingue généraliste. Pour chaque élément, lorsque plusieurs traductions sont possibles, plusieurs traductions candidates sont générées.
   – *neuro* est traduit par *neuro* et *névro*
   – *logy* est traduit par *logie*
4. L'étape de recomposition est simple : l'ordre des éléments cibles est le même que celui des éléments sources, ainsi, *histopathology* ne peut être traduit en français que par *histopathology* et non *pathohistology*.
   – il y a deux traductions candidates : *neurologie* et *névrologie*
5. Au final, la traduction candidate est retenue si elle apparaît dans la liste de composés néoclassiques cibles extraits du corpus.
   – la traduction candidate retenue est *névrologie*

La méthode est testée sur quatre langues : de l'anglais vers le français, l'allemand et l'espagnol. La liste d'éléments néoclassiques alignés a été créée manuellement. Les entrées françaises ont été reprises de Béchade (1992) puis traduites vers l'anglais, l'allemand et le français. La précision obtenue varie entre 96 % et 97 %. Le rappel varie entre 30 % et 37 %. En l'utilisant l'anglais comme langue pivot, Harastani *et al.* créent également des lexiques français-allemand, espagnol-français et allemand-espagnol avec une précision variant entre 97 % et 100 %. Le rappel varie entre 18 % et 35 %.

Weller *et al.* (2011) proposent une méthode similaire pour traduire les composés néoclassiques à ceci près que la structure des composés est légèrement différente : `R1 trans? R2 suffixe?`, c'est-à-dire :

R1 Une racine de type 1 - équivalent aux ICF d'Harastani *et al.*.

trans ? Un élément transitionnel comme *o* ou *i* que l'on retrouve par exemple dans *Kalorimetrie*. Ces éléments permettent de traiter des cas d'allomorphie, par exemple la racine */hydr/* peut se réaliser sous la forme *hydr-* (*hydravion*) ou *hydro-* (*hydrologie*).

R2 Une racine de type 2 - équivalent aux FCF d'Harastani *et al.*.

9. Ceux qui ne correspondent pas à cette forme ne sont pas traduits.

suffixe ? Un suffixe optionnel.

La méthode est testée sur deux paires de langues (allemand-français, allemand-anglais) et dans les deux sens de traduction. Lorsque les traductions candidates sont filtrées sur des termes extraits du corpus, la précision varie de 97 % à 99 %[10]. Concernant les traductions candidates non présentes dans le corpus, il s'avère que 50 % à 78 % d'entre elles sont correctes. Aucune information n'est donnée concernant le rappel.

3.1.3.3 Traduction de mots construits par composition populaire

Les composés populaires correspondent simplement à des mots découpables en d'autres mots. À la différence des composés néoclassiques, les composés populaires sont donc décomposables en éléments qui peuvent être employés de façon autonome. Il se distinguent des unités polylexicales en ceci que ces mots ne sont pas séparés par des espaces. Par exemple le nom allemand $Bleiisotope$ est composé à partir du mot $Blei$ *'plomb'* et du mot $Isotope$ *'isotope'*. Ce procédé est très courant en allemand et dans les langues germaniques en général.

Weller *et al.* (2011) se sont justement penchées sur le cas des composés nominaux allemands. Weller *et al.* se concentrent uniquement sur les structures NOM_1 NOM_2 qui sont traduites en français par la structure NOM PRÉPOSITION NOM et en anglais par les structures NOM NOM et NOM PRÉPOSITION NOM, ce qui permet de générer des traductions fertiles. Par exemple :

1. $Korrosionsschultz$ est découpé en $Korrosion$ et $schultz$. On voit ici que le découpage prend en compte l'élément de liaison s dans $Korrosion\underline{s}schultz$.
2. $Korrosion$ est traduit en anglais par $corrosion$ et $schultz$ est traduit par $protection$ grâce à un dictionnaire bilingue généraliste.
3. On identifie dans la liste des termes extraits du corpus cible tous les termes de structure NOM NOM et NOM PRÉPOSITION NOM qui contiennent $corrosion$ et $protection$, ce qui permet de retrouver les traductions $corrosion\ protection$ et $protection\ against\ corrosion$.

La méthode n'a pas été entièrement évaluée. Seul le nombre de traductions obtenues a été évalué, pas leur exactitude. Sur un ensemble de départ de 2 000 noms allemands (qui ne sont pas nécessairement des composés nominaux), les résultats obtenus sont :

- Pour la traduction vers le français :
  - 86 noms ont pu être traduits par d'autres méthodes (dictionnaire bilingue, méthode des composés néoclassiques, recherche de variantes graphiques).
  - Sur les 1 914 noms restants, 152 (8 %) ont pu être traduits par décomposition.
- Pour la traduction vers l'anglais :
  - 636 noms ont pu être traduits par d'autres méthodes.
  - Sur les 1 364 noms restants, 248 (18 %) ont pu être traduits par décomposition.

La différence de résultat dans les deux langues s'explique par la taille des dictionnaires utilisés (30 000 entrées pour l'allemand-français et 820 000 entrées pour l'allemand-anglais).

Une autre approche de la traduction des mots composés consiste à passer par des langues pivots comme le font Garera et Yarowsky (2008). Leur but est de traduire des mots composés d'une langue source LS vers une langue cible LC en utilisant plusieurs dictionnaires bilingues associant des termes en langue cible à leur traduction dans une troisième langue qui sert de langue pivot (LP).

10. La précision n'est pas indiquée directement dans l'article, nous la calculons à partir du tableau 8, p.91 : in $\frac{correct}{TL-terms}$

La traduction se déroule ainsi :

1. Identification des mots composant l'unité à traduire :
   – $ekurudhë$ est décomposé en $hekur$ et $udhë$.
2. Génération d'une glose en langue cible, en traduisant chacun des mots :
   – $\{hekur,\ udhë\}$ devient $\{iron,\ path\}$.
3. Le même processus de découpage puis génération de glose en langue cible est appliqué aux entrées en LP des dictionnaires bilingues, par exemple :
   – $eisenbahn \rightarrow \{eisen,\ bahn\} \rightarrow \{iron,\ path\}$;
   – $ferrovia \rightarrow \{ferro,\ via\} \rightarrow \{iron,\ path\}$;
   – $järnvag \rightarrow \{järn,\ vag\} \rightarrow \{iron,\ path\}$.
4. Pour chaque entrée LP dont la glose est identique à celle générée pour la langue source, on sélectionne de la traduction associée (l'ordre des éléments peut être inversé de façon à couvrir les cas où l'ordre des éléments n'est pas conservé d'une langue à l'autre) :
   – $eisenbahn \rightarrow railroad$;
   – $ferrovia \rightarrow railroad$;
   – $järnvag \rightarrow railway$.
5. Les traductions candidates sont ordonnées en fonction de : $\frac{freq(g,e_c)}{freq(g)}$ où $freq(g, e_c)$ est le nombre de fois où le mot composé a été traduit par e en passant par la glose g et $freq(g)$ le nombre de fois où la glose g a été générée à partir des entrées des dictionnaires :
   – $ekurudhë \rightarrow railroad$.

L'intérêt de ce passage par une langue pivot est que cela permet de gérer des cas de divergences lexicales. Dans l'exemple, on voit que ceci permet de traduire le mot composé albanais $erkurudhë$ vers l'anglais $railway$ alors qu'un des deux éléments qui le compose n'est pas directement traduisible vers l'anglais ($rail$ n'est pas la traduction de $hekur$ qui signifie *'fer'*). Bien que ce ne soit pas mis en évidence dans l'article, cette méthode permettrait également de gérer les cas où un des deux termes n'a pas un sens compositionnel : $Krankenhaus$ *'sick house'* peut être traduit en $hospital$ en passant par $Sjukhus$ *'sick house'*. Toutefois, son désavantage est de nécessiter des exemples de traductions de mots composés en plusieurs langues (Garera et Yarowsky travaillent avec entre 10 et 50 langues). Les auteurs parviennent à générer une traduction candidate pour 13,20 % des mots composés. Après ordonnancement, ils obtiennent une précision de 19,4 % sur le Top 1 et 36,3 % sur le Top 10.

3.1.4 Filtrage des traductions générées

Une fois les traductions générées, il est courant qu'un terme source se trouve associé à plusieurs traductions candidates. Il est aussi possible que le processus de traduction génère des termes agrammaticaux en langue cible. Il convient donc de filtrer ces traductions. Nous avons rencontré trois méthodes de filtrage dans l'état-de-l'art : (i) recherche d'une attestation de la traduction générée dans le corpus cible (avec éventuellement un ordonnancement en fonction de la fréquence) ; (ii) similarité des contextes d'apparition (application de la méthode distributionnelle) et (iii) apprentissage supervisé.

3.1.4.1 Recherche d'une attestation

Cette méthode consiste à rechercher l'attestation de la traduction générée dans le corpus cible. C'est la méthode employée par exemple par Morin et Daille (2010), Harastani *et al.*

(2012) et Weller *et al.* (2011) qui travaillent avec des corpus comparables : des termes sont préalablement extraits de la partie cible du corpus et c'est sur cette liste de termes cibles que sont filtrées les traductions générées. Dans ce cas, il est possible qu'une traduction possible soit rejetée car non extraite par l'extracteur de terme alors qu'elle est présente dans le corpus. Grefenstette (1999), Cartoni (2009b) et Léon (2008) utilisent un moteur de recherche ainsi que le nombre de réponses retournées par le moteur pour s'assurer que les traductions générées existent bien en langue cible.

Le recours à Internet est à manipuler avec précaution : la masse de données est telle qu'il est tout à fait possible d'y trouver des attestations de mots *a priori* impossibles ou incorrects en langue cible (fautes d'orthographes, jeux de création lexicale...). C'est pourquoi, en général, les auteurs s'appuient aussi sur la fréquence des attestations. Grefenstette, par exemple, retient la traduction la plus fréquente. Léon utilise une série de filtres statistiques comme le rapport de fréquence obtenus à partir du Web. Une traduction ne sera retenue que si sa fréquence est au moins supérieure à 1/10 000^e de la fréquence du terme source. L'idée est que terme source et traduction doivent avoir des fréquences comparables, le rapport de 1/10 000 s'explique par le fait que l'anglais (langue cible) est nettement plus représenté que le français (langue cible) sur Internet.

Un deuxième risque est que le Web contient des textes de natures très différentes, tant au niveau du domaine qu'au niveau du registre de langue ou des buts communicatifs. Même si une traduction générée est attestée, il convient de s'assurer qu'elle apparaît dans des contextes d'usages similaires au terme d'origine, au moins en ce qui concerne le champ thématique. Par exemple, si on traduit $chemin\ de\ fer$ par $iron\ path$ en anglais, cette association de mots renvoie 26 600 résultats lorsqu'elle est requêtée sur Google au moment de la rédaction. Or, elle ne correspond pas du tout à la traduction de $chemin\ de\ fer$ mais au titre d'un album de musique. Une solution est donc de procéder à une comparaison des contextes d'apparition.

3.1.4.2 Filtrage basé sur la similarité des contextes d'apparition

Cette méthode de filtrage exploite le même principe que la méthode distributionnelle : plus les contextes sont similaires, plus il est possible que termes source et cible soient des traductions l'une de l'autre.

Léon (2008) compare ce qu'elle appelle les « *mondes lexicaux* » (*op. cit.*, p. 190) du terme source et de sa traduction candidate. Le monde lexical d'un terme est construit en requêtant ce terme sur un moteur de recherche et en récupérant, sur les 1 000 premiers snippets de résultats, les 50 noms et les 50 adjectifs les plus fréquents. Comme dans l'approche distributionnelle, les mondes lexicaux sont traduits via un dictionnaire bilingue de façon à être comparables. La similitude entre monde lexical du terme source (S) et monde lexical de la traduction candidate (T) est donnée par le coefficient Jaccard :

$$Similitude(S,T) = \frac{|M(S) \cap M(T)|}{|M(S) \cup M(T)|} \tag{3.1}$$

où $M(X)$ représente l'ensemble des mots du monde lexical de X. Seules les paires de traductions ayant une similitude supérieure à un seuil donné sont retenues.

Baldwin et Tanaka (2004) procèdent différemment. Ils utilisent un corpus comparable construit à la volée à partir du web et de paires de traductions appartenant à un même domaine. Les termes sources sont utilisés comme requête dans un moteur de recherche pour construire le corpus source et les termes cibles sont utilisés comme requête dans un moteur de recherche

pour construire le corpus cible. L'alignement se fait entre termes extraits du corpus source et termes extraits du corpus cible. Seuls les termes appartenant au même domaine que les paires de traduction de départ sont conservés. L'appartenance au domaine est évaluée sur la base du coefficient Jaccard (dans ce cas, $M(X)$ correspond aux documents renvoyés par le moteur dans lequel le terme X apparaît).

Garera et Yarowsky (2008) ont également testé l'apport de la similarité des contextes pour ordonner des traductions candidates. Leurs expériences ont porté sur le couple allemand-suédois à partir du corpus parallèle Europarl (15 et 21 millions de mots respectivement) et non du Web. L'apport de cette stratégie est faible : la précision sur le Top1 passe de 19,6 % à 20,1 % et de 38,8 % à 39,1 % sur le Top10.

3.1.4.3 Filtrage par apprentissage supervisé

Baldwin et Tanaka (2004) traduisent des composés nominaux du japonais vers l'anglais en suivant la méthode compositionnelle. La recomposition du terme en langue cible est assurée par des patrons de traduction. Un patron de traduction est de forme $\mathtt{N}_2^E$ `in` $\mathtt{N}_1^E$ où $\mathtt{N}_i^E$ correspond à un nom anglais qui est la traduction du mot japonais de rang *i*. Des variantes morphologiques peuvent être utilisées de façon à faire correspondre la traduction du nom japonais et patron de traduction. Par exemple, dans *kaNkei · kaizeN*, *kaNkei* peut se traduire par *relation, connection, relationship* et *kaizeN* peut se traduire par *improvement, betterment*. Les traductions candidates générées sont : *relation improvement, betterment of relationship, improvement connection, relational betterment* parmi lesquelles la traduction *relational betterment* a été générée en exploitant la parenté morphologique entre *relation* et *relational*.

Une première technique de filtrage est basée sur un score nommé CTQ (Corpus-based Translation Quality). Pour une traduction générée par le patron t et contenant les mots w_1 et w_2, le score CTQ prend en compte :

- $p(w_1, w_2, t)$: la probabilité d'avoir w_1 et w_2 dans le patron t
- $p(w_1, t)$: la probabilité d'avoir w_1 dans le patron t.
- $p(w_2, t)$: la probabilité d'avoir w_2 dans le patron t.

Les probabilités sont calculées sur un corpus en langue cible. Ce score correspond donc à un modèle de langue. Le score CTQ s'obtient ainsi :

$$CTQ(w_1, w_2, t) = \alpha \cdot p(w_1, w_2, t) + \beta \cdot p(w_1, t) \cdot p(w_2, t) \tag{3.2}$$

où α et β sont des coefficients pondérateurs ($0 \geq \alpha, \beta \geq 1$ et $\alpha + \beta = 1$).

Dans une seconde expérience, Baldwin et Tanaka ordonnent les traductions à partir d'un modèle appris sur des exemples de traductions : un exemple positif est une paire de traduction correcte ; un exemple négatif est une paire où la traduction candidate est fausse. Chaque exemple est associé à trois types de variables prédictives :

Traits issus du corpus Il s'agit du score CTQ ainsi que de diverses fréquences d'occurrences et de cooccurences du patron de traduction et des mots inclus dans la traduction (8 traits en tout).

Traits issus des dictionnaires bilingues Il s'agit principalement de probabilités de traduction (6 traits en tout).

Traits basés sur les patrons Il s'agit de deux traits qui prennent en compte le type de patron et l'élément du composé qui agit comme tête (2 traits en tout).

La méthode d'apprentissage employée est celle des Machines à Vecteur de Support (SVM). Cette méthode donne en sortie une valeur continue entre +1 et -1 indiquant si l'individu classifié appartient plutôt à la classe positive (traduction correcte) ou plutôt à la classe négative (traduction incorrecte). C'est cette valeur qui est utilisée pour ordonner les traductions candidates.

De cette manière, Baldwin et Tanaka (2004) obtiennent 43 % de traductions correctes dans le sens japonais → anglais et 51 % de traductions correctes dans le sens anglais → japonais alors que le score CTQ employé seul donne 37 % dans le sens japonais → anglais et 42 % dans le sens anglais → japonais.

Dans cette dernière sous-section, nous avons vu que la sélection de traductions candidates pouvait être guidée par les données : à partir du Web, de corpus comparables ou encore grâce à des exemples de traductions issus d'un dictionnaire bilingue. Nous abordons maintenant des approches dans lesquelles la génération elle-même est faite à partir de règles issues d'exemples de traductions.

3.2 Approches empiriques

Les approches décrites ci-dessous sont qualifiées d'"empiriques" car, tout comme les approches empiriques de la TA, elles reposent sur l'analyse de grands volumes d'exemples de traductions. Nous décrivons deux types d'approches : la traduction par inférence analogique (Langlais *et al.*, 2009) et l'apprentissage de règles de réécriture de caractères (Claveau, 2009) qui peut être associé à l'apprentissage de familles morphologiques (Claveau et Kijak, 2011).

3.2.1 Traduction par inférence analogique

Lepage (2003) nous apprend que l'analogie est un concept ancien, énoncé en premier lieu par le philosophe grec Aristote sous la forme : « *A est à B ce que C est à D* ». L'analogie est introduite en linguistique par Saussure où elle est définie comme « *l'opération par laquelle, étant données deux formes d'un même mot, et seulement une forme d'un second mot, on crée la forme manquante* » (*op. cit.*, p. 92). C'est, par exemple, le raisonnement par lequel nous pouvons retrouver le pluriel de $cheval$ à partir d'un exemple comme $journal \rightarrow journaux$. Nous devons alors résoudre l'équation analogique suivante :

$$[journal : journaux = cheval : x]$$

dont la solution est :

$$\Rightarrow x = chevaux$$

L'analogie est donc une relation de proportion entre quatre éléments qui se note ainsi :

$$[x : y = z : t]$$

et se glose par « x *est à* y *ce que* z *est à* t ». Une équation analogique est une analogie dans laquelle il manque le quatrième terme :

$$[x : y = z :?]$$

Toujours selon Lepage, l'analogie formelle (i.e. qui s'applique à des chaînes de caractères) a été utilisée avec succès en TAL dans le domaine de la TA (paradigme de la traduction par l'exemple) et en prononciation automatique (transformation d'une chaîne de graphèmes en phonèmes).

Les travaux de Langlais *et al.* (2009) proposent d'utiliser l'analogie pour traduire des termes monolexicaux appartenant au domaine médical. La traduction par inférence analogique se fait à partir d'un ensemble de termes de langue source I, d'un ensemble de termes de langue cible O et de relations de traduction entre ces termes. Partons, par exemple, d'un ensemble de mots sources $\{constitution,\ profession,\ constitutionnel,\ professionnel\}$, d'un ensemble de mots cibles $\{constitution,\ profession,\ constitutional,\ professional\}$ et de relations de traductions $\{(constitution,\ constitution),\ (profession,\ profession),\ (constitutionnel,\ constitutional)\}$. Nous cherchons à établir la relation de traduction manquante entre le mot source $professionnel$ et le mot cible $professional$.

La procédure de traduction par inférence analogique d'un mot source i est :

1. Collecter l'ensemble des triplets (x, y, z) tels que x,y et z forment une analogie avec i, dans l'exemple : $(constitution,\ profession,\ constitutionnel)$ pour $professionnel$;
2. Traduire ces triplets, dans notre exemple, on obtient $(constitution,\ profession,\ constitutional)$
3. Sélectionner l'ensemble des mots cibles $\{o_1, o_2..o_n\}$ qui forment une analogie avec les triplets obtenus dans l'étape 2 ; dans l'exemple on obtient $professional$;
4. Sélectionner la bonne traduction parmi $\{o_1, o_2...o_n\}$; une seule solution dans l'exemple : $professional$.

Les étapes 1. et 3. posent des équations analogiques que Langlais *et al.* (2009) résolvent de la façon suivante :

1. Soit une équation à résoudre : $[x\ : y = z\ :?]$
   - ex. : $[\ constitution : profession = constitutionnel : ?]$
2. Faire plusieurs mélanges aléatoires de y et z en gardant l'ordre des lettres :
   - ex. : $constiprofesstuiontion$, $pcronosftietsutsiiononnel$, etc.
3. Pour chacun de ces mélanges, établir son complémentaire par rapport à x :
   - ex. : $constiprofesstuiontion \setminus constitution \rightarrow professionnel,\ prosfestnionel,\ etc.$
   - ex. : $pcronosftietsutsiiononnel \setminus constitution \rightarrow porfiesonnel,\ professionnel,\ etc.$
4. Le(s) complémentaire(s) obtenus représentent les solutions possibles de l'équation analogique.

Il faut noter que l'approche analogique est confrontée à des problèmes de coût calculatoire, notamment au niveau du mélange des chaînes de caractères (résolution de l'équation analogique, étape 2) et au niveau du traitement de tous les triplets formant une analogie avec le mot à traduire (traduction par inférence analogique, étapes 1 et 3). De plus, la traduction par inférence analogique génère plusieurs traductions candidates pour un même terme source qu'il convient de filtrer dans une troisième étape. Langlais *et al.* effectuent la sélection de la meilleure traduction candidate par apprentissage automatique. Pour cela, ils utilisent un classifieur binaire basé sur l'algorithme de perceptron appelée *voted-perceptron* (Freund et Schapire, 1999). Le modèle de classification est appris sur des exemples de relations analogiques alignées comme dans l'équation 3.3 où la relation analogique r est alignée avec la relation analogique r', x étant la traduction de x', y la traduction de y' et ainsi de suite.

$$(r, r') = ([x : y = z : t], [x' : y' = z' : t']) \tag{3.3}$$

Les variables prédictives utilisées sont le degré de l'analogie, le nombre de fois où une forme est générée, les ratios de longueur entre t et t' et le score de vraisemblance de la traduction (calculé à partir de n-grammes appris sur corpus). Une expérimentation est menée sur 1 000 termes issus de l'UMLS et sur six langues : anglais, espagnol, finnois, français, russe, suédois. Les résultats vont de 53,6 % de traductions justes pour le sens finnois → anglais à 64,3 % de traductions justes pour le sens espagnol → anglais.

3.2.2 Apprentissage de règles de réécriture de caractères

Claveau (2009) cherche à traduire des termes médicaux monolexicaux. Il constate que ces derniers sont généralement construits à partir de racines gréco-latines présentes dans de nombreuses langues et que les règles de dérivation morphologique sont également régulières de langue à langue, ce qui donne des paires de traduction aux graphies proches, par ex. : $ophtalmorragie \rightarrow ophthalmorragia$ et $leucorragie \rightarrow leukorrhagia$. Les variations graphiques sont estimées suffisamment régulières pour permettre un apprentissage de règles de réécriture, c'est-à-dire des règles qui s'appliquent à une ou des sous-chaînes du terme à traduire et produisent l'équivalent de cette sous-chaîne en langue cible. Par exemple, si on applique les règles de réécriture $leuco \rightarrow leuko,\ rragie \rightarrow rrhagia$ à $leucorragie$, on obtient $leukorrhagia$.

Les règles de réécriture sont apprises sur une liste de paires de traductions. Le processus est identique à celui de la traduction automatique statistique, si ce n'est qu'il est basé sur des suites de lettres et non des suites de mots. On y retrouve les trois modules d'un système de traduction automatique statistique : (i) alignement ; (ii) d'apprentissage d'un modèle de transfert pour la traduction (règles de réécriture) et (iii) apprentissage d'un modèle de langue.

Les paires de traductions sont alignées au niveau des lettres avec l'outil DPalign [11] qui a l'avantage de pouvoir aligner des séquences n'ayant pas le même alphabet. Les règles de réécritures sont inférées à partir des alignements : à chaque fois que, dans une paire de traductions, deux lettres alignées ne sont pas identiques, toutes les règles de réécriture qui peuvent permettre de faire le lien entre ces deux lettres sont générées. Par exemple, pour la différence e/a dans le couple $(leucorragie,\ leukorrhagia)$, il est possible de générer plusieurs règles : $e \rightarrow\ e,\ ie \rightarrow\ ia,\ gie \rightarrow\ gia$, etc. Le choix de la meilleure règle est fait sur la base du ratio entre le nombre de fois où l'entrée de la règle correspond à une sous-chaîne d'un mot source et le nombre de fois où la règle peut s'appliquer en entier à une paire de mots. Le modèle de langue, quant à lui, correspond à des probabilités de n-grammes [12] de lettres apprises sur la partie cible des paires de traductions.

Pour traduire un terme, toutes les règles de réécriture possibles lui sont appliquées, ce qui crée plusieurs traductions candidates. La traduction la plus probable est sélectionnée grâce au modèle de langue :

$$P(w) = \prod_{i=1}^{m} P(l_i | l_{i-n+1}, ..., l_{i-1}) \tag{3.4}$$

où $P(l_i|l_{i-n+1}, ..., l_{i-1})$ correspond à la probabilité d'avoir la $i$ème lettre du mot sachant les n lettres qui la précèdent.

Claveau (2009) obtient 85,4 % de traductions correctes dans le sens français → anglais et 84,8 % dans le sens anglais → français à partir de 5 400 exemples tirés de l'UMLS (Lindberg *et al.*, 1993). Si la technique est testée sur d'autres couples de langues, les résultats baissent

11. http ://search.cpan.org/ cjfields/BioPerl-1.6.1/Bio/Tools/dpAlign.pm
12. Un n-gramme est une suite n éléments, ici des lettres.

en fonction des similarités morphologiques et graphiques entre langues : de 87,9 % pour le sens portuguais → espagnol à 57,5 % pour le sens anglais → russe.

3.2.3 Traitement de la variation morphologique

Pour finir, nous citerons les travaux de Claveau et Kijak (2011) dont la technique de traduction est basée sur un découpage de l'unité à traduire en *morphes* c'est-à-dire des ≪ *signes linguistiques élémentaires (segments)* ≫ qu'ils distinguent des *morphèmes* qui sont des ≪ *classes d'équivalence avec un signifiant identique et des signifiés proches* ≫ [13]. Ces travaux peuvent être rapprochés des approches compositionnelles décrites dans le paragraphe 3.1.3 qui traduisent des unités monolexicales en passant par un découpage morphologique. À la différence de ces approches symboliques, le système proposé par Claveau et Kijak (2011) apprend les équivalences de traduction entre morphes à partir d'exemples. Il est aussi capable de gérer des cas de fertilité.

Partant d'une base d'exemples de termes médicaux alignés (8 000 paires extraites de l'UMLS), Claveau et Kijak (2011) apprennent des probabilités d'alignement entre chaînes de caractères grâce à une adaptation de l'algorithme espérance-maximisation (Dempster, 1977). Cet algorithme leur permet d'obtenir des alignements avec une chaîne de caractères vide (cas de traductions fertiles) mais ne prend pas en compte la distortion (cas où l'ordre des éléments alignés est différent).

L'algorithme d'espérance-maximisation fonctionne en deux phases qui sont répétées jusqu'à obtenir une convergence :

Espérance (initialisation) La phase d'espérance compte les alignements possibles entre des sous-chaînes de caractères du terme source et des sous-chaînes de caractères du terme cible. Ces sous-chaînes de caractères sont ce que Claveau et Kijak (2011) appellent des morphes. Ces comptes sont stockés dans une table d'alignement de morphes notée γ

Maximisation (initialisation) La phase de maximisation calcule les probabilités d'alignement des morphes en normalisant les comptes de γ, ces probabilités sont stockées dans une table notée δ :

$$\delta(s,t) = \frac{\gamma(s,t)}{\sum_x \gamma(s,x)}$$

où s est un morphe source et t est un morphe cible.

Espérance (itération 1) Les probabilités stockées dans δ sont réutilisées pour pondérer les comptes des alignements, une nouvelle table γ est produite.

Maximisation (itération 1) À nouveau, les comptes sont normalisés pour produire une nouvelle version de la table δ.

... les itérations se succèdent jusqu'à avoir atteint le critère d'arrêt ...

Arrêt les itérations cessent lorsqu'il n'y a plus ou peu de différence entre la table δ obtenue à l'itération i et la table δ obtenue à l'itération $i+1$.

Une des limites de cette approche est de ne pas prendre en compte la variation allomorphique c'est-à-dire le fait qu'un même morphème se réalise sous différentes formes. Cette variation provoque une dispersion des probabilités d'alignement. Si nous prenons l'exemple de la traduction du japonais vers le français, qui est le cas traité par Claveau et Kijak

13. ≪ *we distinguish morphs, elementary linguistic signs (segments), from morphemes, equivalence classes with identical signified and close significants* ≫ (*op. cit.*, p. 237)

(2011), on sait qu'en français le sens de *'bactérie'* peut se réaliser sous la forme $\ll bactérie \gg$, $\ll bactéri\text{-} \gg$ comme dans $bactéricide$ ou encore $\ll bactério\text{-} \gg$ comme dans $bactériologique$. Or, ces trois morphes vont être chacun aligné séparément avec leur équivalent japonais kin [14] :

SOURCE	CIBLE	PROBABILITÉ
kin	bactérie	0,4
kin	bactério	0,3
kin	bactéri	0,2

Le calcul des probabilités d'alignement de la phase de maximisation est donc modifié de façon à prendre en compte tous les équivalents morphologiques :

$$\delta(s,t) = \frac{\sum_{m \in M} \gamma(s,m)}{\sum_x \gamma(s,x)}$$

où M est l'ensemble des morphes associés à t, c'est-à-dire dans notre exemple, si $t = bactérie$, $M = \{bactérie, bactéri, bactério\}$.

Comme décrit dans l'algorithme 1, cet ensemble de morphes M est obtenu en appliquant des règles de réécriture à t elles-mêmes apprises à partir des alignements de l'itération précédente : si deux morphes sont alignés avec un même kanji dans la table γ et que leur plus longue sous-chaîne commune ($plscc$) est supérieure à une longueur donnée, alors on considère qu'ils correspondent au même morphème. Par exemple, $dermo$ et $dermato$ sont alignés au même kanji et partagent la sous-chaîne $derm$. De cette paire de morphes, on peut déduire la règle de réécriture $r = plscc(m_1, m_2) \ominus ato \oplus o$, c'est-à-dire $\ll$ *réécrire le suffixe* ato *en* $o \gg$. Appliquée au morphe $hémato$, cette règle permettra de l'identifier comme un variante de $hémo$ et vice-versa. Ce système de règles de réécriture exploite les analogies formelles entre morphes, dans notre exemple : $dermato : dermo = hémato : hémo$. Dans leurs expériences, **?** ne prennent en compte que des analogies de degré 3 : les règles de réécriture apprises sont de la forme `préfixe` $\oplus$ `base` $\oplus$ `suffixe`.

Algorithme 1 Étape de maximisation modifiée pour la prise en compte des variantes morphologiques

Require: γ
  for all sous-chaîne s t.q. $\gamma(s, \cdot)$ **do**
    for all m_1, m_2 t.q. $\gamma(s, m_1) > 0$ and $\gamma(s, m_2) > 0$ and $plscc(m_1, m_2) > seuil$ **do**
      construire la règle de préfixation et de suffixation r à partir de m_1 et m_2
      incrémenter le score de r
    end for
    for all sous-chaîne t t.q. $\gamma(s, t) > 0$ **do**
      construire l'ensemble M des morphes associés à t avec l'aide des n règles de réécriture les plus fréquentes issues de la précédente itération
      calculer $\delta(s, t)$
    end for
  end for
  return δ

Peu de détails sont donnés sur la manière de traduire les termes médicaux français vers le japonais si ce n'est que les probabilités de traductions des morphes dans δ sont exploitées

14. Nous utilisons l'équivalent en script roman des kanjis (romanji).

dans un algorithme de type Viterbi (Viterbi, 1967). Sur 128 termes à traduire, 92 ont obtenu une traduction (71,8 %). Sur ces 92 traductions, 58 correspondaient à celles de l'UMLS (63 %). En prenant en compte les traductions apparaissant dans des ressources médicales en ligne, le nombre de traductions correctes passe à 82 (89,1 %).

3.3 Évaluation des méthodes de génération de traductions

La génération de traduction est évaluée d'une façon similaire à l'alignement de termes par la méthode distributionnelle à ceci près qu'il est important de pouvoir évaluer le pouvoir génératif des méthodes, c'est-à-dire leur capacité à générer au moins une traduction candidate pour chaque terme source, quelle que soit l'exactitude de la traduction. Cette information est donnée par la couverture, notée C, qui est calculée ainsi :

$$C = \frac{|ST|}{|S|} \tag{3.5}$$

$$ST = \{s : |\mathcal{T}(s)| > 1\}$$

où S est l'ensemble des termes sources et $\mathcal{T}(s)$ est l'ensemble des traductions générées par le système pour le terme source s.

La précision indique, pour les termes sources pour lesquels le système a généré une traduction, la proportion de ceux qui ont reçu au moins une traduction correcte. Comme plusieurs traductions peuvent être générées, la précision peut-être calculée sur les N meilleures traductions :

$$P_N = \frac{|SR_N|}{|ST|} \tag{3.6}$$

$$SR_N = \{s : \mathcal{T}_N(s) \cap \mathcal{R}(s) \neq \emptyset\}$$

où $\mathcal{T}_N(s)$ est l'ensemble des N premières traductions de s et $\mathcal{R}(s)$ est l'ensemble des traductions de référence de s. Comme la génération de traduction offre une bonne précision, c'est la précision sur le Top1 qui est utilisée dans la plupart des cas ($N = 1$). Parfois, la précision est calculée sur l'ensemble des traductions proposées par le système ($N = |T(s)|$).

Le rappel indique la fraction des termes sources pour lesquels le système a pu générer au moins une traduction exacte parmi les N meilleures traductions :

$$R_N = C \times P_N = \frac{|SR_N|}{|S|} \tag{3.7}$$

Enfin, la F1-mesure permet de rendre compte du compromis entre précision et rappel :

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{3.8}$$

Certains auteurs prennent en compte deux niveaux d'exactitude pour la traduction. Un premier niveau est la traduction dite de *référence*, c'est-à-dire une traduction qui existe dans une ressource linguistique de référence (par exemple, l'UMLS dans le domaine médical) ou qui est validée par un expert comme étant la traduction canonique. Pourtant, il existe de nombreux cas où une traduction générée par le système, bien qu'elle ne corresponde pas à la définition d'une traduction de référence, reste tout à fait acceptable et utilisable. Ce concept de traduction

acceptable se retrouve sous le nom de « *silver-standard* » chez Baldwin et Tanaka (2004), de traduction « *acceptable* » chez Léon (2008), et de traduction « *incertaine* » chez Cartoni (2005).

Les tableaux 3.2 et 3.3 récapitulent les résultats obtenus par les techniques de traduction compositionnelle et les approches empiriques. C correspond à la couverture. P, R et F1 correspondent à la précision, au rappel et à la F1-mesure lorsque l'on prend en compte uniquement les traductions de référence sur le Top1 sauf dans le cas de Harastani *et al.* (2012) et Weller *et al.* (2011) qui considèrent l'ensemble de toutes les traductions d'un terme source (résultats marqués d'une astérisque). P_A, R_A, $F1_A$ correspondent à la couverture, précision, rappel et F1-mesure lorsque l'on prend en compte les traductions de référence et les traductions acceptables (toujours Top1). Les sigles UPL et UML correspondent aux unités poly- et mono-lexicales respectivement. Le sigle LG signifie langue générale.

Les résultats ne sont pas comparables car les expérimentations reposent sur des jeux de données différents mais on observe que, globalement, les méthodes de génération de traduction, si elles n'ont pas la capacité de proposer une traduction pour tous les termes sources qui leur sont soumis, proposent des traductions plus précises que l'approche distributionnelle. Avec les méthodes de génération de traduction, la précision varie entre 19 % et 94 % sur le Top1 ; alors qu'avec l'approche distributionnelle, la précision varie entre 13 % à 65 % pour le Top1 et 30 % à 89 % sur le Top10.

Le principal inconvénient de la génération de traduction reste sa couverture : beaucoup de termes ne peuvent être traduits via cette méthode, surtout dans les approches compositionnelles qui tendent à se focaliser sur un type de structure bien précis (par exemple 8 % de traductions générées dans le cas des composés nominaux allemands de Weller *et al.* (2011)). Les approches empiriques affichent globalement des résultats meilleurs : 54 % à 85 % de précision sur le Top1 avec 72 % à 100 % de couverture contre 19 % à 94 % de précision sur le Top1 avec 8 % à 100 % de couverture pour les approches compositionnelles. Toutefois, ces dernières sont gourmandes en données spécialisées : au bas mot, 5 400 paires de termes médicaux alignés pour Claveau (2009) et jusqu'à 19 800 pour Langlais *et al.* (2009). *A contrario*, les approches compositionnelles ne requièrent "que" des dictionnaires bilingues (qui sont généralement facilement disponibles) et quelques patrons de traductions.

RÉFÉRENCE	MÉTHODE FILTRAGE	LANGUES	ÉLÉMENTS À TRADUIRE NB.	TYPE	DOMAINE	RESSOURCES POUR TRADUCTION TAILLE	TYPE	RÉSULTATS C	P	R	F1	P_A	R_A	$F1_A$
Grefenstette (1999)	Web + fréquences	DE → EN ES → EN	724 1140	UPL	LG	36,7k	dico LG		,87 ,86					
Robitaille *et al.* (2006)	Web + similarité contextes	FR → JA	194	UPL	IA/TAL	50k 50k 96k	dico LG dico scientifique thésaurus JA, LG	1	,81	,82	,81			
Baldwin et Tanaka (2004)	classifieur binaire	EN → JA JA→ EN	750 750	UPL	presse	550k	dico LG	,92 ,98	,51 ,44	,47 ,42	,49 .43	,78 ,84	,72 ,82	,75 ,83
Morin et Daille (2010)	score de similarité contextuelle (corpus 800k mots)	FR→ JA	829	UPL	médical	173k	dico LG, médical, technique		,88					
Morin et Daille (2012)	fréquence (corpus de 530k mots)	EN → FR	836	UPL	médical	246k	dico LG	,17	,73	,27				
	fréquence (220k mots)	EN → DE	964	UPL	médical	171k	dico LG	,09	,89	,16				
	score de similarité contextuelle (corpus de 530k mots)	EN → FR	836	UPL	médical	246k	dico LG + vecteurs de contexte	,61	,42	,50				
	termes extraits du corpus (220k mots)	EN → DE	964	UPL	médical	171k	dico LG + vecteurs de contexte	,53	,44	,48				
Léon (2008)	Web + fréquences + contextes	FR→ EN	1075	UPL	LG		dico LG	,83	,89	,74	,81	,94	,78	,85
Cartoni (2005)	aucun	IT → FR	115	mots préfixés	presse		dico LG + RCL	1	,94	,94	,94	1	1	1
Cartoni (2009b)	Web + fréquences	IT → FR	30376	mots préfixés	presse		dico LG + RCL			.42 à ,94				
Harastani *et al.* (2012)	termes extraits du corpus	EN → FR	1068	composés	environnement	146k + 83	dico LG +	,37	,98*	.36*	,53*			
		EN → DE	3538	savants		70k + 61	racines savantes	,36	,96*	.35*	,51*			
		EN → ES	2126			62k + 58		,3	,97*	,29*	,45*			
Weller *et al.* (2011)	termes extraits du corpus	DE → EN	364	composés	environnement		racines		,99*					
		EN → DE	315	savants			savantes		,97*					
		DE → EN	1364	composés	environnement	820k	dico LG	,18						
		DE → FR	1914	populaires		30k		,08						
Garera et Yarowsky (2008)	probabilité	SQ, BG, CS, FA, DE, HU, RU, SK, SV[15] → EN	10273	composés populaires	LG		dico LG	,13	,19	,03	,05			

TABLE 3.2 – Résultats de l'état de l'art - génération de traductions par traduction compositionnelle

RÉFÉRENCE	MÉTHODE FILTRAGE	LANGUES	ÉLÉMENTS À TRADUIRE NB.	TYPE	DOMAINE	RESSOURCES POUR TRADUCTION TAILLE	TYPE	RÉSULTATS C	P	R	F1	P_A	R_A	$F1_A$
Claveau (2009)	probabilité	FR → EN	1000	UML	médical	5,4k	dico médical	1	,85	,85	,85			
		EN → FR	1000			5,4k		1	,85	,85	,85			
		PT → ES	1000			5,4k	UMLS	1	,88	,88	,88			
		EN → RU	1000			5,4k		1	,58	,58	,58			
Langlais *et al.* (2009)	classifieur binaire	FR → EN	1000	UPL,UML	médical	17,3k		1	,57	,19	,29			
		SP → EN	1000			19k	UMLS	1	,63	,23	,33			
		FI → EN	1000			19,8k		1	,54	,21	,3			
Claveau et Kijak (2011)	probabilité	FR → JA	128	UPL,UML	médical	6,4k	UMLS	,72	.63	.45	,53	,89	,64	,74

TABLE 3.3 – Résultats de l'état de l'art - génération de traductions fondée sur les données

3.4 Perspectives de recherche

Nous avons constaté dans le chapitre précédent que les traducteurs jugeaient les lexiques extraits via les méthodes distributionnelles peu utilisables (trop de traductions candidates, précision trop faible). Ces derniers préfèrent avoir peu de traductions mais que celles-ci soient plus fiables. Après cet état-de-l'art sur les méthodes de génération de traductions, nous constatons que cette approche semble mieux adaptée à l'extraction de lexiques destinés à la traduction spécialisée.

Nous avons fait état de deux manières de générer des traductions. D'une part, on trouve la traduction compositionnelle que l'on peut décrire comme une approche "experte" basée sur des dictionnaires de langue générale et une analyse linguistique du terme à traduire. D'autre part, nous rencontrons des méthodes empiriques qui utilisent des ressources spécialisées bilingues comme exemples de traductions. Le tableau 3.4 synthétise les forces et les faiblesses des méthodes distributionnelle, compositionnelle et empirique.

	DISTRIBUTIONNELLE	GÉNÉRATION	
		COMPOSITIONNELLE	EMPIRIQUE
précision	-	+	+
couverture	+	-	-
indépendant de la fréquence des termes à traduire	-	+	+
fonctionne sans exemples	+	+	-

TABLE 3.4 – Comparatif des méthodes d'acquisition automatique de lexiques bilingues

Nous choisissons de continuer nos recherches dans le cadre de la traduction compositionnelle pour plusieurs raisons :

- Par rapport à l'approche distributionnelle, la traduction compositionnelle est moins dépendante de la taille du corpus. En effet, une fois la traduction générée, une seule occurrence dans le corpus cible peut suffire pour valider la traduction, alors que l'approche distributionnelle nécessite que termes sources et cibles soient suffisamment fréquents pour obtenir des vecteurs de contexte représentatifs. De plus, nous savons grâce aux travaux de Morin et Daille (2010) que la méthode compositionnelle donne de meilleurs résultats que la méthode distributionnelle lorsqu'il s'agit de traduire des termes avec un sens compositionnel.
- Concernant les approches empiriques, ces dernières nécessitent qu'il existe déjà des lexiques spécialisés ou des terminologies bilingues. Ce n'est pas un hasard si de telles approches semblent cantonnées au domaine du médical et exploitent toutes la même ressource (UMLS). Ces approches ne permettent pas d'aborder un nouveau domaine sans passer par un travail de constitution d'un lexique bilingue spécialisé de taille conséquente (au bas mot 5 400 entrées). Elles sont donc restreintes aux domaines pour lesquels il existe déjà un minimum de ressources disponibles. Or, notre but est d'être capable d'amorcer l'acquisition de lexiques bilingues lorsqu'il n'existe pas encore de ressources spécialisées.

La principale limite de la traduction compositionnelle est sa couverture : seules certaines unités lexicales peuvent être traduites. Ces unités doivent être complexes, c'est-à-dire décomposables en sous-unités. C'est le cas des unités polylexicales ou encore des unités monolexicales issues d'un processus de construction morphologique. De plus, leur équivalent

en langue cible doit également être décomposable et il doit y avoir bijection entre composants du terme source et composants du terme cible (figure 3.1).

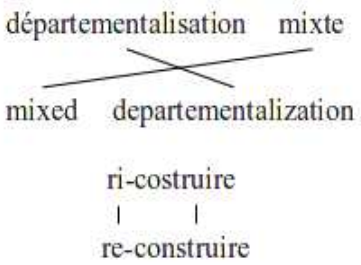


FIGURE 3.1 – Exemples d'alignements identifiables avec la traduction compositionnelle

Des paires de traductions dont un des éléments n'est pas complexe ou dont le sens des composants est trop éloigné ne pourront pas être traduits comme par exemple l'alignement entre $water\ color$ *'eau couleur'* et $aquarelle$ ou entre $pink\ eye$ *'rose œil'* et $conjonctivite$.

Malgré cela, la traduction compositionnelle reste avantageuse pour identifier des traductions d'unités lexicales spécialisées puisqu'il est estimé que les mots complexes forment plus de 60 % des nouveaux termes rencontrées dans les domaines techniques et scientifiques (Namer et Baud, 2007) [16].

La plupart des difficultés de la traduction compositionnelle ont été abordées - avec plus ou moins de succès - par les travaux antérieurs :

- Le non-parallélisme des structures syntaxiques est aisément traité par les méthodes compositionnelles grâce aux permutations des composants traduits, à l'approche "sac de mots" ou aux patrons de traduction.
- Les divergences morphologiques peuvent être traitées grâce à l'utilisation d'un mot ou d'une base lexicale appartenant à la même famille morphologique que le composant à traduire - voir par exemple l'utilisation des adjectifs relationnels chez Morin et Daille (2010) et Cartoni (2009a).
- Les divergences lexicales sont gérées avec l'inclusion de synonymes ou de mots sémantiquement proches par Robitaille *et al.* (2006) ou par la prise en compte de toutes les traductions proposées par un lexique probabiliste (Š. Vintar, 2010).

La question de la fertilité reste la moins abordée à l'heure actuelle. Seuls Weller *et al.* (2011) et Claveau et Kijak (2011) en parlent explicitement. Chacun se focalise sur un cas particulier. Weller *et al.* (2011) se soucient de l'équivalence composé nominal $\leftrightarrow$ syntagme nominal ($Elektronenmikroskop \leftrightarrow electron\ microscope$). Claveau et Kijak (2011) établissent des correspondances entre racine néoclassique et mot graphiquement proche ($bactéri\text{-} \leftrightarrow bactérie$).

Par ailleurs, le traitement des unités monolexicales par l'approche compositionnelle manque de généricité. Chaque auteur se soucie d'un cas particulier de construction (préfixation, composition savante, composition populaire). Or, un mot peut être le résultat de plusieurs processus : $post\text{-}ovariectomie$, $clinico\text{-}pathologic$, $histiologically$, etc.

Enfin, le filtrage des traductions générées est très souvent basé sur un seul critère (attestation de la forme, fréquence, similarité contextuelle) et il n'existe pas d'étude, hors approches par apprentissage automatique, qui cherche à combiner ces critères.

Ces trois points (fertilité, indépendance à la structure morphologique du terme source,

16. ≪ *complex words form more than 60 % of the new terms found in techno-scientific domains* ≫ (*op. cit.*, p. 3)

combinaison des critères pour le filtrage) sont les pistes de recherches que nous avons choisi d'aborder dans ce mémoire. Nos recherches nous ont mené à développer une variante de la méthode de traduction compositionnelle que nous nommons méthode *morpho-compositionnelle*. Le cadre de nos recherches ainsi que les hypothèses de travail sont présentés au chapitre suivant.

Deuxième partie

Contributions à la traduction compositionnelle

Chapitre 4

Cadre méthodologique de la traduction morpho-compositionnelle

Sommaire

Introduction

Le but de ce chapitre est d'exposer le cadre méthodologique de nos recherches. La première section (4.1) définit notre méthode de traduction ainsi que les notions employées. Dans la section 4.2, nous nous efforçons de mettre en avant nos contributions. La méthodologie d'évaluation est présentée en section 4.3. Une synthèse du chapitre est donnée en section 4.4.

4.1 Méthode de traduction morpho-compositionnelle

Nous plaçant dans le cadre de la traduction compositionnelle d'unités monolexicales, nous souhaitons proposer une approche inspirée des travaux de Cartoni (2009b); Weller *et al.* (2011) et Harastani *et al.* (2012) qui soit suffisamment générique pour traiter divers types de constructions morphologiques ainsi que les cas où l'unité monolexicale peut être traduite par une unité polylexicale. De plus, nous souhaitons aussi proposer une méthode d'ordonnancement des traductions produites qui prennent en compte divers critères et pas seulement sa fréquence ou la similarité entre sa distribution et celle du terme source.

La méthode de traduction proposée, que nous nommons *méthode morpho-compositionnelle* se décompose en cinq étapes :

1. **Décomposer** le terme source en morphèmes ou en éléments "approchants" :
   – $post\text{-}menopausal$ est découpé en $post$ et $menopausal$
2. **Traduire**, à l'aide d'un dictionnaire bilingue, chacun de ces éléments en langue cible, la traduction peut faire appel à des **variantes morphologiques** ou à des **synonymes** :
   – $post$ peut être traduit par $post$ ou $après$
   – $menopausal$ peut être traduit par $ménopause$
3. **Recomposer** les éléments traduits de façon à générer un candidat terme en langue cible qui peut être composé d'un ou plusieurs mots :
   – $post$ et $ménopause$ peuvent être recomposés de 4 façons : $postménopause$, $ménopausepost$ (1 mot) et $post\ ménopause$, $ménopause\ post$ (2 mots).
   – $après$ et $ménopause$ peuvent aussi être recomposés de 4 façons : $aprèsménopause$, $ménopauseaprès$, $après\ ménopause$, $ménopause\ après$
4. **Rechercher une attestation** des candidats termes dans un corpus en langue cible :
   – seuls $postménopause$, $post\ ménopause$, $après\ (la)\ ménopause$ et $ménopause\ après$ peuvent être attestés dans un corpus
5. **Ordonner** les candidats extraits du corpus à partir d'un score obtenu par la combinaison de plusieurs critères :
   (a) $postménopause$
   (b) $après\ la\ ménopause$
   (c) $post\ ménopause$
   (d) $ménopause\ après$

Dans la suite de ce mémoire, nous employons le terme *génération de traduction* pour référer aux étapes 1 à 4 et le terme *ordonnancement des traductions* pour référer à l'étape 5.

La méthode que nous proposons ne s'appuie pas sur des patrons de traductions et ne repose pas sur des *a priori* quant à la structure morphologique du terme cible. Les étapes 1, 2 et 3 génèrent toutes les hypothèses de traduction possibles. À l'issue de l'étape 3, nous avons donc un grand nombre de traductions candidates, dont certaines sont des "monstres linguistiques" (ex. : $ménopauseaprès$). Les étapes 4 et 5 permettent de restreindre le nombre de traductions candidates en sélectionnant uniquement les traductions attestées en langue cible (étape 4) puis en les ordonnant de la plus à la moins plausible (étape 5).

Par ailleurs, notre méthode autorise la traduction d'un morphème lié par un morphème libre. Dans notre exemple, il s'agit de la traduction du préfixe $post$ par la préposition $après$. Ce type d'équivalences traductionnelles n'a, à notre connaissance, jamais été utilisé en ce qui concerne la traduction compositionnelle.

4.1.1 Positionnement

Des équivalences entre morphèmes libres et liés ont déjà été utilisées par Claveau et Kijak (2011) dans le cadre des approches empiriques. Dans ce travail, ils établissent des familles morphologiques comme $\{bactério\text{-},\ bactéri\text{-},\ bactérie\}$ qui sont alignés avec un même kanji et peuvent lui servir de traduction. Claveau et Kijak indiquent d'ailleurs que plusieurs traductions générées sont des paraphrases, ce qui montre l'intérêt de l'équivalence morphème lié $\leftrightarrow$ morphème libre pour générer des traductions fertiles.

Dans la même veine, Déléger (2009) génère des paraphrases dans le but de faciliter la compréhension des textes médicaux et notamment la compréhension des composés savants. Elle travaille à partir de corpus comparables monolingues, c'est-à-dire à partir de deux corpus de textes appartenant à un domaine de spécialité et traitant de la même thématique (ex. : tabac, cancer, diabète) mais dont un contient des textes destinés à des spécialistes (corpus scientifique) et l'autre contient des textes destinés à des non spécialistes (corpus vulgarisé). Des composés savants sont extraits du corpus scientifique puis leurs paraphrases sont générées en utilisant l'analyseur morpho-sémantique DÉRIF (Namer, 2005) qui est capable de générer des gloses de mots morphologiquement complexes (par exemple, $gastrite$ est glosé en *'inflammation de l'estomac'*). Les gloses sont ensuite transformées en patrons et projetées sur le corpus vulgarisé : les suites de mots appariées avec le patron sont considérées comme des paraphrases.

Pour générer la glose, l'analyseur DÉRIF procède en deux temps. Tout d'abord, une décomposition hiérarchique du mot morphologiquement complexe est produite et les processus morphologiques ayant contribué à la création du mot sont identifiés. Par exemple, $gastralgie$ est décomposé en [1] :

```
[ [ gastr N* ] [ algie N* ] NOM ]
```

et $prétraitement$ peut être décomposé de deux manières :

```
(1) [ pré [[ traiter VERBE] ment NOM] NOM]
(2) [ [ pré [traiter VERBE ] VERBE] ment NOM]
```

Puis, se basant sur l'hypothèse compositionnelle [2], DÉRIF génère une pseudo-définition (ou glose) à partir des résultats de l'analyse des mots morphologiquement complexes. Ainsi, $gastralgie$ est glosé en *'douleur (du – liée au) estomac'* et $prétraitement$ peut être glosé en *'(Période – Lieu) qui précède le traitement'* (décomposition 1) ou en *'(Action - résultat de l'action) de prétraiter'* (décomposition 2). Dans le cas de $gastralgie$, DÉRIF est capable d'établir un lien entre une racine classique et son équivalent lexical en français : $gastr$ a été relié à $estomac$ et $algie$ à $douleur$. DÉRIF traite à la fois les composés savants comme $gastralgie$ mais aussi les mots obtenus par l'affixation d'une racine classique comme $hépatique$ qui est glosé en *'en relation avec le foie'*.

En recherche d'information crosslingue, Schulz *et al.* (2006) ont mis au point le système MORPHOSAURUS basé non pas sur les mots mais sur ce que Schulz *et al.* appellent des « *subwords* » et qu'ils définissent comme « *la plus petite unité porteuse de sens dans un terme appartenant à un domaine de connaissance* » [3] - en somme, des morphèmes spécialisés. Le

1. Les décompositions ont été obtenues sur la version en ligne de DÉRIF : http ://www.cnrtl.fr/outils/DeriF/requete.php.

2. « *Les théories en morphologie lexicale permettent de déduire la définition d'un mot morphologiquement complexe en fonction de celui de ses constituants. Donc, un système implémentant une telle approche théorique (comme DériF, cf. §4) est à même de calculer la pseudo-définition de mots inconnus à partir des procédés morphologiques mis en œuvre.* » (*op. cit.*, p. 65)

3. « *the minmal meaning-bearing constituent of a domain-specific term* » (*op. cit.*, p. 1685)

système MORPHOSAURUS s'appuie sur un lexique multilingue dans lequel chaque *subword* est associé à un identifiant de sens, quelle que soit sa langue. Par exemple, *nephr-, ren-, kidney* et *riñon*, qui ont le sens de *'rein'*, sont tous associés au même identifiant. Schulz *et al.* distinguent plusieurs types de *subwords* selon la façon dont ils se combinent aux autres *subwords* :

- « ***Stems*** » **:** racines qui portent le contenu sémantique principal du mot et qui correspondent approximativement à des racines classiques ou à des mots indécomposables (*hepat, diaphys, head*)
- « ***Prefixes*** » **:** éléments placés avant une racine (*de-, re-*)
- « ***Proper prefixes*** » **:** préfixes qui ne peuvent être préfixés (*peri-, down-*)
- « ***Infixes*** » **:** éléments transitionnels (*o* dans *gastrointestinal*)
- « ***Suffixes*** » **:** éléments placés après une racine (*-a, -tomy*)
- « ***Proper suffixes*** » **:** suffixes qui ne peuvent être suffixés (terminaisons verbales comme *-ing, -ieron*)
- « ***Invariants*** » **:** éléments qui correspondent à des mots et qui ne doivent pas être utilisés comme *subwords* auquel cas ils provoqueraient des ambiguïtés lors de la décomposition (*ion, gene*)

Certains suffixes dérivationnels et flexionnels comme *-ation, -s* ainsi que les auxiliaires et les verbes modaux ne sont pas pris en compte.

Nous observons que sur les trois travaux de recherche pré-cités, deux ont recours à des théories de la morphologie : Claveau et Kijak s'appuient sur Mel'čuk (2006) et Namer s'appuie sur Corbin (1987). Schulz *et al.*, quant à eux, manipulent des notions inspirées de la morphologie et qu'ils ont adaptées à leurs objectifs : les *subwords* correspondent dans leur globalité à des morphèmes à l'exception de la catégorie « *Invariants* » et du fait que certains suffixes et morphèmes grammaticaux sont ignorés par le système MORPHOSAURUS.

La littérature fait état de diverses analyses du fait morphologique, cette diversité de points de vue se traduit aussi par une grande variété terminologique :

- Mel'čuk (2006), par exemple, distingue les *morphèmes*, élément de sens (signifiant), des *morphes*, segments linguistiques qui sont une réalisation possible d'un morphème (signifié). Par exemple, le morphème 'PLURIEL' correspond aux morphes /-z/, /-s/, /-ɪz/, /-ən/, en anglais (comme dans *girls, bricks, boxes, oxen*).
- Martinet (1985) emploie uniquement le terme de *monème* qui correspond non pas à une unité de sens mais à un « *effet de sens correspondant à une différence formelle* » (*op. cit.*, p. 33). Les monèmes sont identifiés par commutation : le test de commutation consiste à remplacer un élément par un autre dans la chaîne parlée et à observer si ce remplacement provoque un changement de sens. Un monème correspond à la fois à la réalisation phonique et à l'effet de sens observé. En anglais, il y a donc un seul monème 'PLURIEL' de forme /-z/, /-s/, /-ɪz/ ou /-ən/.
- Enfin, la grammaire française traditionnelle (Riegel *et al.*, 2005) ne distingue qu'un élément, le *morphème* qui est à la fois unité de sens et de forme [4]. Un morphème peut présenter plusieurs variantes graphiques ou orales appelées *allomorphes* : par exemple, le radical du verbe *aller* présente quatre allomorphes : *all-, i-, v-, aill-*. Généralement,

4. « *le morphème est généralement considéré comme l'unité minimale porteuse de sens obtenue par segmentation des énoncés. Il s'agit donc d'un segment préconstruit associant une forme et un sens, mais qui ne peut plus se décomposer en segments de même type* » (*op. cit.*, p. 533)

c'est la variante la plus fréquente ou celle jugée la plus représentative qui est utilisée pour désigner le morphème. En anglais, il y a donc un morphème /-z/ dont le sens est 'PLURIEL' et qui possède trois variantes : /-s/, /-ɪz/ et /-ən/.

En ce qui concerne nos travaux, nous avons choisi d'adopter l'optique de Schulz *et al.* c'est-à-dire que les éléments sur lesquels nous nous basons pour effectuer la traduction peuvent parfois correspondre à des morphèmes et parfois s'en approcher sans pour autant y correspondre totalement. En ce sens, nous avons donc défini plusieurs catégories basées à la fois sur des critères linguistiques et graphiques. Nous avons choisi de nous focaliser sur des éléments avec un sens référentiel ou susceptibles de changer fortement le sens d'un mot.

4.1.2 Définitions

Unité polylexicale Toute unité composée de plusieurs mots lexicaux. Cette catégorie inclut les unités polylexicales "classiques" *cancer du sein, essai clinique* mais aussi toute sorte de syntagmes : *après la ménopause, agir sur le cancer, deux dimensions*....

Unité monolexicale Unité composée d'un seul mot : *cancer, sein, après, ménopause*.

Mot Les mots correspondent à une chaîne de caractères composée d'au moins une lettre et éventuellement des chiffres ou des traits d'union, ex : *anti-p21, bio-rad, cancer, vasomoteur, gastrique*... Ces critères graphiques font que nous ne comptons pas comme mots les catégories linguistiques comme les locutions ou encore les mots composés dont les composants sont séparés par des espaces (*chemin de fer*). Ces éléments entrent dans la catégorie des unités polylexicales. Les mots sont des unités *autonomes* : ils peuvent apparaître de façon isolée dans les textes, i.e. entourés de caractères autres que un chiffre, une lettre ou un trait d'union. Nous distinguons les *mots complexes* et les *mots simples*.

Mot complexe Un mot complexe est un mot décomposable en plusieurs sous-éléments appelés morphèmes, ex. : *anti-p21* → *anti+p21, bio-rads* → *bio+rads, vasomoteur* → *vaso+moteur*. Les mots complexes sont les mots traduits par notre méthode. Nous ne traitons que les mots complexes qui sont des mots lexicaux, c'est-à-dire qu'ils appartiennent aux catégories grammaticales NOM, VERBE, ADJECTIF et ADVERBE.

Morphème Un morphème est un élément qui n'est pas décomposable en sous-éléments. Ce sont approximativement les morphèmes tels que définis dans la tradition grammaticale, puisqu'ils sont indécomposables et correspondent à des éléments de sens : *anti- 'inverse de'*, *-bio- 'relatif à la vie, au vivant'*, *-vaso- 'canal, vaisseau'*, *moteur 'qui produit ou transmet le mouvement'*, *cancer*, *gastrique*. Nous faisons la distinction entre morphèmes libres (mots simples) et liés.

Mot simple ou morphème libre Un mot simple est un mot non décomposable en sous-éléments : *cancer, gastrique, p21, rad, moteur*. Dans nos analyses, nous ne retenons que les mots lexicaux. En tant que mot, il peut fonctionner de façon autonome dans les textes. Cette propriété le distingue des *morphèmes liés*. Un mot simple peut être combiné à d'autres morphèmes pour créer un mot complexe (*moteur* dans *vasomoteur*, *p21* dans *anti-p21*).

Morphème lié Un morphème lié est, comme le mot simple, indécomposable en sous-éléments. À l'inverse du mot simple, il ne peut apparaître de façon autonome dans les textes, ex. : *anti- 'contre'*, *-bio- 'relatif à la vie, au vivant'*, *-vaso- 'canal, vaisseau'*. Ils correspondent aux catégories linguistiques suivantes : ***confixes, préfixes, suffixes***. Nous éliminons donc d'autres catégories comme les infixes (*um* dans *k<u>um</u>'ain*, *'en parlant de celui qui mange'*, tagalog), circonfixes (*ge...t* dans *<u>ge</u>sag<u>t</u> 'dit'*, allemand) ou interfixes (*o* dans *gastr<u>o</u>intestinal*).

Préfixe Un préfixe est un morphème lié toujours placé en position initiale d'un mot complexe, ex : *anti-* dans *<u>anti</u>-p21*. Dans les langues étudiées, plusieurs préfixes peuvent s'agglutiner en début de mot (*anti-réélection*) mais nous ne prenons pas en compte cette possibilité [5].

Confixe Les confixes correspondent aux racines grecques et latines entrant en jeu dans la création des composés savants. On les retrouve sous divers noms dans la littérature : bases supplétives, archéoconstituants, composants néolatins, bases savantes, primitifs supplétifs... (Namer, 2003, citée par Grabar (2004)). Le terme de *confixe*, quant à lui, est emprunté à Martinet (1979). Les travaux linguistiques (Bauer, 1983; Martinet, 1979; Riegel *et al.*, 2005) ont soulevé le fait que certains confixes apparaissent toujours en position initiale (*poly-*, *mono-*), d'autres toujours en position finale (*-cide*, *-vore*), et d'autres encore apparaissent indifféremment dans les deux positions(*-graph-*, *-phil-*). Nous ne faisons pas cette distinction : dans nos analyses, les confixes peuvent apparaître dans n'importe quelle position [6]. Toutefois, comme la frontière entre langue de spécialité et langue générale est poreuse, nous avons également été souples dans nos catégories. Certains éléments d'origine grecque ou latine que l'on rencontre en position initiale mais dont l'usage est courant en langue générale ont été considérés comme préfixes (ex : *multi-, poly-*). D'autres éléments ont été affectés aux deux catégories (ex : *méta-/-méta-, micro-/-micro-*).

Comme nous n'analysons pas les interfixes [7], beaucoup des confixes considérés intègrent directement l'interfixe : nous considérons par exemple *-chondri-, -chondr-* et *-chondro-* comme trois confixes alors que d'un point de vue linguistique, il s'agit du confixe *-chondr-* interfixé avec *-i-* ou *-o-* ou sans interfixe.

Les confixes peuvent être combinés entre eux et/ou avec un mot simple pour former un mot complexe. À cette base lexicale, peuvent venir s'accoler préfixes et suffixes.

Suffixe Un suffixe est un morphème lié toujours placé en position finale d'un mot complexe, ex : *-ment* dans *histologique<u>ment</u>*. Dans les langues étudiées (français, anglais, allemand), les suffixes sont principalement utilisés pour la dérivation (*incorporer → incorporation 'action d'incorporer'*) et la flexion (*incorporation → incorporations*). Les préfixes flexionnels ne sont pas pris en compte puisque nous travaillons avec les formes lemmatisées des mots où les phénomènes de flexion sont neutralisés. Concernant la dérivation, nous nous sommes restreints à un petit jeu de suffixes. Ces suffixes ont été choisis car nous les considérons comme fortement susceptibles d'être traduits par un morphème libre en langue cible, phénomène sur lequel nous souhaitions nous pencher en priorité. Il s'agit des suffixes *-ability 'capacité'*, *-able*

5. Afin de simplifier l'écriture de l'algorithme de découpage morphologique, nous n'avons pas implanté de règles analysant les mots comprenant plusieurs préfixes (ce cas n'apparaît pas dans nos données).

6. Contrairement à Harastani *et al.* (2012) qui distinguent « *Initial Combining Forms (ICFs)* [et] *Final Combining Forms (FCFs)* » (*op. cit.*, p. 74).

7. Contrairement à Weller *et al.* (2011) qui découpent *Kalorimetrie* en *Kalor, i* et *metrie*

'capable', *-hood 'état'*, *-like 'similaire'*, *-ly 'manière'*, *-wise 'sens'*. Par conséquent, la majorité des suffixes n'est pas prise en compte. Par exemple, un mot comme *gastrique* ne sera pas découpé en *gastr* et *ique* alors que d'un point de vue purement linguistique, il est bien composé de deux morphèmes.

Notations

L'occurrence d'un mot ou d'un ensemble de mots dans un texte est notée entre guillemets, ex. : « *breathless* ». Le signifié ou la traduction (glose) d'un mot est noté entre apostrophes, ex. : *'sans souffle'*.

Les préfixes sont notés suivis d'un tiret, ex. : *anti-* ; les confixes sont notés entourés de tiret, ex. : *-gastr-* ; les suffixes sont notés précédés d'un tiret ; ex. : *-able*, les mots n'ont aucun tiret, ex. : *cancer*.

Dans un mot complexe, les frontières entre morphèmes sont notées par le signe plus, ex. : *-gastro-+intestinal, anti-+p21, -histo-+logique+-ment*. Dans une unité polylexicale, les frontières de mots sont notées par un espace ex. : *contre le p21*, *manière -histo-+logique*.

La traduction est notée par une flèche allant du terme source vers le terme cible : *-histo-+logical+-ly* → *-histo-+logique+-ment*.

4.1.3 Hypothèses sous-jacentes

La traduction morpho-compositionnelle s'appuie sur les hypothèses suivantes :

Sens compositionnel Nous faisons l'hypothèse que les mots complexes ont un sens compositionnel calculable à partir du sens des morphèmes qui les composent : *anti-+tarte* à le sens de *'contre le tartre'*.

C'est une hypothèse relativement réaliste. C'est par exemple celle sur laquelle s'appuie l'analyseur DÉRIF. Il existe des contre-exemples comme dans le cas où le sens d'un mot est imagé (*rose des vents*). Toutefois, ces contre-exemples se retrouvent surtout dans la langue générale et c'est une hypothèse raisonnable que de considérer que dans la langue technique, le recours à la métaphore ou à des sens imagés est peu courant.

Traduction compositionnelle Nous faisons l'hypothèse d'un parallélisme entre les langues : si un terme source a un sens compositionnel, alors sa traduction a aussi un sens compositionnel, qui plus est, il y a bijection entre les morphèmes sources et les morphèmes cibles : $anti\text{-}_1\text{-}+abortion_2$ → $anti\text{-}_1+avortement_2$, $contre_1$ *(l')* $avortement_2$.

Ceci semble une hypothèse réaliste dans le cas des composés savants et des préfixés puisque comme l'ont observé Namer et Baud (2007) et Cartoni (2009b), ces processus de création lexicale sont relativement similaires pour les langues d'Europe de l'Ouest comme l'anglais, l'allemand, l'espagnol, le français et l'italien. Par contre, ceci est peut-être moins évident dans le cas des composés populaires (voir par exemple les travaux de Garera et Yarowsky (2008) qui justement traitent ces cas de non-correspondance) et les suffixés.

Fertilité La combinaison des morphèmes en mots est propre à chaque langue. Un terme source correspondant à un seul mot peut donc être traduit par un terme cible composé de

plusieurs mots : *minefield* → *champ (de) mines, cytotoxic* → *toxique (pour les) cellules*.

Distortion Nous considérons que l'ordre des morphèmes n'est pas forcément conservé d'une langue à l'autre : *pathophysiological* → *physiopathologique, tumor-margin* → *marge tumorale*. Cette hypothèse nous distingue de Harastani *et al.* (2012) et de Cartoni (2009b).

Divergence lexicale Le morphème cible peut ne pas être la traduction exacte du morphème source : *information-giving* → *offrir des informations, post-conception* → *après-fécondation*.

Variation morphologique La catégorie grammaticale d'un mot n'est pas forcément conservée lors de la traduction : *antitumor* (NOM) → *antitumoral* (ADJECTIF) ; *post-operative* (ADJECTIF) → *après (l') intervention* (SYNTAGME PRÉPOSITIONNEL).

4.1.4 Intérêt de l'approche pour l'exploitation des corpus comparables et la traduction spécialisée

La prise en compte des phénomènes de fertilité, distorsion, divergence lexicale et variation morphologique aide à tirer parti des spécificités des corpus comparables. La partie cible des corpus comparables étant constituée de textes produits spontanément, ceux-ci ne sont pas influencés par un quelconque texte source et ne comportent donc pas de phénomènes de calque. En conséquence, les termes cibles présents dans le corpus ont bien moins de chances d'avoir une structure similaire à celle du terme source, d'où l'intérêt d'essayer de générer des traductions dont la structure morpho-syntaxique et lexicale est relativement éloignée de celle du terme source. Parfois même, seule la génération d'une variante permettra de retrouver la traduction : *anthracycline-containing* ne peut être traduit que par une variante fertile (*contenant de l'anthracycline* et non **anthracycline-contenant*). De plus, nous savons que les paires de traduction sont bien moins fréquentes dans les corpus comparables que dans les corpus parallèles. Même s'il existe en langue cible une traduction de structure identique à celle du terme source, il est possible que cette traduction canonique ne soit pas présente dans le corpus cible. Identifier des variantes est donc aussi un moyen d'augmenter le nombre de paires de traduction extraites sans trop pénaliser la qualité du lexique final.

Enfin, les variantes, en particulier les variantes fertiles et les variantes morphologiques, sont particulièrement utiles pour la traduction spécialisée. D'une part, en fonction de la structure de la phrase dans laquelle le traducteur insèrera sa traduction, il peut être plus idiomatique d'utiliser une variante fertile ou morphologique. Ensuite, il est fréquent qu'une variante fertile corresponde à une version vulgarisée d'un terme scientifique (*ovariectomy* → *ovariectomie* vs. *ablation des ovaires*), ce qui est utile lorsque le traducteur traduit des textes de vulgarisation scientifique. De plus, les variantes fertiles, qui s'apparentent à des paraphrases, aident le traducteur à comprendre le sens du terme source dans sa langue natale.

4.2 Problématiques abordées et contributions

La méthode proposée permet d'aborder différentes problématiques que sont la génération de traductions fertiles, la couverture de multiples modes de construction morphologique

et l'ordonnancement des traductions candidates. Nous détaillons ci-après chacune de ces problématiques en essayant de mettre en avant nos contributions.

4.2.1 Génération de traductions fertiles

La notion de fertilité a été introduite par Brown *et al.* (1990) dans leurs travaux sur l'alignement de mots. Dans ce cadre, la fertilité *d'un mot* est définie comme le nombre de mots cibles (français) avec lequel un mot source (anglais) est aligné :

> « *We call the number of French words that an English word produces in a given alignment its* fertility *in that alignment* » (*op. cit.*, p. 82)

Dans leur article de 1993, Brown *et al.* présentent cinq modèles d'alignement de mots. Le premier modèle acquiert uniquement des probabilités de transfert, c'est-à-dire que pour pouvoir traduire d'une langue source F vers une langue cible E, on va estimer pour chaque mot source f et chaque mot cible e, la probabilité que f soit la traduction du mot e. Cette probabilité est notée $p(f|e)$. Le second modèle estime en plus des probabilités de distorsion, c'est-à-dire $d(i|j)$: la probabilité qu'un mot cible en position i corresponde au mot source en position j. La notion de fertilité apparaît dans le troisième modèle (IBM3). En plus des probabilités de transfert et de distorsion, ce modèle estime également des probabilités de fertilité, c'est-à-dire pour chaque mot source e, le modèle estime $n(\phi|e)$ la probabilité qu'un ensemble de ϕ mot(s) cible(s) soit la traduction de e [8].

Chez Brown *et al.*, la fertilité est donc une valeur numérique indiquant le nombre de mots cibles par lesquels un mot source peut potentiellement être traduit. Dans le tableau 4.1, on voit qu'il y a une probabilité de 0,342 pour le mot anglais $nodding$ *'hocher la tête'* soit traduit par un ensemble de 4 mots français, une probabilité de 0,293 pour qu'il soit traduit par un ensemble de 3 mots français, etc.

$e =$*nodding*

ϕ	$n(\phi\|e)$
4	0,342
3	0,293
2	0,167
1	0,163
0	0,023

TABLE 4.1 – Probabilités de fertilité de *nodding* - adapté de Brown *et al.* (1993, p. 286)

Dans leurs travaux sur l'acquisition de traductions de termes complexes, Daille et Morin (2005) et Robitaille *et al.* (2006) décrivent la fertilité comme le cas où terme source et terme cible sont de longueurs différentes, la longueur d'un terme complexe correspondant au nombre de mots pleins qui le compose [9] :

> « *fertility : source and target MWTs* [Multi-Word Terms] *can be of different lengths* » (*op. cit.*, p. 228)

L'exemple donné par Robitaille *et al.* est celui de traduction du terme français $table\ de\ vérité$ par le terme japonais $shinri{\cdot}chi{\cdot}hyo$ *'vérité valeur table'* où le composant chi *'valeur'* n'a pas d'équivalent dans le terme source.

8. « *The number of French* [target] *words to which* e *is connected in a randomly selected alignment is a random variable,* ϕ_e*, that we call the fertility of* e*.* »(Brown *et al.*, 1993, p. 275)

9. « *we define the length of a MWT as the number of content words it contains* » (*op. cit.*, p. 227)

Dans nos travaux, nous nous sommes intéressés au cas où la langue cible utilise plus de mots lexicaux que la langue source pour exprimer une même idée. C'est dans ce cas précis que nous parlons de *traductions fertiles* :

Traduction fertile *Soit deux ensembles disjoints S et C où S est un ensemble de termes sources et C est un ensemble de termes cibles. Soit la relation de traduction $T \subseteq S \times C$ et la fonction $l(x)$ indiquant le nombre de mots lexicaux du terme x. L'ensemble des traductions fertiles F est défini comme* $\{(s,c)|(s,c) \in T \text{ et } l(c) > l(s)\}$.

La relation de fertilité F est antisymétrique : $sFc \Rightarrow \neg cFs$. Par exemple, l'équivalence de traduction entre $table\ de\ vérité$ (2 mots lexicaux) et $shinri{\cdot}chi{\cdot}hyo$ (3 mots lexicaux) est une traduction fertile lorsque l'on traduit du français vers le japonais mais pas lorsque l'on traduit du japonais vers le français. La traduction de $growth\ rate$ en $taux\ de\ croissance$ n'est pas considérée comme fertile car le nombre de mots lexicaux reste le même. À noter que nous avons restreint notre champ d'étude aux cas où $l(s) = 1$ et $l(c) \geq 1$ c'est-à-dire la traduction d'une unité monolexicale vers une unité mono- ou poly- lexicale ($post\text{-}menopause \rightarrow post\text{-}ménopause,\ après\ (la)\ ménopause$).

Lors d'une étude des cas de traductions fertiles rencontrés dans la base de données terminologiques TERMIUM [10], nous avons dégagé deux types de fertilité : *fertilité sémantique* et *fertilité de surface*.

4.2.1.1 Fertilité sémantique

Dans le cas de la fertilité sémantique, langue source et langue cible opèrent un découpage sémantique différent. Le terme cible contient plus de morphèmes que le terme source. Ces morphèmes supplémentaires peuvent être à l'origine de mots supplémentaires dans le terme cible.

C'est le cas pour la traduction de $express\ option$ *'option express'* en $option\ voie\ rapide$ où le français rajoute la notion de *'voie'* alors qu'elle semble implicite en anglais. C'est aussi le cas pour la traduction de $snorkeling$ en $plongée\ avec\ tuba$ où l'anglais possède un morphème permettant de référer directement à la plongée avec tuba alors que le français, pour obtenir une expression au sens similaire, doit combiner le morphème $plongée$ avec le morphème $tuba$ à l'aide de la préposition $avec$.

Il n'est pas systématique qu'il y ait un nombre de mots supérieur en langue cible : $clavicule$, un mot composé d'un seul morphème se traduit par $collarbone$ en anglais, un mot composé de deux morphèmes (`collar+bone`, littéralement *'os du col'*). Le traitement de la fertilité sémantique n'est pas possible avec la méthode compositionnelle car cette dernière nécessite qu'il y ait bijection entre morphèmes du terme source et morphèmes du terme cible. De tels cas ne peuvent être traités que par les méthodes distributionnelles ou en passant par une langue pivot (Garera et Yarowsky, 2008) ou par le requêtage de documents non mixtes (Léon, 2008).

4.2.1.2 Fertilité de surface

Dans le cas de la fertilité de surface, langue source et langue cible opèrent un découpage sémantique identique, termes sources et cibles contiennent le même nombre de morphèmes et chaque morphème source a son équivalent cible (et inversement). Les langues varient seulement dans la façon dont elles combinent les morphèmes pour former des mots :

10. http ://www.termiumplus.gc.ca/tpv2alpha/alpha-fra.html - dernière consultation le 23/03/2011

Traduction fertile en surface *Soit une traduction fertile sFc, $M(s)$ l'ensemble des morphèmes du terme s, $M(c)$ l'ensemble des morphèmes du terme c et $\mathcal{T}$ une fonction de traduction. sFc est fertile en surface si pour tout $m_c \in M(c)$ il y a un unique $m_s \in M(s)$ tel que $\mathcal{T}(m_s) = m_c$.*

Un premier cas de fertilité de surface est celui où un mot composé est traduit par un syntagme, par exemple la traduction de $mouthwash$ en $bain\ de\ bouche$. Dans les deux langues, les morphèmes en jeu sont des morphèmes libres. La fertilité est causée par le fait que la langue cible choisit de créer un syntagme plutôt qu'agglutiner les mots pour créer un mot composé. La relation de complément entre les mots 'bain' et 'bouche' est donnée par l'ordre des mots en anglais alors qu'elle est indiquée par la préposition de en français.

Un second cas de fertilité de surface est celui où un morphème lié est traduit par un morphème libre, ex. : $unhindered \rightarrow sans\ entraves$. La fertilité est ici créée par le fait que le morphème lié un- est traduit en français par le morphème libre $sans$ qui est réalisé sous la forme d'un mot.

La traduction compositionnelle peut générer des traductions fertiles en surface. La traduction de mots composés par des syntagmes en langue cible a été traitée par Weller *et al.* (2011) pour les composés nominaux allemands. La variation morphème lié → morphème libre n'a pas été traitée dans le cadre de la traduction compositionnelle. Dans les approches empiriques, seuls Claveau et Kijak (2011) proposent d'apprendre à regrouper les variantes d'un même morphème qu'ils relient ensuite à leur traduction ($\{bactérie,\ bactério\text{-},\ bactéri\text{-}\} \rightarrow kin$) . Toutefois, ils ne se concentrent que sur les composés néoclassiques, ce qui laisse de côté des cas comme $unhindered \rightarrow sans\ entraves$.

Au delà des faits de fertilité, la méthode de génération de traduction proposée ne sera pas restreinte à un petit jeu de structures morphologiques comme c'est les cas pour les travaux de Cartoni (2009b) (uniquement mots préfixés), Harastani *et al.* (2012) (uniquement les composés néoclassiques) et Weller *et al.* (2011) (composés nominaux et néoclassiques). La section suivante détaille les structures morphologiques que nous envisageons de traiter.

4.2.2 Variété des structures morphologiques traduites

Contrairement aux autres approches compositionnelles basées sur un découpage en morphèmes ou en unités équivalentes, la méthode de traduction que nous proposons tente de ne pas se limiter à un petit jeu d'équivalences structurelles entre langue source et cible (préfixé → préfixé, composé populaire → syntagme, composé classique → composé classique).

Son entrée est un mot morphologiquement complexe. Ce mot peut avoir été construit par préfixation '*pretreatment*', composition savante '*densitometry*', suffixation '*childless*', composition populaire '*anastrozole-associated'* ou n'importe quelle combinaison de ces quatre modes de construction. Chaque traduction proposée en sortie est une liste de n mots qui peuvent être morphologiquement construits ou pas. Par exemple, *postoophorectomy* peut être traduit par *postovariectomie, après l'ovariectomie* ou *après l'ablation des ovaires*.

Nous décrivons ci-après ces quatre modes de formation et les illustrons à partir d'exemples de traductions anglais → français tirés de la base de données terminologiques TERMIUM[11] et du concordancier bilingue LINGUEE[12].

11. http ://www.termiumplus.gc.ca/
12. http ://www.linguee.com/

4.2.2.1 Composition populaire

Un mot construit par composition populaire résulte de la juxtaposition de deux ou plusieurs mots. Parfois, un trait d'union marque la frontière entre ces mots, parfois ils sont agglutinés [13]. Il est courant qu'avec le temps et l'usage, le sens compositionnel des mots composés ne soit plus transparent pour les locuteurs. Par exemple *saupoudrer* a été composé à partir de *sau-*, une variante de *sel* et de *poudrer* [14]. On parle alors de *démotivation*. Une des difficultés de la traduction des composés populaires est qu'ils sont généralement composés pour former une unité de sens nouvelle et peuvent ne pas avoir de sens compositionnel, par exemple un *chaise longue* n'est plus littéralement *'une chaise qui est longue'* mais plutôt un *'fauteuil'* (Riegel *et al.*, 2005, p. 547). Ces mots composés sont susceptibles d'être traduits de façon fertile lorsque la langue cible ne concatène pas les traductions des composants :

- *life+span* → *durée de vie*
- *word+coiner* → *forgeur de mots*
- *mal+entendant* → *hearing impaired*

4.2.2.2 Composition savante

La composition savante - que l'on retrouve aussi sous le nom de *composition néoclassique* (Namer et Baud, 2007), de *recomposition* (Riegel *et al.*, 2005) ou de *confixation* (Martinet, 1985) - est un processus similaire à la composition populaire si ce n'est que les composants sont des morphèmes liés correspondant à des éléments d'origine grecque ou latine (confixes). Riegel *et al.* (2005) admettent aussi dans les composés savants les mots formés par une racine classique et un mot du vocabulaire courant de la langue (ex. : *biomasse*).

Ce processus de création lexicale est couramment utilisé dans les domaines scientifiques et techniques, ils constituent à eux seuls près de la moitié des néologismes recensés dans les textes médicaux (Lovis *et al.*, 1998, cités par Namer (2005)). De plus, les règles de construction des composés savants sont extrêmement proches dans toutes les langues européennes, ce qui en facilite la traduction automatique (Iacobini, 2003, cité par Namer (2005)).

Il arrive que certains composés savants, initialement réservés à un domaine technique, glissent dans l'usage courant sous l'effet de la vulgarisation de certains concepts. Costaouec et Guérin (2007) citent le cas du confixe *-télé-*, dont le sens est *'à distance'*, que l'on retrouve dans des mots comme *télévision, téléscope, télépathie*. Une fois devenu courant dans la langue commune, certains confixes peuvent devenir libres comme dans le cas de *-bio-* par exemple. D'autres confixes, de part leur très grande productivité, deviennent des préfixes couramment employés : *anti-* dans *antidote, antigel* ; *archi-* dans *archifou, archiduc*. Cette porosité entre langue de spécialité et langue générale peut rendre difficile la classification de ces morphèmes empruntés aux langues classiques. Enfin, nous notons que les composés savants, en tant que base lexicale libre, sont soumis à la préfixation et à la suffixation (*bibliographiquement*, *anélectrolytique*).

Les confixés peuvent produire des traductions fertiles lorsqu'en langue cible le confixe est traduit soit par un mot dérivé de la même racine classique : *dermoreaction* → *réaction dermique*, soit par un mot de même sens mais appartenant à la langue générale : *ludothérapie*

13. Les théories linguistiques incluent également dans la classe des mots composés des groupes de mots qui font preuve d'une forte intégration (ex : impossible d'y insérer un autre élément : * *un clé verte à molette* vs. *une clé à molette verte*) et possèdent une autonomie syntaxique. Nous intéressant uniquement à la traduction d'unités monolexicales (i.e. composées d'un seul mot), nous laissons de côté ce type de mots composés.

14. http ://www.cnrtl.fr/etymologie/saupoudrer

$\rightarrow$ *play therapy*. Une traduction fertile peut être la variante vulgarisée d'un terme technique : *bunioniectomy* $\rightarrow$ *bunionectomie, ablation des oignons*.

4.2.2.3 Préfixation

La préfixation unit des morphèmes liés appelés *préfixes* à une base lexicale libre (mot simple ou confixé, qui peuvent éventuellement être suffixés). Généralement, le préfixé est toujours de même nature grammaticale que la base. Le préfixe est toujours placé avant la base lexicale et opère un changement sémantique :

- *pré-+adolescence* a le sens de *'avant l'adolescence'*
- *anti-+transpirant* a le sens de *'(qui lutte) contre la transpiration'*

La traduction fertile peut subvenir lorsque le préfixe est traduit par un mot de sens identique en langue cible : *unbleached* $\rightarrow$ *non blanchi* ou lorsqu'il existe sous une forme libre en langue cible : *nonabortive* $\rightarrow$ *non abortif*. Dans certains cas, la génération d'une variante fertile peut permettre de retrouver une paraphrase du terme source en langue cible : *pretreatment* $\rightarrow$ *avant le traitement*.

4.2.2.4 Suffixation

La suffixation unit des morphèmes liés appelés *suffixes* à une base lexicale libre (mot simple, confixé). Le suffixe est toujours placé après la base lexicale. Il est courant que la suffixation ait pour effet de changer la catégorie grammaticale, en plus d'opérer un changement sémantique :

- *physiological+-ly* a le sens de *'de façon, du point de vue physiologique'*
- *breath+-able* a le sens de *'qui peut se respirer'*

Comme pour la préfixation, la traduction fertile peut subvenir lorsque le préfixe est traduit par un mot de sens identique en langue cible : *wingless* $\rightarrow$ *sans ailes*. La génération de traductions fertiles peut aussi permettre de retrouver des paraphrases en langue cible : *extremely* $\rightarrow$ *de façon extrême*.

4.2.3 Ordonnancement des traductions candidates

Concernant l'ordonnancement des traductions candidates, nous avons soulevé le fait que les méthodes existantes étaient généralement fondées sur un seul critère (nombre d'attestations, similarité des contextes d'apparition...). Une des contributions de notre travail sera d'explorer de nouveaux critères d'ordonnancement que nous essaierons de combiner. Nous considèrerons des critères comme la fréquence du terme cible, la similarité entre son contexte et le contexte du terme source, la probabilité de traduction entre la partie du discours du terme source et le(s) partie(s) du discours du terme cible, etc. Parmi les méthodes de combinaison de critères, nous essaierons notamment des algorithmes d'apprentissage automatique et plus spécifiquement des algorithmes de *learning-to-rank* (Liu, 2011).

En traduction compositionnelle, seuls Baldwin et Tanaka (2004) combinent plusieurs critères pour ordonner les traductions. Pour cela, ils entrainent un classifieur SVM à partir d'exemples de traductions exactes et fausses. Chaque exemple correspond à un vecteur de traits listant divers critères associés au terme source, au terme cible ou à la traduction en elle-même et qui sont autant d'indices de la pertinence de la traduction. Un exemple est positif si le terme cible correspond à la traduction de référence ; négatif si le terme cible n'est pas la traduction de

référence. Le classifieur renvoie une valeur continue entre -1 et +1 qui est utilisée pour ordonner les traductions candidates.

Cette approche, basée sur la prédiction d'un score de "pertinence" ou de "qualité" à partir d'une paire `(terme source, terme cible)` est l'équivalent des approches *point-wise* utilisées en *learning-to-rank*.

Le *learning-to-rank* est un type d'apprentissage automatique majoritairement employé en recherche d'information où il sert à ordonner des documents par score de pertinence vis-à-vis d'une requête donnée. Le parallèle avec le problème de l'ordonnancement des traductions candidates est évident : nous cherchons à ordonner des termes cibles via un score de "pertinence de traduction" vis-à-vis d'un terme source donné.

À notre connaissance, les algorithmes de *learning-to-rank* sont encore très peu utilisés pour ordonner des traductions. Nous pouvons citer l'article de Sokolov *et al.* (2012) qui fait état de travaux en traduction automatique statistique visant à améliorer l'ordonnancement des différentes hypothèses de traduction. Leurs premiers résultats sont encourageants : ils montrent que leur modèle d'ordonnancement permet d'obtenir jusqu'à +0,4 de BLEU. Toutefois, leur travaux restent exploratoires : Sokolov *et al.* estiment que les conditions expérimentales (manque d'intégration au module de décodage, petit nombre de traits) ne permettent pas d'évaluer l'apport de leur approche à sa juste valeur.

Liu (2009) distingue trois sortes d'algorithmes de *learning-to-rank* :

- Les approches *point-wise* apprennent à associer à une paire `(requête, document)` soit :
  - une classe comme *pertinent, assez pertinent, pas pertinent*, l'ordonnancement est alors vu comme un problème de classification.
  - un score de pertinence, dans ce cas, l'ordonnancement est vu comme un problème de régression.
- Les approches *pair-wise* cherchent à apprendre des préférences : étant donné un triplet `(requête, document 1, document 2)`, elles indiquent si le document 1 est plus pertinent que le document 2 pour répondre à la requête. Pour les approches *pair-wise* et *point-wise*, l'ordre final des documents n'est pas donné directement : il est déduit à partir des catégories ou score de pertinence (*point-wise*) ou à partir des préférences (*pair-wise*).
- Les approches *list-wise*, quant à elles, apprennent directement la meilleure manière de permuter les documents répondant à une requête, de façon à ce que les documents soient ordonnés du plus au moins pertinent. Les données traitées sont directement de la forme `(requête, {document 1, document 2... document n})`.

La figure 4.1, empruntée à Liu (2009), donne un panorama historique des approches en *learning-to-rank*. D'après Liu (2007, 2009) et Cao *et al.* (2007), il semblerait que les approches *list-wise*, qui sont les plus récentes, soient aussi les plus performantes, du moins pour la recherche d'information. Liu (2009), par exemple, a testé plusieurs algorithmes de *learning-to-rank* sur des jeux de données de référence en recherche d'information. Les mesures d'évaluation utilisées sont la MAP et la NDCG[15]. Les diverses expériences montrent que les approches *list-wise* donnent de meilleurs résultats quelle que soit la mesure d'évaluation (NDCG au rang 1, 3 et 10 et MAP).

Dans la suite de notre travail, nous nous appuierons donc sur les approches *list-wise* pour tenter d'ordonner les traductions générées.

15. MAP et NDCG sont présentées en annexe page 197.

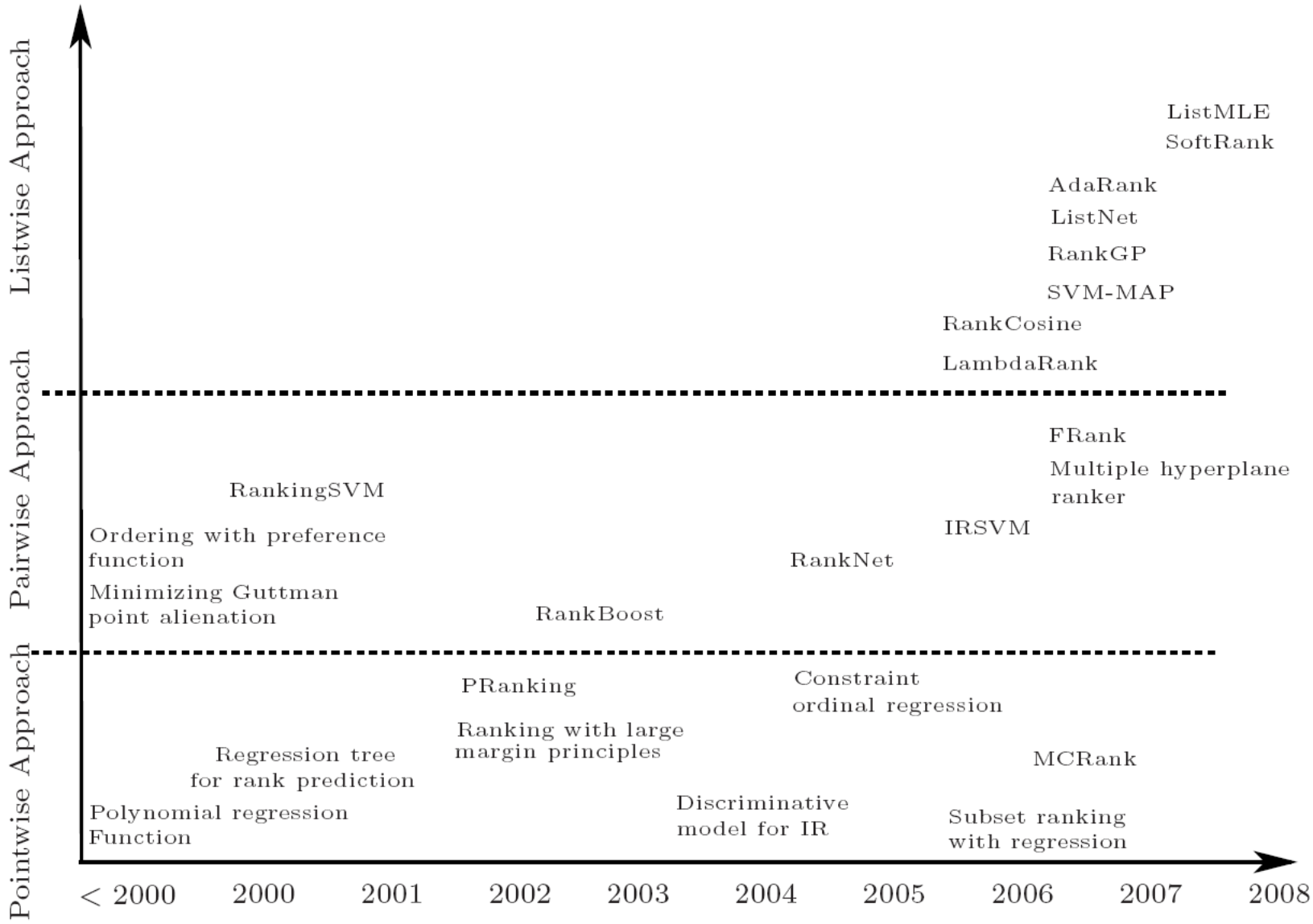


FIGURE 4.1 – Vue historique des approches en *learning-to-rank*- emprunté à (Liu, 2009)

4.3 Méthodologie d'évaluation

L'évaluation de la génération de traduction sera faite à partir de deux références : une référence *a priori* et une référence *a posteriori*. Cette distinction a été introduite par Ozdowska (2006) dans le cadre de l'alignement de mots dans les corpus parallèles :

> « *Globalement, on peut distinguer deux manières de procéder pour constituer les données de référence. La première consiste à annoter manuellement un échantillon de données indépendamment des sorties fournies par un quelconque système d'alignement* [...]. *Le résultat est une référence que l'on qualifiera d'*a priori. *La seconde consiste à juger directement les sorties fournies par le système que l'on cherche à évaluer* [...]. *On parlera de référence* a posteriori. » (*op. cit.*, p. 35)

Notre méthode sera expérimentée sur le corpus médical CANCER DU SEIN décrit en 2.2.2.1 et 5.1. Pour les deux références, nous partirons d'un même ensemble de termes sources (S) dont la traduction n'est pas donnée par notre dictionnaire bilingue généraliste. La référence *a priori* (R) sera construite par projection du méta-thésaurus médical UMLS dans notre corpus. La référence *a posteriori* (P) sera construite par l'annotation manuelle des traductions proposées par notre système. Nous utiliserons ces sorties annotées comme exemples pour l'apprentissage des modèles d'ordonnancement (T) desquelles nous soustrairons les traductions appartenant à la référence *a priori* et qui seront utilisés pour l'évaluation de l'ordonnancement (E). L'ordonnancement sera donc uniquement évalué sur des traductions appartenant à la référence *a priori* et pour lesquelles le système a pu proposer au moins une traduction.

En résumé, nous manipulons cinq ensembles, illustrés par la figure 4.2 :

Évaluation de la génération de traductions
- S correspond aux termes sources à traduire
- R correspond à la référence *a priori*
- P correspond à la référence *a posteriori*, $R \cap P \neq \emptyset$

Évaluation de l'ordonnancement des traductions
- T correspond aux données d'apprentissage du modèle d'ordonnancement : $T = P \setminus R$
- E correspond aux données d'évaluation du modèle d'ordonnancement : $E = P \cap R$

La principale différence entre référence *a priori* et référence *a posteriori* est que les termes sources présents dans la référence *a priori* ont tous une traduction identifiée dans le corpus cible alors que pour la référence *a posteriori*, les termes sources seront simplement des termes rencontrés dans les textes sources et dont la traduction n'est pas donnée par notre dictionnaire bilingue généraliste : nous ne savons pas si leur traduction est présente dans les textes cibles.

4.3.1 Référence *a priori*

L'intérêt principal des références *a priori* est que ces dernières sont souvent construites à partir de ressources linguistiques reconnues et fréquemment employées au sein d'un domaine de recherche : elles permettent donc de se comparer aux autres systèmes. Un second intérêt de la référence *a priori* est qu'elle permet d'évaluer le rappel et d'analyser les cas de silence, ce qui ne peut se faire avec une référence *a posteriori* (les termes sources n'ont pas forcément d'équivalent dans le corpus cible).

Toutefois, ce type de référence n'offre qu'une vision parcellaire des performances d'un système. C'est encore plus vrai dans le domaine de la traduction où l'on sait bien qu'il est

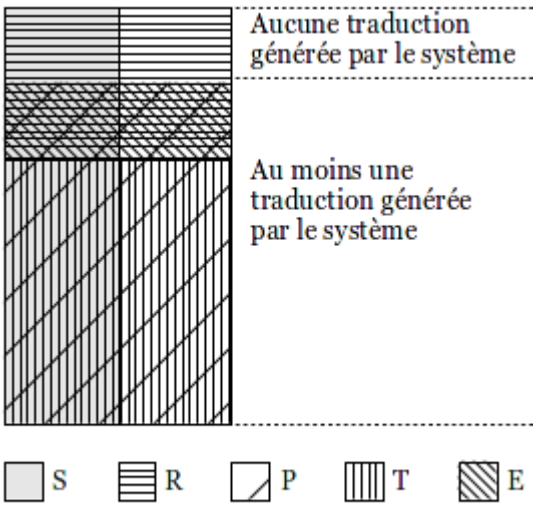


FIGURE 4.2 – Ensembles de données utilisés pour l'évaluation

difficile de lister l'ensemble des traductions possibles d'un terme et que l'utilité d'une traduction n'est évaluable que par rapport à un contexte applicatif donné. Un exemple parlant est celui de l'évaluation menée par Claveau et Kijak (2011) : lorsqu'ils se basent uniquement sur les traductions présentes dans l'UMLS, ils obtiennent une précision de 63 % ; par contre, s'ils comptent comme correctes des traductions non présentes dans l'UMLS mais attestées dans des dictionnaires médicaux en ligne, ils obtiennent 89 % de précision. Nous avons observé exactement le même phénomène sur nos sorties : bien que les traductions de l'UMLS soient parfois très libres, nombre de traductions correctes générées par notre système n'étaient pas répertoriées dans l'UMLS.

L'évaluation *a priori* donne également une image biaisée des lexiques qui seront extraits une fois l'algorithme implanté dans un logiciel de TAO :

- D'une part, le fait de baser l'évaluation uniquement sur des cas où la traduction se trouve effectivement dans le corpus cible aide l'algorithme. Il existe beaucoup plus de traductions fausses dans la référence *a posteriori* puisque pour certains termes, même si une traduction correcte est générée, celle-ci ne sera jamais sélectionnée puis qu'elle n'apparaît pas dans le corpus (toute autre traduction générée et trouvée dans le corpus sera automatiquement fausse).
- D'autre part, la référence n'est pas exhaustive : en proposant une référence rigide, cela éclipse des traductions générées qui seraient correctes bien que non présentes dans la référence.

4.3.2 Référence *a posteriori*

La référence *a posteriori* a pour principal avantage le fait de donner une meilleure estimation des performances que l'on peut attendre du système en situation d'utilisation. Elle permet de répondre à des questions telles que : Combien de termes sources reçoivent effectivement une traduction ? Quelle est la qualité de ces traductions ? Au final, combien de termes sources peuvent espérer recevoir au moins une traduction correcte ?

Les unités linguistiques à traduire ne sont pas des *termes* du domaine à proprement parler [16] mais simplement des mots présents dans le corpus source et dont la traduction n'existe pas dans le dictionnaire bilingue généraliste (ou dont la traduction proposée par le dictionnaire n'est pas dans le corpus cible). L'évaluation *a posteriori* est faite sous l'angle de l'enrichissement de ressources bilingues généralistes à l'aide de traductions spécialisées issues de corpus comparables, ce qui correspond plus à notre cadre applicatif, alors que la référence *a priori* semble plus adaptée à l'évaluation d'outils d'extraction de terminologies bilingues.

Or, comme nous l'avons expliqué dans la section 1.1.3, les besoins du traducteur ne se réduisent pas aux seuls termes techniques mais bien à toute formulation dont il ne connait pas la traduction. Par exemple, $aromatherapy$ peut être considéré comme un terme du domaine médical ; pour autant, ce terme est aussi fréquemment employé dans la langue courante et il serait surprenant qu'un traducteur ne sache pas le traduire. *A contrario*, l'expression $patient\text{-}centred$ n'aurait probablement pas sa place dans une terminologie [17] mais sa traduction peut poser problème : les mots composés de structure NOM-PARTICIPE PASSÉ n'existe pas en français, il faudra donc forcément s'éloigner de la langue source et s'assurer que la traduction proposée correspond bien aux usages linguistiques du domaine.

4.4 Synthèse

Ce chapitre nous a permis de poser le cadre méthodologique de nos recherches. Nous avons présenté les fondements de notre méthode de traduction et nous nous sommes positionnés par rapport aux travaux similaires. Nous avons détaillé les problématiques abordées en mettant en avant nos contributions. Enfin, nous avons exposé et argumenté notre méthodologie d'évaluation. Le chapitre suivant présente les données sur lesquelles nous avons expérimenté notre méthode de traduction.

16. Au sens de dénomination d'un concept du domaine.
17. Elle n'apparaît ni dans l'UMLS ni dans TERMIUM.

Chapitre 5

Données expérimentales

Sommaire

Introduction

Dans ce chapitre, nous présentons les données utilisées pour expérimenter la méthode de traduction présentée dans le chapitre précédent. Nous avons appliqué notre méthode à la traduction de l'anglais vers le français et l'allemand. Nous présentons dans un premier temps les corpus comparables employés (5.1), puis les termes sources sélectionnés pour tester la méthode de traduction (5.2) ainsi que les données utilisées pour l'évaluation de la génération de traduction (5.3) et pour l'apprentissage et l'évaluation du modèle d'ordonnancement (5.4). Enfin, nous présentons les ressources linguistiques sur lesquelles s'appuie notre système de traduction (5.5).

Des extraits des données sont consultables dans l'annexe B. Le corpus allemand, la table de traduction des morphèmes et les lexiques de référence sont librement disponibles sous licence LGPL-LR [1].

5.1 Corpus comparables

Nous avons utilisé des textes spécialisés issus du domaine médical et traitant du cancer du sein. À l'instar de Bowker et Pearson (2002), nous définissons un texte spécialisé comme un texte produit par un expert du domaine et à destination d'autres experts (discours scientifique) ou du grand public (discours vulgarisé).

Les textes ont été collectés manuellement dans des portails scientifiques et des sites Internet d'information à destination des patientes atteintes de cancer du sein et de leurs proches. Les textes en anglais proviennent tous de sources britanniques, les textes en français proviennent de sources françaises et les textes allemands de sources allemandes. Les textes anglais et français traitent tous du cancer du sein chez la femme (les textes traitant du cancer du sein chez l'homme ont été supprimés). Nous avons rencontré beaucoup plus d'articles vulgarisés sur le cancer du sein en anglais qu'en français (il existe de nombreuses associations au Royaume-Uni). Ceci explique le déséquilibre entre le corpus scientifique et le corpus vulgarisé en français (1,45 fois plus de mots pour le corpus scientifique, cf. tableau 5.1). En ce qui concerne le corpus allemand, il a été très dur de trouver des documents, en particulier pour les articles scientifiques. Pour la partie scientifique, nous avons principalement utilisé des résumés d'articles, ce qui explique le grand nombre de documents dans cette partie du corpus (tableau 5.2). Nous avons également eu recours à des textes scientifiques traitant du cancer du sein chez l'homme.

Une fois collectés et passés au format texte, les textes ont été segmentés en mots, lemmatisés et étiquetés avec le logiciel d'analyse linguistique XELDA [2].

	EN	FR	DE
scientifique	198 244 (48 %)	267 180 (59 %)	197 187 (49 %)
vulgarisé	218 336 (52 %)	184 504 (41 %)	201 760 (51 %)
TOTAL	416 580	451 684	398 947

TABLE 5.1 – Composition et taille du corpus en nombre d'occurrences

	EN	FR	DE
scientifique	70	78	103
vulgarisé	272	217	162
TOTAL	342	295	265

TABLE 5.2 – Composition et taille des corpus en nombre de documents

Nous avons également évalué la comparabilité des corpus (tableau 5.3) à l'aide de la mesure de Li et Gaussier (2010) présentée dans la section 1.2.4.1. Cette mesure indique l'espérance

1. http://www.lina.univ-nantes.fr/?Linguistic-resources-from-the,1676.html
2. http ://www.temis.com

de rencontrer la traduction d'un mot source dans le corpus cible (et inversement), par rapport à un dictionnaire bilingue donné. Elle est basée sur la projection d'un dictionnaire bilingue dans le corpus (nous avons utilisé le dictionnaire bilingue décrit en section 5.5.1).

	EN-FR	EN-DE
scientifique	0,71	0,42
vulgarisé	0,69	0,46
TOUT	0,74	0,45

TABLE 5.3 – Comparabilité des corpus étant donné le dictionnaire de l'analyseur XELDA

Le niveau de comparabilité des corpus est relativement faible. À titre de comparaison, Li et Gaussier (2010) ont mené des expériences avec des corpus d'une comparabilité variant entre 0,882 et 0,916. Le dictionnaire employé a été constitué à partir de diverses ressources en ligne et totalise 75 845 paires de traduction (notre dictionnaire en comprend entre 59 495 et 69 285 selon les paires de langues).

De plus, le corpus anglais-allemand est nettement moins comparable que le corpus anglais-français ce qui indique que nous aurons plus de difficultés à en extraire des paires de traduction. Cette moindre comparabilité peut s'expliquer par le fait qu'il a été difficile de trouver des documents en allemand, ce qui a pu conduire à ajouter des textes moins centraux au domaine et donc contenant du vocabulaire périphérique.

5.2 Termes sources

Les termes à traduire ont été extraits du corpus anglais de façon semi-supervisée en suivant le processus suivant :

(i) Nous avons établi une petite liste de morphèmes liés appartenant à la langue anglaise et extrait automatiquement du corpus anglais tous les mots qui contenaient ces morphèmes, ou du moins, dont une sous-chaîne de caractères correspondait à un morphème lié. Par exemple, nous avons extrait les mots $postchemotherapy$ et $poster$ parce qu'ils contenaient la chaîne $post$.

(ii) Ces mots extraits du corpus ont été triés manuellement : ceux qui n'étaient pas morphologiquement construits (erreurs lors de l'extraction comme pour $poster$) ont été éliminés, ceux qui étaient bien le résultat d'une construction ont été gardés et découpés en morphèmes : $postchemotherapy$ a été découpé en $post,\ chemo$ et $therapy$.

(iii) Si le découpage morphologique de la phase (ii) mettait au jour de nouveaux morphèmes liés, nous reprenions le processus en (i) où nous projetions les nouveaux morphèmes sur le corpus de façon à ramener de nouveaux mots construits.

Nous avons également ajouté les mots comportant des traits d'unions (ex. $ER\text{-}positive$) ainsi que les mots correspondant à la concaténation de deux ou plusieurs mots (ex. $mouthwash$). Ces mots composés ont été validés manuellement. De cette façon, nous avons obtenu une liste de 2 025 mots anglais morphologiquement construits. De cette liste nous avons exclu, pour chaque paire de langue, tous les mots anglais qui avaient une traduction dans notre dictionnaire bilingue généraliste et dont la traduction était présente dans le corpus cible. Au final, nous avons donc une liste de 1 839 mots complexes à traduire en français (S^{FR}) et une liste de 1 824 mots à traduire vers l'allemand (S^{DE}).

81 % des termes sources correspondent à une composition populaire ou à une base populaire préfixée et/ou suffixée. Les termes avec un composant néoclassique ne forment que 19 % des termes sources (tableau 5.4). 58 % des termes sources comprennent au moins 1 morphème lié (préfixe ou confixe ou suffixe). La majorité des termes sources ne sont composés que de deux morphèmes, le nombre de morphèmes maximum est 4 (tableau 5.5).

Structure morphologique	S^{FR}	S^{DE}	exemple
composition populaire	42 %	42 %	*acute-phase*
préfixe + base populaire	28 %	28 %	*nondifferential*
composition savante	17 %	17 %	*oncogenesis*
base populaire + suffixe	9 %	8 %	*sleeveless*
préfixe + base populaire + suffixe	2 %	2 %	*abnormally*
base savante + suffixe	1 %	1 %	*chemotherapeutically*
préfixe + base savante	1 %	2 %	*pro-angiogenic*

TABLE 5.4 – Structures morphologiques des termes sources

	S^{FR}	S^{DE}
2 morphèmes	88 %	88 %
3 morphèmes	11 %	11 %
4 morphèmes	1 %	1 %

TABLE 5.5 – Taille des termes sources (nb. de morphèmes)

5.3 Données de référence pour l'évaluation de la génération de traduction

5.3.1 Référence *a priori*

Pour constituer notre référence *a priori*, nous avons utilisé l'UMLS (Lindberg *et al.*, 1993) qui est un méta-thésaurus du domaine médical. Ce méta-thésaurus rassemble des concepts dont les termes appartiennent à plus de soixante vocabulaires contrôlés et classifications utilisés en biomédecine (MeSH, SNOMEDCT, MeDRA...). L'édition 2000 du méta-thésaurus comptait environ 730 000 concepts et 1,5 million de termes dans 17 langues [3].

Ce méta-thésaurus est fréquemment employé comme référence pour la traduction automatique de termes, en particulier pour l'évaluation des méthodes empiriques puisque par la même occasion, il fournit aussi des exemples pour l'apprentissage (Langlais *et al.*, 2008, 2009; Claveau, 2009; Claveau et Kijak, 2011).

Pour constituer notre référence *a priori*, nous avons procédé ainsi :

(i) Extraction, pour chaque terme source s de S^{FR} et S^{DE}, des termes français, respectivement allemands, associés au même concept que s dans l'UMLS

(ii) Nettoyage manuel des termes cibles extraits

(iii) Recherche des termes cibles attestés dans le corpus cible

3. http ://www.nlm.nih.gov/mesh/umlsforelis.html - dernière consultation le 12/02/2012.

(i) Extraction des associations terme source → termes cibles de l'UMLS

Les termes anglais, ainsi que leurs traductions en français et allemand ont été extraits des fichiers MRCONSO.RRF.aa et MRCONSO.RRF.ab de l'UMLS. Le contenu de ces fichiers consiste en 18 colonnes séparées par une barre de disjonction (voir extrait 5.1) :

```
C0001418|ENG|P|L0001418|PF|S0010818|N|A7568548||C2852||NCI|PT|C2852|Adenocarcino
ma|0|N|256|
C0001418|ENG|S|L9068756|PF|S11315684|Y|A17310266|2839509015|443961001||SNOMEDCT|
PT|443961001|Malignant␣adenomatous␣neoplasm|9|N|2304|
C0001418|ENG|P|L0001418|VO|S0585881|Y|A0640520||||SNMI|PT|M-81403|Adenocarcinoma
␣NOS|9|N||
C0001418|ENG|S|L0680539|VO|S0940401|Y|A0996143||M0000355|D000230|MSH|PM|D000230|
Malignant␣Adenomas|0|N|256|
C0001418|ENG|S|L2775107|PF|S3623578|Y|A16362919|573026011|189582009||SNOMEDCT|OF
|189582009|[M]Adenocarcinoma␣NOS␣(morphologic␣abnormality)|9|O||
C0001418|FRE|P|L3246449|PF|S3773563|N|A11067791|||10001141|MDRFRE|LT|10001141|Ad
énocarcinome|3|N||
C0001418|FRE|S|L5705743|PF|S6538780|Y|A9155521||M0000355|D000230|MSHFRE|EN|D0002
30|Carcinome␣glandulaire|3|N||
C0001418|FRE|S|L6175020|PF|S7052488|Y|A11051669|||10001141|MDRFRE|LT|10001166|Ad
énocarcinome␣SAI|3|N||
C0001418|GER|P|L1229669|PF|S1471615|N|A10139059|||10001141|MDRGER|LT|10001141|Ad
enokarzinom|3|N||
C0001418|GER|S|L1226879|PF|S1468825|Y|A1419477||||WHOGER|PT|1289|ADENOCARCINOM␣N
NB|2|N||
```

Extrait 5.1 – Fichiers MRCONSO.RRF.* du méta-thésaurus UMLS

L'identifiant de concept est donné par la première colonne, la langue est donnée par la deuxième colonne et le terme est dans la 15ème colonne. Dans cet extrait, on voit que le concept C0001418 est associé aux termes suivants :

- **Anglais :** *Adenocarcinoma, Malignant adenomatous neoplasm*, *Malignant Adenomas, [M]Adenocarcinoma NOS (morphologic abnormality)*
- **Français :** *Adénocarcinome SAI, Carcinome glandulaire*
- **Allemand :** *Adenokarzinom, ADENOCARCINOM N NB*

Cette étape d'extraction a donné une liste de 261 termes anglais associés à 771 termes allemands et une liste de 259 termes anglais associés à 768 termes français.

(ii) Nettoyage manuel des termes cibles extraits

Une des difficultés d'utilisation de cette ressource est qu'il s'agit d'un thésaurus et non d'une ressource faite pour la traduction. Par conséquent, comme on le voit dans l'exemple, les termes associés à un concept sont des formes linguistiques assez variées qui correspondent bien au concept mais qui parfois ne sont pas des traductions exactes les unes des autres. Cette particularité fait que certains travaux n'utilisent pas directement les concepts de l'UMLS mais plutôt les liens établis au sein de chaque vocabulaire contrôlé dont est constitué l'UMLS. Par

exemple, Langlais *et al.* (2008) constituent trois ressources : une à partir des liens de traduction du MesH, une autre à partir de ceux du MedRA et une autre à partir du SNOMED CT.

Malgré cela, nous avons souhaité utiliser les identifiants de concept, justement parce que nous nous intéressons à des traductions non canoniques. Ce choix nous a permis d'obtenir des équivalences de traduction relativement libres. Par exemple, pour le terme anglais *discomfort 'inconfort'*, nous obtenons les termes allemands *unwohlsein 'malaise'* et *fuehlt sich nicht wohl 'se sent pas à l'aise'*. Pour *self-image 'image de soi'*, nous obtenons en français *perception de soi, autoperception*.

Toutefois, nous avons noté que certaines traductions extraites de l'UMLS avait un sens très éloigné du terme source. Par exemple, *epirubicin-vinorelbine* est associé à *épirubicine* alors que l'épirubicine et la vinorelbine sont deux médicaments différents, le premier étant un anthracycline (inhibe la synthèse de l'ARN et ADN) ; le second un anti-mitotique (inhibe la divison des cellules) [4]. Le terme *epirubicin-vinorelbine* est employé pour désigner un traitement combinant les deux médicaments [5]. Nous avons donc examiné tous les termes cibles proposés et éliminé ceux qui, après consultation des ressources adéquates [6], nous semblaient avoir un sens trop éloigné du terme source. En cas de doute, nous avons préféré laisser un terme cible potentiellement erroné plutôt que de l'enlever.

Nous avons également observé que l'UMLS donnait, pour certains termes cibles, la forme singulier et pluriel (*polyamine → polyamine, polyamines*) et parfois une forme indiquant la variante féminine du terme (*intolerant → intolérant(e)*). Ces formes ont aussi été corrigées et/ou éliminées manuellement.

Après vérification manuelle, notre liste d'équivalences contenait 261 termes anglais associés à 767 termes allemands (4 termes cibles ôtés) et 259 termes anglais associés à 732 termes français (36 termes cibles ôtés).

(iii) Recherche des termes cibles attestés dans le corpus cible

Dans cette troisième et dernière étape, nous avons recherché des attestations des termes cibles dans les textes en langue cible. Pour aider à l'appariement, les termes ont été normalisés, du côté des entrées de l'UMLS comme du côté des mots du corpus : casse en minuscule, tirets ôtés, lemmatisation. Tous les termes cibles non attestés dans le corpus cible ont été ôtés de la référence, ainsi que les termes sources pour lesquels aucun des termes cibles donnés par l'UMLS n'était attesté dans le corpus.

Nous avons ainsi obtenu une liste de 126 termes anglais alignés avec 163 termes français (R^{FR}) et 90 termes anglais alignés avec 104 termes allemands (R^{DE}).

Le tableau 5.6 récapitule les différentes étapes de la construction de la référence *a priori*.

4. source : http://en.wikipedia.org/wiki/Epirubicin et http://en.wikipedia.org/wiki/Vinorelbine
5. source : http://www.ncbi.nlm.nih.gov/pubmed/ ?term=epirubicin-vinorelbine
6. fiches encyclopédiques collaboratives avec liens interlangues : http://en.wikipedia.org/
concordancier bilingue : http://www.linguee.fr/
contextes monolingues anglais spécialisés : http://www.ncbi.nlm.nih.gov/pubmed/
dictionnaire médical anglais : http://medical-dictionary.thefreedictionary.com/

		EN→ FR	EN→ DE
(i)	traductions trouvées dans l'UMLS	261 → 771	259 → 768
(ii)	traductions après nettoyage manuel	261 → 767	259 → 732
(iii)	traductions attestées dans le corpus cible (R)	126 → 163	90 → 104

TABLE 5.6 – Étapes de la construction de la référence *a priori*

5.3.2 Référence *a posteriori*

Pour constituer la référence *a posteriori*, nous avons fait annoter les sorties du système par des traducteurs et des locuteurs [7]. Une partie des traductions a été annotée par plusieurs annotateurs de façon à calculer l'accord inter-annotateur (une sélection aléatoire de 100 traductions pour chaque paire de langue a été annotée par au moins deux annotateurs). La mesure d'accord utilisée est le $Kappa$ de Carletta (1996) (cf. annexe A.6). Nous obtenons un $Kappa$ de 0,71 pour l'anglais-français et 0,77 pour l'anglais-allemand. En cas de désaccord entre deux annotateurs, c'est l'annotation faite par l'annotateur le plus consensuel [8] qui a été retenue.

Quatre valeurs ont été utilisées pour l'annotation, chacune correspondant à un "coût" d'utilisation pour le traducteur : EXACT, ACCEPTABLE, PROCHE et FAUX. Le tableau 5.7 résume les critères d'annotation. Dans tous les cas, le terme cible peut être plus long que le terme source (traductions fertiles).

VALEUR	sens	opérations traducteur
EXACT	$=$	0 ou adaptation au type de discours
ACCEPTABLE	$=$	transformation morpho-syntaxique
PROCHE	$\simeq$	recherches complémentaires éventuelles
FAUX	$\neq$	autres recherches obligatoires

TABLE 5.7 – Valeurs pour l'annotation des traductions

EXACT

Le terme cible correspond à la traduction exacte du terme source, i.e. c'est le terme spécialisé "consacré" ou un équivalent vulgarisé.

- il y a **égalité sémantique** (ni ajout ni perte de sens) → le terme cible indique au traducteur le sens du terme source
- le traducteur peut utiliser le terme cible **tel quel**.

Exemples :

- $pathophysiological \rightarrow physiopathologique$
- $cardiotoxicity \rightarrow toxicité\ cardiaque,\ toxicité\ pour\ le\ cœur$
- $tumour\text{-}margin \rightarrow marge\ tumorale$

7. Les sorties en anglais ont été annotées par une étudiante (M2) de l'Institut Supérieur d'Interprétation et de Traduction (ISIT), l'auteure de la présente thèse et une troisième personne maîtrisant l'anglais. Les sorties en allemand ont été annotées par une traductrice et deux étudiants (M2) de l'ISIT.

8. C'est-à-dire celui ayant le meilleur $Kappa$.

ACCEPTABLE

Le terme cible correspond à un dérivé morpho-syntaxique de la traduction exacte :
- il y a **égalité sémantique** (ni ajout ni perte de sens) → le terme cible indique au traducteur le sens du terme source
- le traducteur devra obligatoirement effectuer une **transformation morpho-syntaxique** pour reconstituer la traduction exacte

Exemples :
- $dosimetry \rightarrow dosimétrique$ (traduction exacte : $dosimétrie$)
- $cytoprotection \rightarrow protéger\ les\ cellules$ (traduction exacte : $protection\ des\ cellules$, $cytoprotection$)
- $œstrogene\text{-}sensitive \rightarrow sensibilité\ à\ l'œstrogène$ (traduction exacte : $sensible\ à\ l'œstrogène$)

PROCHE

Le terme cible ne correspond ni au terme consacré, ni à un équivalent vulgarisé ni à une variante morpho-syntaxique mais il reste utile pour la traduction :
- il y a **proximité sémantique** (intersection ou inclusion du sens) → le terme cible apporte au traducteur des éléments de compréhension du terme source
- le traducteur devra **éventuellement** faire des **recherches complémentaires** pour trouver la traduction du terme source

Exemples :
- $desirability \rightarrow désir$
- $high\text{-}dose \rightarrow dose\ un\ peu\ plus\ élevée$

FAUX

Le terme cible est inutile pour la traduction :
- il n'y a **aucune proximité sémantique** → le terme cible ne fournit pas d'éléments de compréhension du terme source au traducteur
- le traducteur devra trouver la traduction par un **autre moyen**

Exemples :
- $immunoscore \rightarrow immunomarquer$
- $risk\text{-}reducing \rightarrow risque\ de\ réduction$

Au final, P^{FR}, la référence *a posteriori* pour l'anglais-français, est constituée de 730 termes sources anglais associés à 2 129 traductions candidates annotées manuellement. P^{DE}, la référence *a posteriori* pour l'anglais-allemand, est constituée de 654 termes sources anglais associés à 2 016 traductions candidates annotées manuellement [9].

9. Les chiffres de 730, respectivement 654, correspondent aux nombres de termes sources pour lesquels le système a pu générer au moins une traduction candidate en français, respectivement en allemand.

5.4 Données pour l'apprentissage et l'évaluation du modèle d'ordonnancement

Nos données d'apprentissage (T) correspondent aux termes sources (et à leurs traductions candidates générées par le système et annotées) qui n'appartiennent pas à la référence *a priori* ($T = P \backslash R$). Elles contiennent 647 termes anglais associés à 1 970 traductions candidates en français et 588 termes anglais associés à 1 829 traductions candidates en allemand.

Nos données d'évaluation (E) correspondent aux termes sources (et à leurs traductions candidates générées par le système et annotées) qui appartiennent à la référence *a priori* ($E = P \cap R$). Elles contiennent 83 termes anglais associés à 159 traductions candidates en français et 66 termes anglais associés à 187 traductions candidates en allemand.

5.5 Ressources linguistiques

5.5.1 Dictionnaire bilingue généraliste

Le dictionnaire bilingue généraliste utilisé par le générateur de traduction est le dictionnaire fourni avec l'analyseur linguistique XELDA (version 2.8.1). Ce dictionnaire comporte 37 655 entrées anglaises alignées avec 59 495 traductions en français et 69 285 en allemand.

5.5.2 Dictionnaire de synonymes

Le dictionnaire de synonymes utilisé par le générateur est aussi fourni avec l'analyseur linguistique XELDA (version 2.8.1). Il comporte 5 064 entrées associées à 7 596 synonymes pour l'anglais, 2 387 entrées associées à 3 169 synonymes pour le français, 4 209 entrées associées à 4 883 synonymes pour l'allemand.

5.5.3 Table de traduction de morphèmes liés

À notre connaissance, il n'existe pas de dictionnaire bilingue de morphèmes liés. Constituer une ressource de ce type qui soit quasi-exhaustive sort du cadre de notre travail de thèse. C'est pourquoi nous nous sommes contentés de décrire dans un fichier les traductions possibles des morphèmes liés contenus dans les termes sources [10]. Ces traductions pouvaient être des morphèmes liés ou des morphèmes libres. Par exemple, le confixe *-phyto-* peut être traduit en français par *-phyto-* ou *plante*. Pour constituer cette ressource, les traducteurs se sont aidés de dictionnaires morphologiques monolingues spécialisés présents sur Internet [11] et d'un dictionnaire encyclopédique (Drosdowski, 2006).

La ressource indique pour chaque morphème lié anglais sa nature (préfixe, confixe, suffixe) et la liste de ses traductions ainsi que leurs natures (préfixe, confixe, suffixe, mot). Dans la grande majorité des cas, la forme liée de la traduction est de la même nature que le morphème source, i.e. un préfixe est traduit par un préfixe, un confixe est traduit par un confixe, un suffixe

10. La table anglais-français a été faite par l'auteure de la présente thèse, la table anglais-allemand a été faite par une traductrice.

11. http ://medical-dictionary.thefreedictionary.com/,
http ://georges.dolisi.free.fr/Terminologie/Menu/terminologie_medicale_menu.htm

est traduit par un suffixe. La seule exception concerne le suffixe anglais *-less 'sans'* qui peut être traduit par les préfixes privatifs *ab-* ou *a-*.

Allomorphes et formes interfixées sont traités comme des entrées indépendantes. Par exemple, le confixe anglais dérivé du grec *-plasis- 'action de façonner, modeler'* est présent sous la forme de trois entrées : *-plasia-*, *-plasty-* et *-plasy-*. Le confixe *-patho-* a trois traductions liées possibles en allemand : *-path-*, *-pathie-* et *-patho-*.

L'extrait 5.2 donne un aperçu de ces tables de traduction. Les confixes sont notés suivis du symbole :c, les suffixes sont notés avec :s, les préfixes avec :p et les mots avec :w. La totalité de la ressource est consultable en annexe B.3.3.

EN	DE	FR
patho:c	behandlung:w, krankheit:w, leiden:w, path:c, pathie:c, patho:c	maladie:w, path:c, pathie:c, patho:c, souffrance:w
phyto:c	pflanze:w, pflanzen:w, phyt:c, phyto:c	bourgeon:w, excroissance:w, phyt:c, phyto:c, plante:w, végétal:w
plasia:c	plasie:c, plastisch:c, plastischer:w, umformbarkeit:w, verformung:w	modeler:w, plase:c, plasie:c, plasique:c
plasty:c	plasie:c, plastisch:c, plastischer:w, umformbarkeit:w, verformung:w	plastie:c, plastique:c, plastique:w, réparation:w
plasy :c	plasie:c, plastisch:c, plastischer:w, umformbarkeit:w, verformung:w	
less:s	a:p, ab:p, abs:p, los:s, nicht:w, ohne:w	a:p, ab:p, aucun:w, privé:w, sans:w

.

Extrait 5.2 – Extrait de la table de traduction des morphèmes anglais → français

Les tableaux 5.8 et 5.9 indiquent la taille des tables de traductions. Au total, ces dernières comportent 242 entrées en anglais qui sont associées à 1001 traductions en français et 1081 en allemand.

	# entrées EN	#traductions FR				
		préfixes	confixes	suffixes	mots	TOTAL
Préfixes	50	97	0	0	163	260
Confixes	185	0	410	0	310	720
Suffixes	7	2	0	6	13	21
TOTAL	**242**	**99**	**410**	**6**	**486**	**1001**

TABLE 5.8 – Taille des tables de traduction des morphèmes liés anglais → français (nb. d'entrées et traductions)

	# entrées EN	#traductions DE				
		préfixes	confixes	suffixes	mots	TOTAL
Préfixes	50	87	0	0	194	281
Confixes	185	0	385	0	382	767
Suffixes	7	3	0	13	17	33
TOTAL	**250**	**90**	**385**	**13**	**593**	**1081**

TABLE 5.9 – Taille des tables de traduction des morphèmes liés anglais → allemand (nb. d'entrées et traductions)

5.5.4 Lexiques pour la décomposition des termes sources

Pour la décomposition des termes sources, nous utilisons un lexique de morphèmes liés et un lexique de morphèmes libres. Le lexique de morphèmes liés est simplement constitué des entrées de la table de traduction des morphèmes.

Le lexique de morphèmes libres est composé des entrées du dictionnaire bilingue généraliste, des entrées du dictionnaire de synonymes et de mots attestés dans le corpus anglais. Découpage en mots et lemmatisation ont été effectués avec l'analyseur XELDA. Afin d'obtenir un maximum d'entrées dans notre lexique de morphèmes libres, nous avons, lorsque cela était possible, scindé les mots rencontrés dans les ressources en d'autres mots. Par exemple, si le mot $ataxia\text{-}telangiectasia$ était présent dans le corpus, nous avons ajouté trois mots à notre lexique : $ataxia\text{-}telangiectasia,\ ataxia$ et $telangiectasia$. L'algorithme 2 décrit le processus.

Algorithme 2 Extraction d'une liste de morphèmes libres

Require: C (mots corpus), D (entrées dictionnaires)
  $words \leftarrow \emptyset$
  for all $word\_a$ **in** $C \cup D$ **do**
    $words$.add($word\_a$)
    for all $word\_b$ **in** split_on_hyphens_and_spaces($word\_a$) **do**
      $words$.add($word\_b$)
    end for
  end for
  return $words$

5.5.5 Familles morphologiques

Nous avons acquis automatiquement des familles de mots morphologiquement proches à l'aide de l'algorithme de racinisation de Porter (1980). Cet algorithme, qui est destiné à la recherche d'information, construit des familles de mots sur la base de racines communes obtenues à l'aide de règles de désuffixage-recodage. Par exemple, les mots $elaborately,\ elaborate,\ elaboration$ sont tous racinisés en $elabor$: on considère alors qu'ils appartiennent à la même famille morphologique.

Pour extraire les familles morphologiques nous avons racinisé, pour chaque langue, tous les mots du corpus ainsi que les entrées du dictionnaire bilingue et les entrées du dictionnaire

de synonymes, préalablement scindés avec l'algorithme 2. Nous avons obtenu 5 835 familles morphologiques anglaises (2,51 mots en moyenne par famille), 7 049 familles morphologiques françaises (2,45 mots en moyenne par famille) et 7 348 familles morphologiques allemandes (2,15 mots en moyenne par famille).

L'algorithme de racinisation produit parfois des erreurs. Par exemple, les mots *ironically 'ironiquement'*, *ironical 'ironique'* et *iron 'fer'* sont tous reliés à la racine *iron*. Par ailleurs, les mots *individualistic*, *individualist*, *individualisation, individualised* et *individualise* sont répartis sur deux familles alors qu'ils devraient appartenir à la même. Nous avons évalué la qualité des familles morphologiques. Pour cela, nous avons extrait, pour chaque langue, 50 familles morphologiques acquises automatiquement et avons observé :

- **Le taux de faux positifs :** pourcentage de paires de mots classés dans une même famille alors qu'ils n'appartiennent pas à une même famille, ex : classement de *iron* et *ironically* dans une même famille.
- **Le taux de faux négatifs :** pourcentage de paires de mots non classés dans une même famille alors qu'ils appartiennent à une même famille, ex : classement de *individualisation* et *individualised* dans deux familles différentes.

Les résultats montrent que le raciniseur a tendance à manquer des rapprochements morphologiques plutôt que de créer des rapprochements erronés, à l'exception de l'allemand qui semble obtenir de bons résultats [12]. Les détails de l'évaluation ainsi que des extraits des familles morphologiques sont donnés en annexe B page 221.

	EN	FR	DE
Faux positifs	0,19%	0,08%	0,02%
Faux négatifs	14,77%	15,45%	0%

TABLE 5.10 – Évaluation des familles morphologiques

5.5.6 Dictionnaire de cognats

Afin d'augmenter les possibilités de traduction, nous avons extrait du corpus des paires de cognats que nous utilisons comme un dictionnaire bilingue spécialisé propre au corpus. Pour identifier les cognats, nous avons repris la technique de Hauer et Kondrak (2011) : il s'agit d'un classifieur - LibSVM de Chang et Lin (2011) - entraîné sur des paires de traductions extraites de dictionnaires bilingues. Les variables prédictives sont les suivantes :

- Distance d'édition (Levenshtein, 1966)
- Plus long préfixe commun
- Nombre de bigrammes communs
- Longueur du terme source
- Longueur du terme cible
- Différence entre la longueur du terme source et la longueur du terme cible

Pour constituer les jeux d'apprentissage, nous avons utilisé quatre dictionnaires présents sur le site du FREE DICTIONARIES PROJECT [13] :

- Universal dictionary (maintenu par le projet DICTS.INFO)
- Wiktionary (maintenu par le projet WIKTIONARY.ORG)

12. Pour l'allemand, une partie des paires à été ôtée de l'évaluation car nous n'avons pas su déterminer si les mots appartenaient ou pas à la même famille.
13. http : //www.dicts.info/uddl.php

– Omegawiki (maintenu par le projet OMEGAWIKI.ORG)
– Wikipedia (maintenu par le projet WIKIPEDIA.ORG)

De ces quatre dictionnaires nous avons extrait toutes les paires `(terme source, terme cible)` telles que la distance d'édition entre le terme source et le terme cible est inférieure ou égale à quatre. Si terme source et terme cible étaient des traductions, ils constituaient un exemple positif ; sinon ils constituaient un exemple négatif. Nous avons sélectionné autant d'exemples positifs que négatifs afin que le modèle ne favorise pas l'une ou l'autre des classes (COGNAT vs. NON COGNAT). Nous avons obtenus 42 404 exemples pour l'anglais-français et 14 798 exemples pour l'anglais-allemand.

L'apprentissage a été effectué avec la libraire de fouille de données WEKA (Hall *et al.*, 2009) qui propose un paquetage pour LibSVM. Deux modèles ont été appris : un pour identifier les cognats anglais-français, un autre pour les cognats anglais-allemand. La moyenne des taux d'erreur obtenus par validation croisée à 10-blocs sont de 3,49 % pour la classification anglais-français et 6,93 % pour la classification anglais-allemand.

Nous avons créé le dictionnaire spécialisé grâce à l'algorithme 3. Pour chaque mot du corpus source, nous retenons comme traduction tout mot cible avec lequel la distance d'édition est inférieure ou égale à 4 et dont le classifieur SVM indique qu'il s'agit d'un cognat. Si plusieurs mots cibles satisfont ces critères, nous retenons celui qui a la plus petite distance d'édition avec le mot source. Les mots du corpus ont été préalablement scindés avec l'algorithme 2.

Algorithme 3 Extraction d'un dictionaire de cognats à partir d'un corpus comparable

Require: C_s (corpus source), C_t (corpus cible), $Classifier$ (classifieur cognats)
  $Dictionary \leftarrow \emptyset$
  for all w_s **in** C_s **do**
    $translation \leftarrow \emptyset$
    $score \leftarrow 4$
    for all w_t **in** C_t **do**
      $ed =$edit_distance(w_s, w_t)
      if $ed \leq score$ **and** $Classifier$.isCognate(w_s, w_t) **then**
        $translation \leftarrow w_t$
        $score = ed$
      end if
    end for
    if $translation \neq \emptyset$ **then**
      $Dictionnary$.add$(w_s, translation)$
    end if
  end for
  return $Dictionary$

Au final, nous avons obtenu 6 708 paires de traduction pour l'anglais-français et 6 391 pour l'anglais-allemand.

5.6 Synthèse

Ce chapitre nous a permis de présenter les données employées pour expérimenter la méthode de traduction présentée au chapitre 4 : corpus comparables, termes sources, lexiques

de références, données d'apprentissage et ressources linguistiques. Dans le chapitre suivant, nous décrivons les résultats obtenus par la génération de traduction.

Chapitre 6

Formalisation et évaluation de la génération de traductions candidates

Sommaire

Introduction

Dans ce chapitre, nous présentons la mise en œuvre de la méthode de génération traduction décrite dans le chapitre 4 (étapes 1 à 4). L'objectif de cette méthode est de pouvoir traduire

diverses sortes de mots morphologiquement complexes et de pouvoir générer des traductions fertiles.

Nous commençons par détailler l'algorithme de génération de traduction (section 6.1). Puis, nous présentons les résultats de l'évaluation du module de découpage morphologique dans la section 6.2. La section 6.3 évalue la qualité des traductions générées. Nous y décrivons diverses expériences dont le but est de mettre au jour l'apport des données et des stratégies de génération. Les limites de notre travail ainsi que les perspectives de recherche sont discutées dans la section 6.4.

6.1 Algorithme de génération de traductions

La méthode de traduction présentée au chapitre précédent est implantée grâce à l'algorithme 4. Cet algorithme prend en entrée un terme source monolexical ($source\_term$) et produit en sortie zéro, une ou plusieurs unités mono- ou poly-lexicales en langue cible ($translations$).

L'algorithme s'appuie sur les ressources décrites au chapitre précédent, que nous notons ainsi :

- $Trans$ est une ressource de traduction liant composants en langue source et composants en langue cible : table de traduction des morphèmes, dictionnaire bilingue généraliste et dictionnaire de cognats.
- Var^{src}, resp. Var^{tgt}, est une ressource qui permet de gérer la variation en langue source (src), resp. cible (tgt) : familles morphologiques, dictionnaire de synonymes.
- $Corpus^{src}$ un corpus segmenté en mots, lemmatisé et étiqueté en langue source.
- $Comp^{src}_{type}$, resp. $Comp^{tgt}_{type}$, est une liste de composants en langue source, resp. cible, où $type$ égale $pref$ pour les préfixes, $conf$ pour les confixes, $suff$ pour les suffixes et $word$ pour les mots (ces listes correspondent aux entrées des ressources $Trans$ et Var^{src} et aux mots extraits du corpus en langue source).
- $Stop^{tgt}$ est une liste de mots outils en langue cible.

Pour plus de clarté, nous accompagnons la description de l'algorithme avec l'exemple de la traduction du terme anglais $cytotoxic$ vers le français à partir des données suivantes ($src = en$ et $tgt = fr$) :

En tant que méthode de traduction compositionnelle, la fonction de génération de traductions se décompose en quatre sous-fonctions qui s'appliquent séquentiellement :

- Décomposition $\mathcal{D}$ (section 6.1.1)
- Traduction $\mathcal{T}$ (section 6.1.2)
- Recomposition $\mathcal{R}$ (section 6.1.3)
- Sélection $\mathcal{S}$ (section 6.1.4)

La traduction de $cytotoxic$ est donc donnée par une composition de fonctions, chacune s'appliquant sur le résultat fourni par la précédente :

$$Translation(\text{“cytotoxic”}) = \mathcal{S}(\mathcal{R}(\mathcal{T}(\mathcal{D}(\text{“cytotoxic”}))))$$

Algorithme 4 Génération de traductions

Require: $source\_term$, $target\_corpus$

```
translations ← ∅
for all {c_1, ...c_i} in D(source_term) do                    ▷ 1. Décomposition
    for all {t_1, ...t_j} in T(c_1) × ... T(c_i) do           ▷ 2. Traduction
        if i ≠ j then
            continue
        end if
        for all {w_1, ...w_k} in R ({t_1, ...t_j}) do         ▷ 3. Recomposition
            for all match in S({w_1, ...w_k}, target_corpus) do   ▷ 4. Sélection
                add match to translations
            end for
        end for
    end for
end for
return translations

function D(source_term)
    decompositions ← ∅
    for all {m_1, ...m_i} in SPLIT(source_term) do            ▷ 1.1 Découpage morphologique
        for all {c_1, ...c_j} in CONCATENATE ({m_1, ...m_i}) do   ▷ 1.2 Concaténation
            add {c_1, ...c_j} to decompositions
        end for
    end for
    return decompositions
end function

function R({t_1, ...t_i})
    recompositions ← ∅
    for all {t_1, ...t_i} in PERMUTATE ({t_1, ...t_i}) do     ▷ 3.1 Permutation
        for all {w_1, ...w_j} in CONCATENATE ({t_1, ...t_i}) do   ▷ 3.2 Concaténation
            if FILTER ({w_1, ...w_j}) then                    ▷ 3.3 Filtrage
                add {w_1, ...w_j} to recompositions
            end if
        end for
    end for
    return recompositions
end function
```

$Comp^{en}_{conf} = \{$-cyto-$\}$
$Comp^{en}_{word} = \{$cytotoxic, cytotoxicity, toxic$\}$
$Comp^{fr}_{conf} = \{$-cyto-$\}$
$Comp^{fr}_{word} = \{$cellule, toxique$\}$
$Trans = \{$
$\{$-cyto- $\rightarrow$ -cyto-, cellule$\}$,
$\{$toxic $\rightarrow$ toxique$\}$,
$\{$cytotoxicity $\rightarrow$ cytotoxicité$\}$
$\}$
$Var^{en} = \{$cytoxic $\rightarrow$ cytoxicity$\}$
$Stop^{fr} = \{$pour, le$\}$
$Corpus^{fr} =$ "le/DET cytotoxicité/N être/AUX le/DET propriété/N de/PREP ce/DET qui/PRO être/AUX toxique/A pour/PREP le/DET cellule/N ./PUN"

TABLE 6.1 – Exemple de données pour la traduction de $cytotoxic$ vers $toxique\ pour\ les\ cellules$ et $cytotoxicité$

6.1.1 Décomposition

La fonction de décomposition $\mathcal{D}$ comprend elle-même en deux fonctions nommées SPLIT (découpage morphologique) et CONCATENATE (concaténation des morphèmes) :

$\mathcal{D}$("cytotoxic"')
= CONCATENATE(SPLIT("cytotoxic"'))

6.1.1.1 Découpage morphologique (SPLIT)

Cet étape décompose le terme source en morphèmes en suivant l'algorithme 5. Le découpage se fait par simple projection des entrées des ressources $Comp^{src}$, $Comp^{src}_{conf}$, $Comp^{src}_{suff}$, $Comp^{src}_{word}$ sur la chaîne de caractères représentant le terme source. On prend également en compte des contraintes de longueur sur les sous-chaînes appariées avec les entrées des ressources ($\mathcal{L}0$, $\mathcal{L}1$, $\mathcal{L}2$). Par exemple, la chaine $string[1:n]$, composée de n caractères, peut être découpée en une base lexicale $string[1:i]$ et un suffixe $string[i+1:n]$ si $string[i+1:n] \in Comp^{src}_{suff}$ et $string[1:i] \geq \mathcal{L}2$. Les contraintes de longueur utilisées par l'algorithme 5 ont été paramétrées empiriquement ($\mathcal{L}0 = 5$, $\mathcal{L}1 = 4$, $\mathcal{L}2 = 4$).

Un terme source est d'abord décomposé en un préfixe optionnel + une base$_1$, laquelle est décomposée en une base$_2$ + un suffixe optionnel. Pour finir, la base$_2$ est décomposée en un ou plusieurs confixes ou mots. Lorsque plusieurs découpages sont possibles, seulement ceux ayant donné le plus grand nombre de composants sont retenus.

CONCATENATE(SPLIT("cytotoxic"'))
= CONCATENATE({cyto, toxic})

Algorithme 5 SPLIT : Découpage morphologique

Require: $source\_term$, $Comp_{pref}$, $Comp_{conf}$, $Comp_{suff}$, $Comp_{word}$
$\mathcal{L}0 = 5$; $\mathcal{L}1 = 4$; $\mathcal{L}2 = 4$;

$lemmas\_splits \leftarrow [\,]$ ▷ Découpage sur les tirets puis découpage de chaque sous-élément
for all $lemma$ **in** SPLIT_ON_HYPHENS($source\_term$) **do**
  $prefix =$ GET_PREFIX($lemma$) ▷ Extraction du préfixe optionnel
  $base =$ REMOVE_PREFIX($prefix, lemma$)
  $suffix =$ GET_SUFFIX($base$) ▷ Extraction du suffixe optionnel
  $base =$ REMOVE_SUFFIX($lemma, suffix$)
  $splits \leftarrow \emptyset$ ▷ Plusieurs décompositions possibles pour la base
  for all $(c_1, ...c_n)$ **in** GET_COMPONENTS($base$) **do** ▷ $(c_1, ...c_n)$ est la liste des composants de la base
    add $(prefix, c_1, ..., c_n, suffix)$ to $splits$
  end for
  add $splits$ to $lemmas\_splits$
end for

▷ Combinaison de tous les découpages de chaque sous-élément
$splits = \{(s_1, ..., s_n) | s_1 \in lemmas\_splits[1], ..., s_n \in lemmas\_splits[n]\}$
▷ Retourne les combinaisons avec le plus grand nombre de morphèmes
return $\{split | split \in splits$, length($split$) = max ($\{$length($split$)$|split \in splits\}$) $\}$

function GET_PREFIX($lemma$)
  for all $pref$ **in** sorted by descending length ($Comp_{pref}$) **do**
    $base =$ REMOVE_PREFIX($prefix, lemma$)
    if $lemma = pref$ or ($lemma$ starts with $pref$ and length($base$)$\geq \mathcal{L}0$ and ($base \in Comp_{words}$ or $base$ can be decomposed into words or confixes)) **then**
      return $pref$
    end if
  end for
end function

function GET_COMPONENTS($base$)
  $C = \{w | w \in Comp_{words}$ and length(w) $\geq \mathcal{L}2\} \cup Comp_{conf}$
  return $splits = \{(c_1, ..., c_n) | c_i \in C$ for all $1 \leq i \leq n$, and $c_1 + ... + c_n = base\}$
end function

function GET_SUFFIX($lemma$)
  for all $suff$ **in** sorted by descending length ($Comp_{suff}$) **do**
    $base =$ REMOVE_SUFFIX($lemma, suffix$)
    if $lemma$ ends with $suff$ and length($base$) $>$ length($suff$) and length($base$) $\geq \mathcal{L}1$ **then**
      return $suff$
    end if
  end for
end function

6.1.1.2 Concaténation des morphèmes (CONCATENATE)

La fonction CONCATENATE génère toutes les concaténations possibles des morphèmes issus du découpage morphologique. Par exemple, si le terme source abc a été découpé en trois morphèmes {a,b,c}, alors il existe quatre façons de concaténer ces morphèmes : {abc}, {a,bc}, {ab,c}, {a,b,c}. Pour un terme source ayant été découpé en n morphèmes, il existe 2^{n-1} concaténations possibles.

CONCATENATE({cyto, toxic})
$=$ {cyto, toxic}, {cytotoxic}

Cette étape de concaténation des morphèmes permet d'augmenter les possibilités de trouver des traductions. Prenons l'exemple où nous souhaitons traduire le terme $non\text{-}cytotoxic$ et où nos ressources contiennent des traductions pour non, $cyto$ et $cytotoxic$ mais pas pour $toxic$. Si nous nous basons uniquement sur la sortie du découpage morphologique, c'est-à-dire $\{non\text{-}, \text{-}cyto\text{-}, toxic\}$, la traduction de $non\text{-}cytotoxic$ échoue car il n'y a pas de traduction pour $toxic$. Par contre, si nous passons par une étape de concaténation intermédiaire, nous obtenons aussi la décomposition $\{non\text{-}, cytotoxic\}$. Comme nos ressources contiennent des traductions pour non et $cytotoxic$, nous pouvons traduire le terme source.

Dans le cas où nous aurions quand même eu la traduction $toxic \rightarrow toxique$, la concaténation permet de générer plusieurs traductions possibles et donc des variantes potentiellement intéressantes pour le traducteur. Par exemple, si le dictionnaire possède les traductions $non\text{-} \rightarrow non$, $\text{-}cyto\text{-} \rightarrow cellule$, $toxic \rightarrow toxique$, et $cytotoxic \rightarrow cytotoxique$, alors, nous pouvons générer à la fois la traduction $non\ toxique\ pour\ les\ cellules$ et $noncytotoxique$.

Notons enfin que la concaténation regénère la forme entière du terme source (i.e. non découpée). Une fois de plus, ceci permet d'augmenter les possibilités de traduire le terme, par exemple, en passant par un cognat et/ou une variante morphologique. Si nous savons que $noncytotoxic$ appartient à la même famille morphologique que $noncytotoxicity$ et que nous savons que $noncytotoxicity$ se traduit par $noncytotoxicité$, alors nous pouvons établir un lien de traduction direct entre $noncytotoxic$ en $noncytotoxicité$ ce qui reste une traduction acceptable. Cette stratégie est une solution de repli intéressante lorsque le terme n'a pas pu être découpé en morphèmes ou que la traduction de l'un des morphèmes a échoué.

6.1.2 Traduction

La fonction de traduction donne une traduction pour chaque décomposition générée par $\mathcal{D}$. Nous appuyant sur le principe de compositionnalité, nous considérons que la traduction du tout est fonction de la traduction des parties : $\mathcal{T}(a, b) = \mathcal{T}(a) \times \mathcal{T}(b)$. Pour une décomposition $\{c_1, ...c_n\}$ ayant n composants, il existe donc $\prod_{i=1}^{n} |\mathcal{T}(c_i)|$ traductions possibles.

$\mathcal{T}$({cyto, toxic}, {cytotoxic})
$= \mathcal{T}$(cyto) $\times\ \mathcal{T}$(toxic), $\mathcal{T}$(cytotoxic)
$=$ {cyto, toxique},{cellule, toxique}, {cytotoxicité}

Les traductions sont obtenues en utilisant les ressources $Trans$ et Var. Si le composant est identifié comme un morphème lié (préfixe, confixe, suffixe), ses traductions seront celles données par la table de traduction des morphèmes. Si le composant est un mot, alors nous utilisons le dictionnaire généraliste et le dictionnaire de cognats. Dans les cas où le mot est

composé d'une seule lettre ou correspond à un sigle, nous ajoutons également ce mot tel quel comme traduction. Ceci permet de retrouver des traductions de termes comportant des noms de gènes, sigles ou abréviations empruntés directement à l'anglais, ex. : $\underline{p}\text{-}value \rightarrow valeur\ de\ p$, $\underline{er}\text{-}negative \rightarrow negativ\ er$[1]. Les deux ressources bilingues peuvent être combinées aux ressources monolingues (familles morphologiques, synonymes) de façon à gérer les cas de variation.

En résumé, la traduction d'un mot peut être :

- une traduction directe, obtenue en consultant les ressources bilingues :
  - $\mathcal{T}(c) = Trans(c)$, ex. : $toxic \rightarrow toxique$.
- une traduction indirecte obtenue en :
  - traduisant le mot puis en recherchant une variante de la traduction :
    - $\mathcal{T}(c) = Var^{tgt}(Trans(c))$, ex. : $toxic \rightarrow toxique \rightarrow toxicité$ pour une variante morphologique ou $toxic \rightarrow toxique \rightarrow vénéneux$ pour un synonyme.
  - recherchant une variante du mot source puis en traduisant cette variante en langue cible :
    - $\mathcal{T}(c) = Trans(Var^{src}(c))$, ex. : $toxic \rightarrow toxicity \rightarrow toxicité$ pour une variante morphologique ou $toxic \rightarrow poisonous \rightarrow vénéneux$ pour un synonyme.

Pour un jeu de composants donné ($\{c_1, ..., c_i\}$), si l'un des composants ne peut être traduit, la traduction du tout échoue.

6.1.3 Recomposition

La fonction de recomposition $\mathcal{R}$ prend en entrée les traductions générées par $\mathcal{T}$ et les recompose en une suite d'unités lexicales. La recomposition se déroule en trois étapes : permutation des composants traduits (PERMUTATE), concaténation en mots (CONCATENATE) et filtrage (FILTER) :

$\mathcal{R}$({cyto,toxique},{cellule,toxique},{cytotoxicité})
=FILTER(CONCATENATE(PERMUTATE({cyto,toxique},{cellule,toxique}, {cytotoxicité})))

6.1.3.1 Permutation des composants traduits (PERMUTATE)

Cette étape génère, pour une traduction de n éléments, les $n!$ permutations de ces éléments. Par exemple, la traduction{A,B,C} donne six permutations : {A,B,C}, {A,C,B}, {B,A,C}, {B,C,A}, {C,A,B} et {C,B,A}. Cette phase permet de prendre en compte le phénomène de distortion : l'ordre des unités traduites peut être différent d'une langue à l'autre. Il est coûteux d'utiliser des fonctions de complexité $O(n!)$ mais dans notre cas nous manipulons de petits ensembles (au maximum quatre composants).

FILTER(CONCATENATE(PERMUTATE({cyto,toxique},{cellule,toxique}, {cytotoxicité})))
=FILTER(CONCATENATE({cyto,toxique}, {toxique,cyto}, {cellule, toxique},
{toxique, cellule}, {cytotoxicité}))

1. ER est l'abréviation en anglais de *Estrogen Receptor 'récepteur des œstrogènes'*.

6.1.3.2 Concaténation en mots (CONCATENATE)

Une fois les composants permutés, nous générons, pour chacune des permutations, toutes les concaténations possibles de ses composants (comme nous le faisons après le découpage morphologique). Cette étape sert à reformer les mots cibles à partir des composants traduits.

FILTER(CONCATENATE({cyto,toxique}, {toxique,cyto}, {cellule, toxique},
{toxique, cellule}, {cytotoxicité}))
=FILTER({cyto, toxique}, {cytotoxique}, {toxique, cyto},{toxiquecyto}, {cellule, toxique},
{celluletoxique}, {toxique, cellule},{toxiquecellule},{cytotoxicité})

6.1.3.3 Filtrage (FILTER)

Cette étape filtre les sorties de CONCATENATE à l'aide d'heuristiques. Par exemple une séquence d'unités lexicales en langue cible $L = \{l_1, ..., l_n\}$ sera éliminée si l'une des unités lexicales correspond à un morphème lié ou si la séquence finit par certains mots outils. Par exemple, la recomposition {cytotoxique} sera acceptée mais pas {-cyto-, toxique} car -$cyto$- est un morphème lié et il ne peut apparaître en tant que morphème libre. Une séquence comme {traitement, après} est éliminée en français car finissant par la préposition $après$ (*a priori*, un terme cible finissant par un mot outil est mal formé, c'est pourquoi nous éliminons ces cas-là).

FILTER({cyto, toxique}, {cytotoxique}, {toxique, cyto},{toxiquecyto}, {cellule, toxique},
{celluletoxique}, {toxique, cellule},{toxiquecellule},{cytotoxicité})
= {cytotoxique}, {toxiquecyto}, {cellule, toxique}, {celluletoxique}, {toxique, cellule},
{toxiquecellule}, {cytotoxicité}

Ces suites de composants concaténés correspondent aux unités lexicales qui seront projetées dans le corpus cible grâce à la fonction de sélection. Par exemple, la suite {toxique$_A$, cellule$_B$} correspond à l'unité lexicale $toxique$ suivie de $cellule$. La suite {cytotoxique$_{AB}$} correspond à une seule unité lexicale : $cytotoxique$.

6.1.4 Sélection

La fonction de sélection $\mathcal{S}$ tente d'apparier les suites d'unités lexicales générées par $\mathcal{R}$ avec les lemmes des mots du corpus cible. Nous appelons $L = \{l_1, ...l_n\}$ une suite d'unités lexicales générées par $\mathcal{R}$. Nous appelons $W = \{w_1, ...w_m\}$ une suite de mots du corpus cible, $l(w_k)$ est le lemme du mot w_k et $p(w_k)$ la partie du discours du mot w_k. L s'apparie avec W si il existe une séquence strictement croissante d'indices $I = \{i_1, ...i_n\}$ tels que $l(w_{i_j}) = l_j$ et $\forall j, 1 \leq j \leq n$ et $\forall i, 1 \leq |i_{j-1} - i_j| \leq \mathcal{L}3$ et $\forall w_k |\ k \notin I,\ l(w_k) \in Stop^{tgt}$; $\mathcal{L}3$ ayant été fixé à 3 empiriquement.

$\mathcal{S}$({cytotoxique}, {toxiquecyto}, {cellule, toxique}, {celluletoxique}, {toxique, cellule},
{toxiquecellule}, {cytotoxicité})
= {"toxique/A pour/PREP le/DET cellule/N", "cytotoxicité/N"}

En d'autres termes, L est une sous-séquence de lemmes de W et nous autorisons au maximum $\mathcal{L}3$ mots outils entre deux mots qui s'apparient avec les unités lexicales de L. Ainsi {toxique, cellule} s'apparie avec $\ll toxique\ pour\ les\ cellules \gg$ mais pas avec $\ll toxique\ étendu$

$aux\ cellules \gg$ (présence d'un mot lexical entre $toxique$ et $cellule$) ni avec $\ll toxique\ pour\ aucune\ de\ ces\ cellules \gg$ (plus de trois mots outils entre $toxique$ et $cellule$).

Définition d'une traduction candidate

Pour chaque suite d'unités lexicales L, nous collectons dans le corpus cible toutes les séquences de mots $W_1, W_2, ...W_p$ qui se sont appariées avec L selon la définition donnée en supra. Nous considérons que deux séquences $W1$ et $W2$ correspondent à une même traduction candidate si $|W1| = |W2|$ et $\forall(w1_i, w2_j)$ tels que $w1 \in W1, w2 \in W2, i = j$ alors $l(w1_i) = l(w2_j)$ et $p(w1_i) = p(w2_j)$, i.e. si deux séquences de mots correspondent à la même suite de paires `(lemme, partie du discours)`, alors ces deux séquences sont une et une seule traduction candidate. Ceci nous permet d'ignorer les différences de flexions, par exemple $\ll toxique\ pour\ \underline{la}\ cellul\underline{e} \gg$ et $\ll toxique\ pour\ \underline{les}\ cellul\underline{es} \gg$ correspondent à la même traduction candidate : "toxique/A pour/PREP le/DET cellule/N". Par contre, $\ll toxique\ \underline{pour}\ les\ cellules \gg$ et $\ll toxique\ \underline{envers}\ les\ cellules \gg$ sont deux traductions candidates différentes.

Cette première section nous a permis de présenter la totalité de l'algorithme de génération de traductions. Les sections suivantes concernent l'évaluation de l'algorithme. Nous avons d'abord évalué l'algorithme de découpage morphologique (section 6.2), puis l'algorithme de traduction dans son ensemble (section 6.3).

6.2 Évaluation du découpage morphologique

Pour évaluer la fonction de découpage morphologique SPLIT, nous avons observé :
- Le nombre de mots qui n'avaient pas pu être découpés ;
- Parmi les mots qui ont pu être découpés :
  - Le nombre de ceux qui avaient reçu au moins un découpage correct ;
  - Parmi ceux qui n'ont pas reçu un découpage correct [2] :
    - combien ont été surdécoupés, ex. : {ligation, de, pendent} au lieu de {ligation, dependent} ;
    - combien ont été sous-découpés, ex. : {cyclophosphamide, based} au lieu de {cyclo, phosphamide, based}.

Ces résultats sont donnés dans le tableau 6.2. La liste S^{FR} (1839 mots), resp. S^{DE} (1824 mots) correspond à la liste des termes sources à traduire en français, resp. en allemand, présentée dans la section 5.2. Les résultats sont stables quelle que soit la liste de mots découpés.

Nous observons qu'environ 3,15 % à 3,23 % des termes ne sont pas découpés. Pour les termes ayant pu être découpés, 93 % d'entre eux ont reçu au moins un découpage correct. Il y a peu de surgénération : seuls deux termes ont reçu deux découpages différents, et parmi ces découpages, au moins un était correct :
- $grandparent$ est découpé en `grand+parent` et en `grandpa+rent`
- $lymphoedema$ est découpé en `-lymph-+oedema` et en `-lympho-+edema`

Pour les termes ayant été mal découpés, 76,85 % à 77,27 % d'entre eux ont été surdécoupés.

2. Nous avons eu des cas où le nombre de morphèmes était correct mais le découpage faux, ex. : {grandpa,rent} au lieu de {grand, parent} mais dans ces cas-là, le découpage correct avait également été généré.

Par exemple, *ligation-dependent* a été découpé en `ligation+de-+pendent`. Le sur-découpage n'est pas dommageable car grâce à la concaténation, nous pouvons retrouver le découpage correct. Par contre, 12,12 % à 12,37 % des cas ont été sous-découpés. Par exemple, *cyclophosphamide-based* est découpé en `cyclophosphamide+based`.

	S^{FR}	S^{DE}
non découpés	3,15 %	3,23 %
découpés	96,85 %	96,77 %
↪ nb. découpages / terme	1,001	1,001
↪au moins 1 découpage correct	93,82 %	93,88 %
↪aucun découpage correct	6,18 %	6,12 %
↪ sur-découpage	77,27 %	76,85 %
↪ sous-découpage	22,73 %	23,15 %

TABLE 6.2 – Résultats obtenus par la fonction de découpage morphologique SPLIT

6.3 Évaluation des traductions générées

Dans cette section, nous présentons l'évaluation des traductions générées avec l'algorithme 4. Nous commençons par présenter les mesures et références utilisées pour l'évaluation (6.3.1) puis nous présentons les résultats de plusieurs expériences visant à mettre en avant les apports des ressources et des stratégies de génération (6.3.2 à 6.3.6). Nous finissons par une analyse qualitative des traductions générées (6.3.7).

6.3.1 Références et mesures d'évaluation

Pour évaluer les traductions générées, nous utilisons les deux références présentées dans le chapitre précédent, section 5.3 : référence *a posteriori* et référence *a priori*.

6.3.1.1 Référence *a posteriori*

Cette référence est constituée des sorties du système annotées manuellement (cf. section 5.3.2). Les termes sources correspondent aux listes S^{FR} et S^{DE} décrites en section 5.2.

Pour l'évaluation avec la référence *a posteriori* nous employons les mesures suivantes :

Couverture (C)

La couverture correspond à la fraction de termes sources pour lesquels le système a pu générer une traduction, quelle que soit son exactitude :

$$C = \frac{|ST|}{|S|} \tag{6.1}$$

$$ST = \{s : |\mathcal{T}(s)| > 1\}$$

où S est l'ensemble des termes sources et $\mathcal{T}(s)$ est l'ensemble des traductions générées par le système pour le terme source s.

Précision (P)

La précision indique la fraction de termes de ST pour lesquels le système a généré au moins une traduction correcte, c'est-à-dire annotée comme EXACT ou ACCEPTABLE par les juges :

$$P = \frac{|SC|}{|ST|} \tag{6.2}$$

$$SC = \{s : s \in ST, \mathcal{A}(s) \cap \{\text{EXACT, ACCEPTABLE}\} \neq \emptyset\}$$

où $\mathcal{A}(s)$ est l'ensemble des annotations manuelles affectées aux traductions de s. Par la suite, nous faisons la différence entre une précision basée uniquement sur les traductions exactes (P_E) et une précision basée sur les traductions exactes ou acceptables (P_{EA}).

Utilisabilité (U)

Par utilisabilité, nous entendons indiquer la fraction de termes sources pour lesquels le système a pu générer au moins une traduction exacte. L'utilisabilité est donnée par le produit de la couverture et de la précision :

$$U = \frac{|SC|}{|S|} = C \times P \tag{6.3}$$

Par la suite, nous distinguons également U_E (traductions exactes uniquement) et U_{EA} (traductions exactes ou acceptables).

Résultats obtenus

Les résultats obtenus sont donnés dans le tableau 6.3. Nous notons que la génération est moins bonne pour l'anglais-allemand : aussi bien au niveau de la couverture (40 % *vs.* 36 %) que de la précision (48 % *vs.* 59 %). Il est difficile de se comparer aux autres travaux de traduction automatique d'unités monolexicales tant les approches et les données employées sont différentes (focalisation sur une structure morphologique particulière, utilisation d'exemples de traductions). Par exemple, Cartoni (2005) obtient 94 % de précision et tous les termes obtiennent au moins une traduction. Cependant, Cartoni se concentre uniquement sur la traduction de mots préfixés. Harastani *et al.* (2012) obtiennent également une très bonne précision (de 96 % à 98 %) pour une couverture plus faible (30 % à 37 %) mais ils se concentrent uniquement sur la traduction de composés classiques. Sur les composés populaires, Garera et Yarowsky (2008) obtiennent une moyenne de 19 % de précision et de 13 % de couverture sur 9 couples de langues. Nous proposons dans la section 6.3.2 de comparer notre méthode à des approches ciblées sur un seul type de structure morphologique et ce, sur la base de notre jeu de données.

	C	P_E	U_E	P_{EA}	U_{EA}
EN-FR	,40	,59	,24	,69	,28
EN-DE	,36	,48	,17	,56,	,20

TABLE 6.3 – Évaluation *a posteriori* de la génération de traduction

6.3.1.2 Référence *a priori*

Cette référence est un lexique bilingue obtenu en projetant le méta-thésaurus médical UMLS dans les textes de notre corpus (cf. section 5.3.1). Nous notons SR l'ensemble des termes sources appartenant à ce lexique. Les mesures d'évaluation sont les mesures standard de précision, rappel et F1-mesure.

Précision (P)

La précision indique la fraction de termes de ST pour lesquels le système a pu générer au moins une traduction qui correspond à celle donnée par l'UMLS.

$$P = \frac{|SR|}{|ST|} \tag{6.4}$$

$$SR = \{s : s \in ST, \mathcal{T}(s) \cap \mathcal{R}(s) \neq \emptyset\}$$

où $\mathcal{T}(s)$ est l'ensemble des traductions de s et $\mathcal{R}(s)$ est l'ensemble des traductions de s données par l'UMLS.

Comme nous l'avons évoqué dans la section 4.3, si nous nous basons uniquement sur les traductions de l'UMLS, nous écartons des cas où le système a quand même généré une traduction correcte. Par exemple, pour le terme $mastectomy$ '$mastectomie,\ ablation\ du\ sein$', notre système a généré les traductions $mastektomie$, $entfernung\ des\ brust$, $abschnitt\ ein\ brust$, $ablation\ der\ brust$ qui sont toutes des traductions correctes. Or, la traduction donnée par l'UMLS est $brustamputation$, traduction qui n'a pas été générée par notre système.

Nous proposons donc de calculer également les précisions P_E et P_{EA}. P_E prend en compte les traductions données par l'UMLS et les traductions annotées comme EXACT. P_{EA} prend en compte les traductions données par l'UMLS et les traductions annotées comme EXACT ou ACCEPTABLE par les juges. Le tableau 6.4 donne des exemples de traductions annotées EXACT ou ACCEPTABLE par les juges mais non présentes dans l'UMLS.

Rappel (R)

Le rappel indique la fraction de termes sources pour lesquels le système a pu générer au moins une traduction qui correspond à celle donnée par l'UMLS :

$$R = \frac{|SR|}{|S|} \tag{6.5}$$

Nous calculons également R_E et R_{EA} à partir des traductions annotées EXACT et/ou ACCEPTABLE.

F1-mesure (F1)

Enfin, la F1-mesure permet de rendre compte du compromis entre précision et rappel :

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{6.6}$$

Nous calculons également $\mathrm{F1}_E$ et $\mathrm{F1}_{EA}$ à partir de P_E, R_E et P_{EA}, R_{EA}.

Résultats obtenus

Les résultats obtenus avec l'évaluation *a priori* sont donnés dans le tableau 6.5. Nous y observons de très bonnes précisions (de 0,8 à 0,7 en prenant uniquement en compte les traductions de l'UMLS) alors que dans le cas de l'évaluation *a posteriori*, la précision était entre 0,59 et 0,48. La différence entre la précision P et les précisions P_E et P_{EA} qui prennent en compte les traduction EXACT et ACCEPTABLE montre à quel point se baser uniquement sur les traductions de l'UMLS élimine une partie des traductions correctes générées par le système. Par exemple, dans le cas de P_E, la précision passe de 0,8 à 0,94 pour le français et de 0,7 à 0,88 pour l'allemand. Une fois de plus, nous voyons que les résultats sont moins bons pour l'allemand que pour le français bien que ce ne soit pas aussi marqué que pour l'évaluation *a posteriori*.

La référence *a priori* nous permet de comparer notre approche aux approches empiriques qui évaluent également leurs algorithmes sur l'UMLS. Un des avantages de ces approches est que, à l'instar de notre méthode, elles ne ciblent pas un type de construction morphologique particulier. Par contre, elles nécessitent de nombreux exemples de traductions. Claveau (2009) obtient les meilleurs résultats : 88 % de précision et 88 % de rappel sur des langues très proches (portuguais → espagnol) et 58 % de précision et 58 % de rappel sur des langues éloignées (anglais → russe). Langlais *et al.* (2009) obtiennent les moins bons résultats : 57 % de précision et 19 % de rappel pour la traduction du français vers l'anglais ; 63 % de précision et 23 % de rappel pour la traduction de l'espagnol vers l'anglais. Claveau et Kijak (2011), qui traduisent du français vers le japonais, obtiennent 63 % de précision et 45 % de rappel s'ils comptent uniquement les traductions présentes dans l'UMLS. Ils obtiennent 89 % de précision et 64 % de rappel s'ils prennent en compte toutes les traductions correctes générées par l'algorithme (leurs résultats sont alors comparables aux nôtres).

	P	R	F1	P_E	R_E	$F1_E$	P_{EA}	R_{EA}	$F1_{EA}$
EN-FR	,80	,52	,63	,94	,62	,75	,95	,63	,76
EN-DE	,70	,51	,59	,88	,64	,74	,89	,66	,76

TABLE 6.5 – Évaluation *a priori* de la génération de traduction

Nous avons mené diverses expériences visant à mettre au jour les apports des stratégies de génération et des différentes ressources linguistiques. Pour chaque expérience, nous présentons les résultats de l'évaluation *a posteriori* et *a priori*. Les expériences menées portent sur :

- l'apport de la généricité du modèle (section 6.3.2)
- l'apport des ressources linguistiques (section 6.3.3)
- l'apport de la stratégie de repli (section 6.3.4)
- l'apport des traductions fertiles (section 6.3.5)
- l'apport du corpus vulgarisé (section 6.3.6)

Enfin, nous finissons par une analyse qualitative des silences et des erreurs de notre système (section 6.3.7).

6.3.2 Apport de la généricité du modèle

Dans cette expérience nous comparons notre système à des approches qui ne chercheraient à traduire qu'un type particulier de construction morphologique ou qui ne seraient basées que sur l'identification de cognats. Nous avons choisi quatre bases de comparaison :

Préfixation Cette approche ne traduit que les mots construits par préfixation. Le terme source est découpé (lorsque c'est possible) en préfixe + base, le préfixe est traduit avec la table de traduction des morphèmes (nous ne retenons que les traductions qui sont elles-mêmes des préfixes) et la base est traduite avec le dictionnaire généraliste. Base et préfixe sont concaténés en un seul mot. Cela correspond approximativement à l'approche de Cartoni (2009b).

Les mots sources composés d'un préfixe et d'une base lexicale autonome représentent 31 % de la référence *a posteriori* français (32 % pour l'allemand) et 25 % de la référence *a priori* français (34 % pour l'allemand).

Composition savante Cette approche ne traduit que les mots construits par composition savante. Le terme source est découpé (lorsque c'est possible) en un ensemble de confixes ou un ensemble de confixes et mots simples, les confixes sont traduits avec la table de traduction des morphèmes (nous ne retenons que les traductions qui sont elles-mêmes des confixes) et les mots sont traduits avec le dictionnaire généraliste. Tous les morphèmes traduits sont concaténés en un seul mot. L'ordre des éléments n'est pas conservé. Cela correspond approximativement à l'approche de Harastani *et al.* (2012) à la différence qu'ils distinguent éléments initiaux et éléments finaux et que l'ordre des éléments est conservé.

Les mots sources composés d'au moins 1 confixe et de bases lexicales autonomes ou d'autres confixes représentent 18 % de la référence *a posteriori* français et allemand et 56 % de la référence *a priori* français (52 % pour l'allemand).

Composition populaire Cette approche ne traduit que les mots construits par composition populaire. Le terme source est découpé (lorsque c'est possible) en un ensemble de mots simples et les mots sont traduits avec le dictionnaire généraliste. Les traductions peuvent être concaténées en un seul mot ou en plusieurs mots. L'ordre des mots n'est pas conservé. Cela correspond approximativement à l'approche de Weller *et al.* (2011) si ce n'est qu'elles se basent sur des patrons de traduction prédéfinis.

Les mots sources composés de plusieurs bases lexicales représentent 48 % de la référence *a posteriori* français (49 % pour l'allemand) et 25 % de la référence *a priori* français (21 % pour l'allemand).

Cognat Cette approche traduit n'importe quel mot en lui assignant comme traduction son cognat lorsque ce dernier existe. Les cognats sont identifiés avec la méthode décrite section 5.5.6.

Nous avons comparé ces quatre méthodes de "base" avec notre système qui prend en compte plusieurs structures morphologiques et se fonde aussi sur des ressources permettant de gérer la variation. Les résultats sont donnés dans les tableaux 6.6, 6.7 (référence *a posteriori*) et 6.8, 6.9 (référence *a priori*).

6.3.2.1 Évaluation *a posteriori*

Sur la référence *a posteriori*, nous voyons que la plupart des autres méthodes sont beaucoup plus précises que notre méthode. Pour autant, ces méthodes génèrent très peu de

traductions, ce qui, au final, fait que notre méthode produit le lexique le plus utilisable (au sens de notre mesure d'utilisabilité U), quel que soit le couple de langues.

Si nous comparons les quatres méthodes de base entre elles, nous observons que la composition savante est la plus précise, suivie de près par la préfixation. Ces deux méthodes, par contre, ont une très faible couverture mais qui peut être expliquée par le fait que notre liste de termes sources contient en majorité des composés populaires. La méthode la plus intéressante des quatre se révèle être celle des cognats. C'est celle qui obtient les meilleurs scores d'utilisabilité (U). Composition populaire et cognats ont une précision relativement faible. Les résultats sont meilleurs pour les cognats lorsque l'on prend en compte les traductions acceptables ; en effet, beaucoup de cognats identifiés sont des variantes morphologiques ($aromatherapist \rightarrow aromathérapie$, $comprehensively \rightarrow comprehensive$). Des cas d'erreurs impliquent typiquement des mots préfixés comme $pretreatment \rightarrow treatment$ ou des combinaisons avec un autre mot très court : $fulvestrant\text{-}er \rightarrow fulvestrant$.

Pour les composés populaires, les erreurs sont dans leur vaste majorité causées par les traductions fertiles souvent combinées à un ordre de mots différents du terme source, par exemple :

- $low\text{-}risk$ est traduit par $bas\ car\ cela\ risquer$ qui apparait dans le contexte « $...sauf\ si\ le\ nombre\ de\ plaquettes\ est\ trop\ \boldsymbol{bas\ car\ cela\ risquerait}\ de\ ...$ »
- $strong\text{-}smelling$ est traduit par $fort\ pour\ bien\ sentir$ qui apparait dans le contexte « $Presser\ assez\ \boldsymbol{fort\ pour\ bien\ sentir}\ le\ tissu\ mammaire$ ».

Pour autant, certaines traductions ont un sens assez proche de celui du terme source : $milk\text{-}producing$ est traduit par $lait\ produire$ (contexte : « $canal\ évacuant\ le\ \boldsymbol{lait\ produit}\ par\ la\ glande\ mammaire.$ »).

6.3.2.2 Évaluation *a priori*

Sur la référence *a priori*, les résultats sont à peu près sembables à ceux de l'évaluation *a posteriori* : les autres méthodes sont plus précises mais au final, nous obtenons la meilleure F1-mesure. Nous obtenons aussi toujours le meilleur rappel. La précision des cognats est toujours bien meilleure lorsque l'on prend en compte les traductions EXACT ou ACCEPTABLE. Préfixation et composition ont de très bonnes précisions mais les cognats restent, des quatre méthodes de base, la méthode donnant la meilleure F1-mesure.

La différence avec l'évaluation *a posteriori* est qu'ici, la composition populaire obtient de très bons résultats. En français, sur quatre composés populaires, la traduction donnée par l'UMLS est trouvée directement : $workload \rightarrow charge\ de\ travail$, $lifestyle \rightarrow style\ de\ vie$, $viewpoint \rightarrow point\ de\ vue$, $half\text{-}life \rightarrow demivie$. Pour l'allemand, il est difficile de tirer des conclusions : un seul composé populaire a obtenu une traduction et la traduction générée, bien qu'exacte, n'est pas celle proposée dans l'UMLS : $child\text{-}birth$ est traduit par $geburt\ ein\ kind$ *'naissance d'un enfant'* (traduction annotée EXACT) alors que l'UMLS propose $geburt$ *'naissance'*[3].

6.3.2.3 Synthèse

Nous pouvons conclure qu'il est intéressant de chercher à traduire plusieurs types de structures morphologiques tout en s'appuyant sur des ressources permettant de gérer la variation plutôt que de se concentrer sur un seul type de construction : même si les traductions

3. « *Die* $\boldsymbol{Geburt\ eines\ Kindes}$ *läutet umfassende Veränderungen in der weiblichen Brust ein.* » *'La naissance d'un enfant annonce des changements majeurs dans la poitrine féminine.'*

TERME SOURCE	TRADUCTION UMLS	TRADUCTION GÉNÉRÉE	ANNOTATION
heterozigozity	heterozygot (DE)	heterozygozity	EXACT
radiography	roentgen (DE)	radiograpisch	ACCEPTABLE
lumpectomy	ablation d'une tumeur (FR)	lumpectomie	EXACT
co-repressors	co-répresseur (FR)	corépression	ACCEPTABLE

TABLE 6.4 – Différences entre les références *a priori* et *a posteriori*

	C	P_E	U_E	P_{EA}	U_{EA}
Composition savante	,03	**,95**	,03	**1**	,03
Cognat	,13	,66	,08	,81	,10
Composition populaire	,05	,63	,03	,65	,03
Préfixation	,02	,90	,02	,97	,02
Notre méthode	**,40**	,59	**,24**	,69	**,28**

TABLE 6.6 – Comparaison avec d'autres méthodes de génération, évaluation *a posteriori* anglais-français

	C	P_E	U_E	P_{EA}	U_{EA}
Composition savante	,03	**,96**	,02	**,98**	,02
Cognat	,10	,58	,06	,66	,07
Composition populaire	,04	,55	,02	,62	,03
Préfixation	,03	,86	,02	,92	,03
Notre méthode	**,36**	,48	**,17**	,56	**,20**

TABLE 6.7 – Comparaison avec d'autres méthodes de génération, évaluation *a posteriori* anglais-allemand

	P	R	F1	P_E	R_E	$F1_E$	P_{EA}	R_{EA}	$F1_{EA}$
Composition savante	,83	,20	,32	,97	,23	,37	**1**	,24	,38
Cognat	,76	,37	,50	,89	,44	,59	,92	,45	,61
Composition populaire	**1**	,03	,06	**1**	,03	,06	**1**	,03	,06
Préfixation	,56	,04	,07	,89	,06	,12	**1**	,07	,13
Notre méthode	,80	**,52**	**,63**	,94	**,62**	**,75**	,95	**,63**	**,76**

TABLE 6.8 – Comparaison avec d'autres méthodes de génération, évaluation *a priori* anglais-français

	P	R	F1	P_E	R_E	$F1_E$	P_{EA}	R_{EA}	$F1_{EA}$
Composition savante	**,80**	,18	,29	,95	,21	,35	,95	,21	,35
Cognat	,62	,26	,36	,86	,36	,50	,89	,37	,52
Composition populaire	0.0	0.0	0.0	**1**	,01	,02	**1**	,01	,02
Préfixation	,75	,07	,12	**1**	,09	,16	**1**	,09	,16
Notre méthode	,70	**,51**	**,59**	,88	**,64**	**,74**	,89	**,66**	**,76**

TABLE 6.9 – Comparaison avec d'autres méthodes de génération, évaluation *a priori* anglais-allemand

sont moins précises, au final, le nombre de termes sources ayant reçu au moins une traduction correcte est plus important. Les cognats apparaissent comme une méthode intéressante pour compléter un lexique bilingue.

Notons que notre système n'est pas équivalent à l'union des quatres méthodes auxquelles il a été comparé. La forte couverture de notre système s'explique par le recours à diverses stratégies de traductions qui ne sont pas employées par les quatre méthodes auxquelles nous avons comparé notre système :

- Aucune des quatre méthodes n'emploie de ressources destinées à gérer la variation. Par exemple, la méthode de traduction basée sur préfixation ne pourra établir d'équivalence entre $bioavailable$ et $biodisponibilité$.
- À l'exception de la méthode basée sur les cognats, le dictionnaire de cognat n'est utilisé par aucune autre méthode. Par exemple, la méthode de traduction basée sur la composition populaire ne pourra traduire $taxane\text{-}treated$ en $traitement\ par\ taxane$ puisque la traduction de $taxane$ n'existe que dans le dictionnaire de cognats.
- Seule la méthode de traduction basée sur la composition populaire génère des traductions fertiles. Les autres méthodes ne peuvent pas générer de telles traductions. Par exemple, la méthode basée sur la composition savante ne pourra pas trouver d'équivalence entre $tumorectomy$ et $ablation\ de\ la\ tumeur$ et la méthode basée sur la préfixation ne pourra pas trouver d'équivalence entre $pre\text{-}chemotherapy$ et $avant\ la\ chimiothérapie$.
- Aucune des méthodes auxquelles nous avons comparé notre méthode ne propose de stratégie de repli. L'équivalence $uniformly \rightarrow uniforme$ ne peut être trouvée que par notre méthode.
- Pour finir, notre méthode propose aussi de traduire des termes suffixés. Ceci nous permet de trouver des traductions comme $retrospectively \rightarrow façon\ retrospective$.

6.3.3 Apport des ressources linguistiques

Dans cette expérience, nous avons souhaité évaluer l'apport des ressources permettant de gérer la variation (dictionnaire de synonymes, familles morphologiques) et du dictionnaire de cognats qui permet d'augmenter la taille du lexique bilingue.

Nous avons effectué quatre tests dont les résultats sont donnés dans les tableaux 6.10 à 6.13 :

Base La traduction a été effectuée uniquement avec le dictionnaire généraliste et la table de traduction des morphèmes.

Base + dictionnaire de cognats Dictionnaire généraliste, table de traduction des morphèmes et dictionnaire de cognats.

Base + familles morphologiques Dictionnaire généraliste, table de traduction des morphèmes et familles morphologiques.

Base + dictionnaire de synonymes Dictionnaire généraliste, table de traduction des morphèmes et dictionnaire de synonymes.

Toutes les ressources Dictionnaire généraliste, table de traduction des morphèmes, dictionnaire de cognats, familles morphologique et dictionnaire de synonymes.

6.3.3.1 Évaluation *a posteriori*

Les résultats observés sont quasi-identiques quel que soit le couple de langues. La méthode de base est la plus précise (à l'exception du cas où l'on prend en compte les traductions acceptables pour le français : dans ce cas, la combinaison base + cognats est aussi précise que les ressources de base seules car les cognats identifient un bon nombre de dérivés morphologiques, ex. : $aromatherapist \rightarrow aromatherapy$). Plus on rajoute de ressources, plus la couverture augmente et plus la précision baisse mais au final l'utilisabilité est meilleure sauf pour la combinaison base + synonymes : cette dernière ne produit pas un lexique plus utilisable que les ressources de base seules. Dictionnaire de cognats et familles morphologiques sont des ressources très utiles, ils fournissent notamment beaucoup de traductions jugées comme acceptables.

6.3.3.2 Évaluation *a priori*

Si l'on compte uniquement les traductions générées qui correspondent à celles de l'UMLS, nous retrouvons les mêmes résultats pour les deux couples de langues. Globalement, la meilleure F1-mesure est toujours obtenue avec la combinaison de toutes les ressources. Toutefois, la combinaison base + cognats obtient une meilleure précision que la base seule. Si nous comparons les F1-mesures, nous observons que les familles morphologiques et les synonymes ont peu d'impact par rapport à la base seule. La combinaison base+cognats et la combinaison de toutes les ressources obtiennent des F1-mesures équivalentes (en particulier pour l'anglais-allemand).

6.3.3.3 Synthèse

Globalement, nous voyons que l'ajout de nouvelles ressources permet de traduire plus de termes mais fait baisser la précision. Le dictionnaire de synonymes présente peu d'intérêt. Pour les autres ressources, nous notons une différence entre les résultats de l'évaluation *a posteriori* et l'évaluation *a priori*. Dans l'évaluation *a posteriori*, cognats et dérivés morphologiques ont un impact positif mais les meilleurs résultats sont obtenus avec la combinaison de toutes les ressources. Dans l'évaluation *a priori*, les dérivés morphologiques ont un impact faible (surtout en allemand), seuls les cognats ont un apport réellement intéressant. Nous pensons que cette différence est due à la nature des termes sources à traduire. Dans la référence *a priori*, la plupart des termes à traduire ont une graphie très proche de leur traduction (ex. : $translocation \rightarrow translokation,\ cytogenetic \rightarrow cytogénétique$), ce qui se prête bien à l'identification de traduction par cognats. Dans la référence *a posteriori*, nous retrouvons beaucoup plus de mots composés et de mots qui ne sont pas graphiquement proches de leur traduction ($tumor\text{-}margin \rightarrow tumorrand$). Les mots composés, par exemple, se prêtent bien à la génération de traductions fertiles ; de plus, la fertilité implique souvent une variation morphologique ; ex. : $cytoprotection \rightarrow protéger\ la\ cellule$.

6.3.4 Apport de la stratégie de repli

Dans cette section, nous évaluons l'intérêt de la stratégie de repli décrite en 6.1.1.2. Cette stratégie permet, lorsque le terme n'a pu être décomposé ou qu'un des composants n'est pas traduit, d'essayer de trouver une traduction directement en passant soit par le

	C	P_E	U_E	P_{EA}	U_{EA}
Base	,16	**,73**	,12	**,77**	,12
Base + dictionnaire de cognats	,28	,71	,19	**,77**	,21
Base + familles morphologiques	,27	,56	,15	,66	,18
Base + dictionnaire synonymes	,17	,69	,12	,72	,13
Toutes les ressources	**,40**	,59	**,24**	,69	**,28**

TABLE 6.10 – Apport de ressources linguistiques, évaluation *a posteriori*, anglais → français

	C	P_E	U_E	P_{EA}	U_{EA}
Base	,15	**,60**	,09	**,63**	,10
Base + dictionnaire de cognats	,27	,56	,15	,61	,16
Base + familles morphologiques	,24	,48	,12	,57	,14
Base + dictionnaire synonymes	,17	,55	,09	,60	,10
Toutes les ressources	**,36**	,48	**,17**	,56	**,20**

TABLE 6.11 – Apport de ressources linguistiques, évaluation *a posteriori*, anglais → allemand

	P	R	F1	P_E	R_E	$F1_E$	P_{EA}	R_{EA}	$F1_{EA}$
Base	,78	,29	,42	,93	,34	,50	**,98**	,36	,52
Base + dictionnaire de cognats	**,81**	,51	,62	**,94**	,59	,72	,95	,60	,73
Base + familles morphologiques	,76	,29	,42	**,94**	,37	,53	**,98**	,38	,55
Base + dictionnaire synonymes	,77	,29	,42	**,94**	,35	,51	**,98**	,37	,53
Toutes les ressources	,80	**,52**	**,63**	**,94**	**,62**	**,75**	,95	**,63**	**,76**

TABLE 6.12 – Apport de ressources linguistiques, évaluation *a priori*, anglais → français

	P	R	F1	P_E	R_E	$F1_E$	P_{EA}	R_{EA}	$F1_{EA}$
Base	,71	,27	,39	**,94**	,36	,52	,94	,36	,52
Base + dictionnaire de cognats	**,75**	,49	**,59**	,93	,61	**,74**	**,95**	**,62**	,75
Base + familles morphologiques	,67	,29	,40	,87	,38	,53	,90	,39	,54
Base + dictionnaire synonymes	,63	,27	,38	,87	,37	,52	,87	,37	,52
Toutes les ressources	,70	**,51**	**,59**	,88	**,64**	**,74**	,89	,66	**,76**

TABLE 6.13 – Apport de ressources linguistiques, évaluation *a priori*, anglais → allemand

dictionnaire de cognats soit par la combinaison d'une ressource de variation (synonymes, familles morphologiques) et d'une ressource bilingue (dictionnaire généraliste, cognats). Par exemple, $exactly$ a pu être traduit en $exakt$ en passant de $exactly$ à $exact$ via les familles morphologiques anglaises puis de $exact$ à $exakt$ en passant par les cognats anglais-allemand. Les résultats sont donnés dans les tableaux 6.14 à 6.17.

6.3.4.1 Évaluation *a posteriori*

Si l'on observe les sorties annotées, nous voyons que, globalement, cette stratégie de repli permet d'augmenter la couverture avec une légère baisse de la précision qui n'impacte pas l'utilisabilité finale du lexique. La stratégie est plutôt efficace avec les cognats pour la traduction du français vers l'anglais mais son apport n'est pas évident pour l'allemand. Pour les synonymes et familles morphologiques, il n'y a pas d'amélioration franche. Par contre, l'impact est net lorsque l'on combine toutes les ressources : l'utilisabilité augmente de 3 points pour l'allemand et de 4 points pour le français (traductions exactes et acceptables). La combinaison de toutes les ressources permet d'obtenir des traductions en passant à la fois par le dérivé morphologique et le cognat comme dans l'exemple ci-dessus.

6.3.4.2 Évaluation *a priori*

On observe globalement le même phénomène si l'on se base sur l'UMLS : impact visible avec les cognats (surtout pour l'allemand), faible impact avec les familles morphologiques et les synonymes, impact plus important obtenu avec la combinaison de toutes les ressources. Toutefois, ici, l'impact est beaucoup plus net pour l'anglais-allemand que pour le français-allemand : lorsqu'on prend en compte traductions exactes et acceptables, la mesure $F1_{EA}$ augmente de 5 points pour l'allemand, de 1 point pour le français.

6.3.4.3 Synthèse

Pour résumer l'apport de la stratégie de repli, nous voyons que ce sont les cognats et leur combinaison avec les familles morphologiques qui ont un réel intérêt. L'apport de cette stratégie de repli est surtout marqué pour la traduction vers l'allemand.

6.3.5 Apport des traductions fertiles

Dans cette section, nous effectuons deux comparaisons :

Traductions non fertiles vs. fertiles Ici, nous comparons la qualité de chaque type de traductions : les traductions fertiles sont elles plus exactes que les traductions non fertiles ? Génèrent-elles plus de traductions ?

Traductions non fertiles vs. toutes les traductions Ici, nous observons l'effet de l'ajout des traductions fertiles à un lexique qui ne serait composé que de traductions non fertiles : les traductions fertiles améliorent-elles la qualité finale du lexique ?

	C	P_E	U_E	P_{EA}	U_{EA}
Base + dictionnaire de cognats − repli	,24	**,72**	,17	**,79**	,19
Base + dictionnaire de cognats + repli	**,28**	,71	**,19**	,77	**,21**
Base + familles morphologiques − repli	,26	**,57**	,15	,66	,17
Base + familles morphologiques + repli	**,27**	,56	,15	,66	**,18**
Base + dictionnaire synonymes − repli	,17	**,70**	,12	**,73**	,12
Base + dictionnaire synonymes + repli	,17	,69	,12	,72	**,13**
Toutes les ressources − repli	,35	**,60**	,21	,69	,24
Toutes les ressources + repli	**,40**	,59	**,24**	,69	**,28**

TABLE 6.14 – Apport de la stratégie de repli, évaluation *a posteriori*, anglais → français

	C	P_E	U_E	P_{EA}	U_{EA}
Base + dictionnaire de cognats − repli	,24	**,58**	,14	**,62**	,15
Base + dictionnaire de cognats + repli	**,27**	,56	**,15**	,61	**,16**
Base + familles morphologiques − repli	,22	**,49**	,11	,56	,12
Base + familles morphologiques + repli	**,24**	,48	**,12**	**,57**	**,14**
Base + dictionnaire synonymes − repli	,17	**,56**	,09	**,61**	,10
Base + dictionnaire synonymes + repli	,17	,55	,09	,60	,10
Toutes les ressources − repli	,31	**,50**	,15	,56	,17
Toutes les ressources + repli	**,36**	,48	**,17**	,56	**,20**

TABLE 6.15 – Apport de la stratégie de repli, évaluation *a posteriori*, anglais → allemand

	P	R	F1	P_E	R_E	$F1_E$	P_{EA}	R_{EA}	$F1_{EA}$
Base + dictionnaire de cognats − repli	**,82**	,48	,61	**,96**	,56	,71	**,97**	,57	,72
Base + dictionnaire de cognats + repli	,81	**,51**	**,62**	,94	**,59**	**,72**	,95	**,60**	**,73**
Base + familles morphologiques − repli	,75	,29	,41	,94	,36	,52	,98	,37	,54
Base + familles morphologiques + repli	**,76**	**,29**	**,42**	,94	**,37**	**,53**	,98	**,38**	**,55**
Base + dictionnaire synonymes − repli	,77	,29	,42	,94	,35	,51	,98	,37	,53
Base + dictionnaire synonymes + repli	,77	,29	,42	,94	,35	,51	,98	,37	,53
Toutes les ressources − repli	,78	,48	,60	**,95**	,59	,73	**,97**	,60	,75
Toutes les ressources + repli	**,80**	**,52**	**,63**	,94	**,62**	**,75**	,95	**,63**	**,76**

TABLE 6.16 – Apport de la stratégie de repli, évaluation *a priori*, anglais → français

6.3.5.1 Évaluation *a posteriori*

Quelle que soit la langue (tableaux 6.18 et 6.19), nous observons que les traductions fertiles sont de nettement moins bonne qualité que les traductions non fertiles (la précision perd 6 à 20 points pour l'anglais-français ; 32 à 39 points de précision de l'anglais-allemand). De plus, les traductions fertiles ne permettent pas de générer un plus grand nombre de traductions que les traductions non fertiles.

Toutefois, ce type de traduction est intéressant en combinaison avec les traductions non fertiles car les traductions fertiles apportent un bon complément : elles permettent d'augmenter la couverture et, bien qu'elles fassent baisser la précision, au final, le lexique extrait est plus utilisable. Nous voyons que les traductions fertiles sont plus intéressantes pour l'anglais-français que pour l'anglais-allemand : augmentation de la couverture de 16 points et baisse de la précision de 1 points (traductions exactes) pour le français ; augmentation de la couverture de 12 points et baisse de la précision de 10 points (traductions exactes) pour l'allemand.

Les tableaux 6.20 et 6.21 détaillent l'impact des traductions fertiles lorsqu'elles sont combinées aux traductions non fertiles. Pour la traduction du français vers l'anglais, nous voyons que lorsque des traductions fertiles sont ajoutées à un terme source pour lequel nous n'avions pas pu générer une traduction "classique", dans la majorité des cas, la traduction fertile est correcte (EXACT ou ACCEPTABLE). Ce n'est pas le cas pour l'allemand par contre, où dans les trois-quart des cas, toutes les traductions fertiles sont fausses. Lorsque les traductions fertiles viennent se rajouter aux traductions "classiques", le lexique ne peut qu'en bénéficier :

- Si aucune traduction classique n'est correcte et que :
  - les traductions fertiles amènent au moins une traduction correcte, la précision augmente.
  - aucune des traductions fertiles n'est correcte, la précision reste stable.
- S'il y a au moins une traduction classique correcte et que :
  - il y a une traduction fertile correcte : la précision reste stable, mais une variante a été trouvée.
  - il n'y a aucune traduction fertile correcte : la précision reste stable également.

433 termes sources avec traduction(s) fertile(s)
↪ **289 (67 %) n'avaient aucune traduction**
↪ 160 (55 %) cas où les traductions fertiles ramènent une traduction correcte*
↪ 129 (45 %) cas où les traductions fertiles ramènent du bruit
↪ **144 (33 %) avaient déjà une traduction (non fertile)**
↪ 16 (11 %) cas où les traductions fertiles créent une correction$^{\blacklozenge}$
↪ 60 (42 %) cas où les traductions fertiles ajoutent une variante†
↪ 68 (47 %) cas où les traductions fertiles sont sans effet*

TABLE 6.20 – Impact détaillé des traductions fertiles, évaluation *a posteriori*, anglais-français

	P	R	F1	P_E	R_E	$F1_E$	P_{EA}	R_{EA}	$F1_{EA}$
Base + dictionnaire de cognats − repli	**,76**	,46	,57	,93	,56	,69	,93	,56	,69
Base + dictionnaire de cognats + repli	,75	**,49**	**,59**	,93	**,61**	**,74**	**,95**	**,62**	**,75**
Base + familles morphologiques − repli	,67	,29	,40	,87	,38	,53	,90	,39	,54
Base + familles morphologiques + repli	,67	,29	,40	,87	,38	,53	,90	,39	,54
Base + dictionnaire synonymes − repli	**,67**	,27	,38	**,89**	,36	,51	**,89**	,36	,51
Base + dictionnaire synonymes + repli	,63	,27	,38	,87	**,37**	**,52**	,87	**,37**	**,52**
Toutes les ressources − repli	**,72**	,48	,57	,87	,58	,69	,88	**,59**	,71
Toutes les ressources + repli	,70	**,51**	**,59**	**,88**	**,64**	**,74**	**,89**	,66	**,76**

TABLE 6.17 – Apport de la stratégie de repli, évaluation *a priori*, anglais → allemand

	C	P_E	U_E	P_{EA}	U_{EA}
Traductions non fertiles	,24	**,58**	**,14**	**,75**	**,18**
Traductions fertiles	,24	,52	,12	,55	,13
Traductions non fertiles	,24	,58	,14	**,75**	,18
Toutes les traductions	**,40**	**,59**	, **24**	,69	**,28**

TABLE 6.18 – Apport des traductions fertiles, évaluation *a posteriori*, anglais-français

	C	P_E	U_E	P_{EA}	U_{EA}
Traductions non fertiles	**,24**	**,58**	**,14**	**,69**	**,16**
Traductions fertiles	,20	,26	,05	,30	,06
Traductions non fertiles	,24	**,58**	,14	**,69**	,16
Toutes les traductions	**,36**	,48	**,17**	,56	**,20**

TABLE 6.19 – Apport des traductions fertiles, évaluation *a posteriori*, anglais-allemand

371 termes sources avec traduction(s) fertile(s)
↪ **219 (30 %) n'avaient aucune traduction**
↪ 57 (26 %) cas où les traductions fertiles ramènent une traduction correcte*
↪ 162 (74 %) cas où les traductions fertiles ramènent du bruit
↪ **152 (70 %) avaient déjà une traduction (non fertile)**
↪ 11 (7 %) cas où les traductions fertiles créent une correction$^{\blacklozenge}$
↪ 162 (28 %) cas où les traductions fertiles ajoutent une variante†
↪ 99 (65 %) cas où les traductions fertiles sont sans effet*

TABLE 6.21 – Impact détaillé des traductions fertiles, évaluation *a posteriori*, anglais-allemand

*EXACT OU ACCEPTABLE
$^{\blacklozenge}$les traductions fertiles contiennent une traduction correcte alors que toutes les traductions non fertiles sont fausses
†il y a à la fois une traduction fertile correcte et une traduction non fertile correcte
*toutes les traductions fertiles sont fausses et soit toutes les traductions non fertiles sont fausses, soit une des traductions non fertiles est correcte

6.3.5.2 Évaluation *a priori*

Si l'on se base sur la référence *a priori*, nous observons dans les tableaux 6.22 et 6.23 que les traductions fertiles ne sont un complément vraiment intéressant que pour le français (de 4 à 3 points de plus de F1-mesure selon que l'on se base uniquement sur les traductions de l'UMLS, sur les traductions exactes ou sur les traductions exactes ou acceptables). Voici le détail des traductions fertiles obtenues pour le français (nous indiquons d'une astérisque les traductions fertiles correspondant à la traduction de l'UMLS) :

Au moins une traduction fertile correcte[4] **- pas de traduction classique**
- *g-protein* → *protéine g**
- *lifestyle* → *style de vie**

Au moins une traduction fertile correcte et une traduction classique correcte
- *cardiotoxicity* → *toxicité cardiaque*, cardiotoxicité*
- *cytogénétique* → *génétique de cellule*[5]*, cytogénétique*

Traductions fertiles incorrectes et au moins une traduction classique correcte
- *mammaplasty* → *plastique de sein, mammoplastie*
- *overweight* → *supérieur avec et sans charge, surcharge*

Traductions fertiles incorrectes et traductions classiques incorrectes
- *in-patient*[6] → *pas malade, non malade, inverse chez le patient, [...] impatience*

En ce qui concerne la traduction vers l'allemand (tableau 6.23), l'apport des traductions fertiles n'est pas du tout évident : la F1-mesure varie de - 2 à +1 point lorsqu'on ajoute les traductions fertiles au lexique. Voici le détail des traductions fertiles obtenues pour l'allemand :

Au moins une traduction fertile correcte - pas de traduction classique
- *childbirth* → *geburt ein kind 'naissance d'un enfant'*

5. **cytogenetic** abnormalities / instability → anomalie / instabilité **génétique des cellules**
6. *in-patient* signifie *'patient hospitalisé en journée'*

– *hypercalcaemia → zu viel calcium in das blut 'trop de calcium dans le sang'*

Au moins une traduction fertile correcte et une traduction classique correcte

– *chemo-radiotherapy → strahlen und chemotherapie 'radio- et chimiothérapie', radiochemotherapy, chemoradiotherapy*
– *self-examination → selbst untersuchen 's'examiner', selbstuntersuchung 'auto-examen'*

Traductions fertiles incorrectes et au moins une traduction classique correcte

– *childhood → zustand der kind 'état de l'enfant', kindheit*
– *gynaecomastia → frau brust 'femme poitrine', frau mit brust 'femme avec poitrine', [...] gynäkomastie*

Traductions fertiles incorrectes - pas de traduction classique

– *breathless → ohne atmen 'sans respirer'*
– *ultrasound → über die fest 'sur le solide'*

Traductions fertiles incorrectes et classiques incorrectes

– *workplace→ stellen sich noch die aufgabe'rendre la tâche encore', aufgabestellung 'tâche'*

6.3.5.3 Synthèse

Nous avons observé que, quel que soit le type d'évaluation, les traductions fertiles étaient moins intéressantes pour l'allemand. Une explication partielle est que le corpus anglais-allemand est moins comparable que le corpus anglais-français (0,45 *vs.* 0,74) : les traductions anglais-allemand sont globalement de moins bonne qualité que les traductions anglais-français.

Une autre explication implique le type morphologique des langues étudiées. En effet, en tant que langue latine, le français aura plus facilement tendance à utiliser des syntagmes composés de plusieurs mots (typiquement les structures NOM$_1$ PRÉPOSITION NOM$_2$ ou NOM ADJECTIF) alors que l'anglais et l'allemand, langues germaniques, créeront plus facilements de mots composés (de structure NOM$_2$+NOM$_1$ ou ADJECTIF+NOM). Par exemple, des composés tels *anthracycline-containing* (anglais) ou *Anthracyclin-enthaltende* (allemand) sont courants dans ces langues alors qu'en français, il n'existe pas de traduction non fertile pour ces composés : une traduction correcte serait *contenant de l'anthracycline*, des équivalents non fertiles comme **anthracycline-contenant* ou **contenant-anthracycline* sont agrammaticaux en français.

6.3.6 Apport du corpus vulgarisé

Dans cette section, nous souhaitons évaluer l'intérêt d'intégrer des textes vulgarisés à un corpus spécialisé. Généralement, les termes à traduire sont plutôt des termes appartenant au discours scientifique, du moins, c'est ce que nous avons observé sur nos données (cf. tableau 6.24).

termes sources présents uniquement dans les textes scientifiques	69 %
termes sources présents uniquement dans les textes vulgarisés	20 %
termes sources présents dans les textes scientifiques et vulgarisés	11 %

TABLE 6.24 – Présence des termes sources dans les corpus

Or, lorsque l'on a recours aux corpus comparables, c'est surtout parce que l'on manque de données parallèles dans les langues que l'on souhaite traiter. Mais parfois, même des données

comparables sont difficiles à trouver en quantité suffisante. Comme nous l'avons vu pour la collecte du corpus allemand, il a été difficile de collecter des textes spécialisés représentatifs du discours *scientifique*. Dans quelle mesure peut-on augmenter la taille du corpus en lui ajoutant des textes spécialisés mais appartenant plutôt au discours *vulgarisé* ? Prend-on le risque d'une baisse de la qualité du lexique extrait ?

Nous avons comparé les situations suivantes :

Corpus scientifique Prise en compte des traductions présentes uniquement dans le corpus scientique

Corpus vulgarisé Prise en compte des traductions présentes uniquement dans le corpus vulgarisé

Corpus entier Prise en compte de toutes les traductions.

Les tableaux 6.25 à 6.28 indiquent les résultats obtenus. Globalement, nous pouvons observer que lorsqu'on compare corpus scientifique et corpus vulgarisé, les résultats sont toujours meilleurs avec le corpus scientifique, à la fois en termes de couverture / précision (évaluation *a posteriori*) que de précision / rappel (évaluation *a priori*). Il est difficile d'en tirer une conclusion à partir du corpus anglais-français puisque le corpus scientifique est plus volumineux que le corpus vulgarisé [7] : il y a plus de possibilités d'y trouver des traductions. Par contre, pour l'anglais-allemand, les tailles sont comparables et il est net que les textes scientifiques sont plus intéressants que les textes vulgarisés.

Lorsque l'on compare corpus scientifique seul et corpus entier, les résultats montrent bien qu'il est intéressant d'avoir recours à des textes vulgarisés pour augmenter un corpus scientifique. Pour l'allemand, l'ajout du corpus vulgarisé provoque une légère baisse de la précision mais au final la qualité du lexique reste meilleure avec le corpus entier (utilisabilité pour la référence *a posteriori*, F1-mesure pour la référence *a priori*).

6.3.6.1 Fertilité et type de discours

Nous avons également souhaité savoir si un lien pouvait être établi entre type de discours et fertilité des traductions. Par exemple, nous pouvons supposer que les textes vulgarisés contiennent plus de traductions fertiles. Pour vérifier cela, nous avons comparé le nombre de fois où traductions fertiles et non fertiles ont été trouvées dans le corpus scientifique et dans le corpus vulgarisé. Comme les corpus scientifiques et vulgarisés sont de tailles différentes, nous avons normalisé ces comptes avec le nombre de mots distincts dans chaque corpus :

$$freq(f, c) = \frac{|t \in f \cap c|}{|words(c)|} \tag{6.7}$$

où f est le type de traductions (i.e. fertile ou non fertile), c est un corpus (i.e. corpus scientifique ou vulgarisé), $|t \in f \cap c|$ est le nombre de traductions EXACT ou ACCEPTABLE de type f trouvées dans le corpus c et $|words(c)|$ est le nombre de mots distincts dans le corpus c.

Les graphiques 6.1 et 6.2 indiquent la répartition de chaque type de traduction dans chaque type discours. On observe que, quelle que soit la langue, le corpus vulgarisé contient plus de traductions fertiles que le corpus scientifique.

7. Le corpus scientifique français comprend 1,45 fois plus de mots que le corpus vulgarisé anglais (cf. section 5.1)

	P	R	F1	P_E	R_E	$F1_E$	P_{EA}	R_{EA}	$F1_{EA}$
Traductions non fertiles	**,77**	**,48**	**,60**	**,92**	**,58**	**,71**	**,95**	**,60**	**,73**
Traductions fertiles	,38	,05	,08	,75	,10	,17	,75	,10	,17
Traductions non fertiles	,77	,48	,60	,92	,58	,71	,95	,60	,73
Toutes les traductions	**,80**	**,52**	**,63**	**,94**	**,62**	**,75**	,95	**,63**	**,76**

TABLE 6.22 – Apport des traductions fertiles, évaluation *a priori*, anglais-français

	P	R	F1	P_E	R_E	$F1_E$	P_{EA}	R_{EA}	$F1_{EA}$
Traductions non fertiles	**,74**	**,51**	**,61**	**,90**	**,62**	**,74**	**,92**	**,63**	**,75**
Traductions fertiles	0.0	0.0	0.0	,31	,06	,09	,38	,07	,11
Traductions non fertiles	**,74**	,51	**,61**	**,90**	,62	,74	**,92**	,63	,75
Toutes les traductions	,70	,51	,59	,88	**,64**	,74	,89	**,66**	**,76**

TABLE 6.23 – Apport des traductions fertiles, évaluation *a priori*, anglais-allemand

	C	P_E	U_E	P_{EA}	U_{EA}
Corpus vulgarisé	,25	,54	,14	,64	,16
Corpus scientifique	,35	,58	,20	,68	,24
Corpus entier	**,40**	**,59**	**,24**	**,69**	**,28**

TABLE 6.25 – Apport du corpus vulgarisé, évaluation *a posteriori*, anglais → français

	C	P_E	U_E	P_{EA}	U_{EA}
Corpus vulgarisé	,26	,44	,11	,53	,14
Corpus scientifique	,29	**,49**	,14	**,57**	,16
Corpus entier	**,36**	,48	**,17**	,56	**,20**

TABLE 6.26 – Apport du corpus vulgarisé, évaluation *a posteriori*, anglais → allemand

	P	R	F1	P_E	R_E	$F1_E$	P_{EA}	R_{EA}	$F1_{EA}$
Corpus vulgarisé	,76	,31	,44	**,96**	,39	,55	**,96**	,39	,55
Corpus scientifique	,77	,44	,56	,94	,53	,68	**,96**	,54	,69
Corpus entier	**,80**	**,52**	**,63**	,94	**,62**	**,75**	,95	**,63**	**,76**

TABLE 6.27 – Apport du corpus vulgarisé, évaluation *a priori*, anglais → français

	P	R	F1	P_E	R_E	$F1_E$	P_{EA}	R_{EA}	$F1_{EA}$
Corpus vulgarisé	,64	,36	,46	,86	,48	,61	,86	,48	,61
Corpus scientifique	,69	,47	,56	**,89**	,60	,72	**,90**	,61	,73
Corpus entier	**,70**	**,51**	**,59**	,88	**,64**	**,74**	,89	**,66**	**,76**

TABLE 6.28 – Apport du corpus vulgarisé, évaluation *a priori*, anglais → allemand

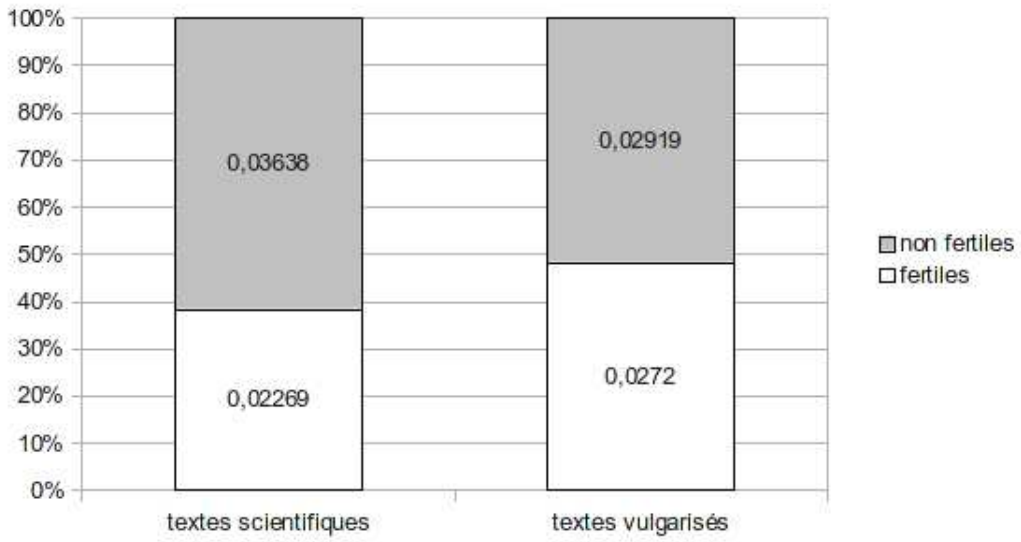


FIGURE 6.1 – Répartition des traductions fertiles et non fertiles dans les textes scientifiques et vulgarisés, anglais → français

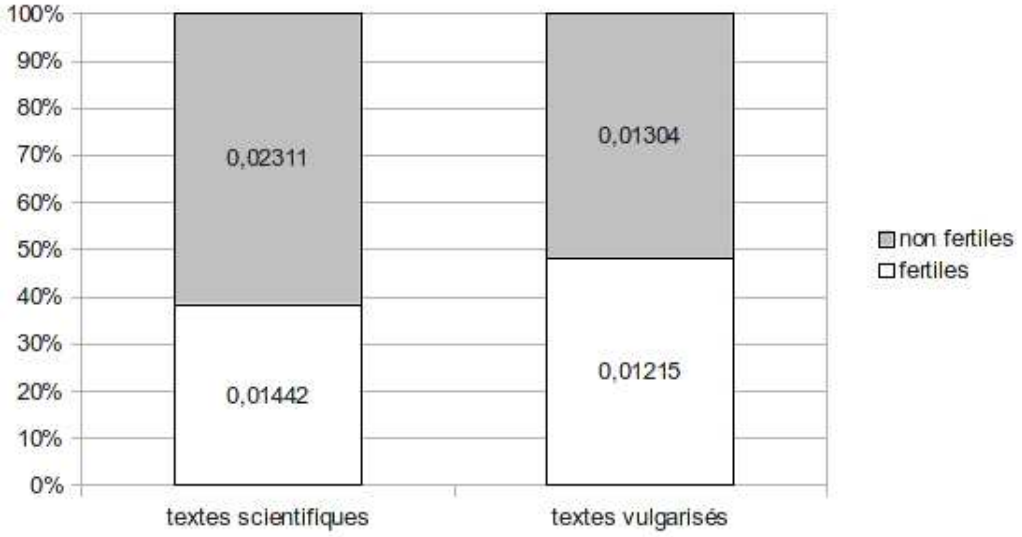


FIGURE 6.2 – Répartition des traductions fertiles et non fertiles dans les textes scientifiques et vulgarisés, anglais → allemand

6.3.7 Analyse qualitative

Notre dernière analyse porte sur les silences (termes sources non traduits) et les erreurs de notre système. Les données analysées consistent, pour chaque couple de langue, en :

- 50 cas de silences extraits aléatoirement parmi les traductions de l'UMLS qui n'ont pas été retrouvées par notre système ;
- 50 cas de traductions erronées extraits aléatoirement parmi les traductions générées par notre système et annotées comme FAUX ou PROCHE par les juges.

6.3.7.1 Analyse des silences

Les résultats de l'analyse sont donnés dans le tableau 6.29.

	EN-FR	EN-DE
non compositionnalité	20 (40 %)	25 (50 %)
↪ sens identique, découpage morphologique différent	10	13
↪ non-correspondance sémantique	10	11
↪ acronyme	0	1
traduction absente dans les ressources	15 (30 %)	18 (36 %)
semi-compositionnalité	12 (24 %)	2 (4 %)
↪ sens proche	8	2
↪ sens éloigné	4	0
recomposition et sélection	3 (6 %)	5 (10 %)
TOTAL	50	50

TABLE 6.29 – Analyse des cas de silence

La majorité des cas de silence (40 % à 48 %) sont dus à des cas qui ne peuvent pas être traités par la méthode compositionnelle :

- Soit le sens du terme source et du terme cible sont identiques mais le découpage morphologique est différent :
  - *collarbone* → *klavikula 'clavicule'* : deux morphèmes en anglais, un seul en allemand.
  - *newborn* → *nourrisson* : deux morphèmes en anglais, un seul en français.
- Soit le découpage sémantique entre les deux langues est différent, un des deux termes possède un ou de(s) élément(s) de sens supplémentaire(s) :
  - *ultrasonography* → *ultrashall 'ultrason'* : morphème supplémentaire *-graphy-* en langue source
  - *underam* → *achselhöle* : *achsel 'axiliaire'* correspond à peu près à *underarm 'sous le bras'* et *höle 'cavité'* n'a pas d'équivalent dans le terme source
  - *anti-emetic* → *médicaments antiémétiques* : *médicaments* n'a pas d'équivalent dans le terme source
- Soit le terme source est traduit par un acronyme :
  - *electrocardiogram* → *ekg*.

La deuxième cause de silence est due à la couverture de nos ressources. Entre 30 % et 36 % des traductions ont échoué car soit le terme n'a pu être découpé, soit un des composants n'a pu être traduit et la stratégie de repli a echoué :

- *mastectomy* → *brustamputation* : pas de correspondance entre *-ectomy* et *amputation*
- *pre-eclampsia* → *präeklampsie* : pas de correspondance entre *pre-* et *prä-* ni entre *eclampsia* et *eklampsie*
- *pharmacokinetics* → *pharmacocinétique* : pas de correspondance entre *kinetics* et *cinétique*
- *headache* → *maux de tête* : pas de correspondance entre *ache* et *mal* (le dictionnaire propose uniquement la locution *faire mal*).

Le troisième type de silence correspond à des cas de semi-compositionnalité : le nombre de morphèmes est identique et il y a correspondance de traduction entre une partie des morphèmes seulement. Pour l'autre partie des morphèmes, la distance sémantique est plus ou moins grande :

- Sens proche (cas de divergence lexicale) :
  - *bloodstream* → *circulation sanguine* : *stream* a la sens de *'courant'* plus que de *'circulation'*
  - *intra-abdominal* → *bauchhöhle* : *-intra-* a le sens de *'à l'intérieur'*, ce qui est un peu éloigné de *höhle 'cavité'*.
- Sens éloigné :
  - *brachytherapie* → *curiethérapie* :*-brachy-* a le sens de *'court, bref dans le temps'* et *curie* vient du nom propre *Curie* qui réfère à Marie Curie.
  - *abnormality* → *malformation* : *formation* n'est pas la traduction de *normality* qui a le sens de *'normalité'*.

Ces cas concernent principalement la traduction anglais → français (24 % *vs.* 4 % pour l'anglais → allemand).

Les derniers cas de silence concernent divers problèmes de recomposition et de sélection. Par exemple, en allemand, nous n'avons pas généré de *-s-* intermédiaire entre les mots d'un composé, nous n'avons donc pas pu retrouver la traduction de *workplace* (*arbeitsplatz*). À ceci s'ajoutent des problèmes liés au prétraitement des textes : caractères mal encodés, mauvaise normalisation de certains termes qui empêche l'appariemment, etc.

6.3.7.2 Analyse des erreurs

Les résultats de l'analyse sont donnés dans le tableau 6.30.

	EN-FR	EN-DE
mauvaise traduction d'un des composants	41 (82 %)	41 (82 %)
↪ sens non adapté au contexte	23	20
↪ variante trop éloignée	16	18
↪ cognat erroné	2	3
insertion de mots outils	8 (16 %)	3 (6 %)
↪ mauvais mots outils	7	2
↪ mots outils inutiles	1	1
ordre des éléments	1 (2 %)	6 (12 %)
TOTAL	50	50

TABLE 6.30 – Analyse des erreurs

La majeur partie des erreurs (82 %) provient de composants traduits de façon erronée :

- La traduction erronée peut provenir du dictionnaire généraliste ou de la table de traduction des morphèmes. Il s'agit alors de cas de polysémie ou de traductions qui ne sont pas valables dans le contexte dans lequel elles ont été trouvées :
  - *patient-related* → *patient et leur famille* : *related* a le sens de *'en rapport avec, relié'* mais aussi *'proche, membre de la famille'*
  - *gynaecomastia* → *brust der frau* : *brust der frau* signifie bien *'sein de la femme'* et la gynécomastie correspond effectivement à un surdéveloppement des seins chez l'homme mais ce n'est pas le sens que prend cette expression en contexte : « *...Die* ***Brust der Frau*** *ist außerdem ein sekundäres Geschlechtsmerkmal...* » *'Le sein de la femme est aussi une caractéristique sexuelle secondaire'*.
- La traduction erronée peut provenir de variantes (dérivés morphologiques, synonymes) qui produisent souvent une traduction au sens trop éloigné du morphème source :
  - *inactivate* → *inactivité* : bien que morphogiquement proche, *inactivité* ne peut pas être considéré comme une traduction correcte de *inactivate* qui a le sens de *'désactiver'*.
  - *incorrect* → *nicht zu ein verbesserung* : *verbesserung* signifie *'correction'* plutôt *correct*. Dans les textes du corpus *nicht zu ein verbesserung* a le sens de *'pas une amélioration'* : « *die Bestrahlung nach brusterhaltender Therapie* ***nicht zu einer Verbesserung*** *des Gesamtüberlebens* » *'l'irradiation après traitement conservateur du sein n'améliore pas la survie globale'*.
- Enfin, dans certains cas, la traduction a été trouvée dans le dictionnaire de cognats et elle est fausse :
  - *infrequently* → *manière plus fréquent* : le dictionnaire de cognats a donné la traduction *infrequent* → *fréquent*.
  - *in-patient* → *patient* : *patient* a été obtenu directement par le dictionnaire de cognats.

D'autres traductions incorrectes sont dues aux mots outils autorisés entre deux mots lors de la recherche du terme cible dans le corpus :

- Soit les mots outils ne sont pas les bons :
  - *prosurvival* → *favorable de survie* dans « *un facteur* ***favorable de survie*** ». La traduction exacte aurait été *favorable à la survie*
  - *hormone-sensitive* → *hormon oder empfindlich* dans « *...abhängig von bestimmten* ***Hormonen oder empfindlich*** *gegenüber bestimmten Substanzen...* » *'sujets à certaines hormones ou sensibles à certaines substances'*. La traduction correcte aurait été *empfindlich auf Hormone 'sensible aux hormones'*
- Soit ils sont inutiles, aucun mot outil n'aurait du être autorisé :
  - *non-mutation* → *non à un mutation « tumeurs du sein héréditaires liées ou* ***non à des mutations*** »
  - *hand-foot* → *hand oder fuss « ...neurotoxische Schädigungen der Nerven an* ***Händen oder Füßen****...* » *'dommages neurotoxiques pour les nerfs dans les mains ou les pieds'*. Le mot composé *hand-foot* fait partie de l'expression *hand-foot syndrome*[8] qui se traduit en allemand par *Hand-Fuß-Syndrom*. La traduction correcte aurait du être *handfuss*.

Le dernier type d'erreurs provient de l'ordre des éléments. Dans les cas observés, ce type d'erreur n'apparaît qu'avec des traductions fertiles :

8. En français : *syndrome main-pied*, rougeur et gonflement des mains et des pieds suite à un traitement de chimiothérapie.

– *pre-pregnancy* → *schwangerschaft vor* au lieu de *vorschwangerschaft 'avant la grossesse'*
– *first-year* → *an lors de leur premier* : ici, l'ordre attendu était plutôt *premier an*

6.4 Discussion

Cette section discute de notre méthode de génération de traduction : nous présentons d'abord un bilan global de nos expériences puis nous exposons les perspectives d'amélioration de la méthode.

6.4.1 Bilan

Référence *a priori* vs. *a posteriori*

Nous avons évalué notre méthode de génération sous plusieurs angles (apport des ressources, des techniques utilisées...) et notamment à l'aide de deux références : référence *a priori* et référence *a posteriori*. Ces deux références sont composées de termes sources différents. Pour la référence *a priori*, les termes sources sont effectivement des termes du domaine puisqu'ils sont présents dans le méta-thésaurus UMLS. En plus de permettre d'évaluer les cas de silence, cette référence permet aussi d'évaluer notre méthode sous l'angle de l'extraction de terminologies bilingues et de nous comparer à d'autres méthodes. La référence *a posteriori*, quant à elle, contient n'importe quel terme source pour peu qu'il n'ait pas d'entrée dans le dictionnaire bilingue généraliste ou que sa traduction n'existe pas dans le corpus cible. L'angle d'évaluation est plutôt celui de l'aide à la traduction spécialisée ou de l'enrichissement de ressources généralistes. Selon la référence utilisée, nous avons obtenu des nuances dans les résultats, en particulier au niveau de la comparaison des méthodes de traduction (composition populaire *vs.* savante *vs.* cognats...), de l'apport des ressources linguistiques et de l'apport de la stratégie de repli. La composition populaire, par exemple, ramène beaucoup plus de bruit avec la référence *a posteriori*. Les combinaisons familles morphologiques et cognats sont intéressantes pour la référence *a posteriori* mais pas pour la référence *a priori* où seuls les cognats ont un impact intéressant.

Comparaison des couples de langues

Concernant les langues, nous avons systématiquement observé de meilleurs résultats pour le couple anglais → français que pour le couple anglais → allemand. Deux choses peuvent expliquer cela : la moindre comparabilité du corpus anglais → allemand et les types morphologiques des langues (pour l'anglais → allemand, la recherche de traductions fertiles est moins pertinente). De plus, d'une façon générale, l'allemand est moins bien traité par notre système. Par exemple, nous ne gérons le cas de l'interfixe `-s-` qui peut s'intercaler entre deux mots dans les composés populaires. Nous avons vu que les familles morphologiques avaient plus d'impact pour le français. Or, il faut noter que pour l'allemand, ces dernières sont bien moins exhaustives (7 348 familles pour l'allemand, composées de 2,15 mots en moyenne contre 7 049 en français, composées de 2,45 mots en moyenne).

Traductions fertiles

La génération de traductions fertiles, bien que de moins bonne qualité que la génération de traductions classiques, se révèle être un facteur d'amélioration du lexique pour l'anglais → français mais l'apport n'est pas évident pour l'allemand. Nous faisons l'hypothèse que ceci est dû au type morphologique des langues mais cela reste à vérifier en menant des expériences sur d'autres langues. D'une façon générale, les traductions fertiles amènent des propositions de traductions originales et qui complètent idéalement la traduction classique en proposant une variante vulgarisée, s'apparentant souvent à une paraphrase. Ces traductions sont autant d'attestations d'usages linguistiques que le traducteur pourra utiliser à sa guise dans son travail. Les extraits 6.1 et 6.2 présentent des exemples de traductions fertiles trouvées par le système accompagnées de contextes d'occurrences pour chaque couple de langues.

6.4.2 Perspectives

Pour finir, l'analyse qualitative nous a permis de dégager les limites de notre approche ainsi que de futures perspectives de travail qui touchent à l'amélioration des ressources linguistiques et au perfectionnement de la méthode de génération de traductions fertiles.

Amélioration des ressources linguistiques

Dictionnaire de synonymes Les diverses expériences ont montré que le dictionnaire des synonymes n'avait pas vraiment d'intérêt. Or, nous avons vu que certains cas de silences sont dus à des cas de semi-compositionnalité où un des composants est traduit par un composant de sens proche bien qu'il ne soit pas sa traduction exacte. Il s'agit de cas de divergence lexicale qui ne peuvent être traités par des synonymes au sens strict. Une piste de recherche pourrait être de recourir à un thésaurus afin d'inclure des mots sémantiquement proches dans les variantes. Une autre piste pourrait être d'acquérir cette liste de mots proches dans les corpus cible et source, en considérant comme sémantiquement proches deux mots apparaissant dans des contextes similaires ou encore en essayant de traduire un mot par les N premiers candidats donnés par la méthode distributionnelle.

Familles morphologiques Concernant les familles morphologiques, il faut noter que l'algorithme de racinisation de Porter (1980) est une méthode assez brutale pour extraire une telle ressource. Dans certains cas, même si deux mots sont apparentés morphologiquement, leurs sens peuvent-être trop éloignés pour que cette paire soit intéressante pour gérer la variation morphologique en traduction. Une solution peut être d'avoir recours à des patrons de variation précis comme le font Morin et Daille (2010)[9]. Nous pensons que pour autant, les familes morphologiques obtenues par racinisation ne doivent pas être éliminées car cela risquerait de faire baisser la couverture de la méthode : il s'agirait juste de donner plus de poids aux variantes morphologiques obtenues par patron.

9. Exemple de règle (Morin et Daille, 2010, p. 86) :
N_1 Adj → N_1 Prep Art$^?$ N_2
$\mathcal{M}$(Adj, N_2) = [-ique, -ie]
$\mathcal{M}$(Adj, N_2) = [-ique, -e]

Anglais → français

– loco-régional → local et régional
  – « ***Loco-regional*** *treatments are potentially curative when disease is confined to the breast and lymph nodes.* »
  – « *A la fin des traitements* ***locaux et régionaux*** *(chirurgie, radiothérapie) et après une chimiothérapie si elle a été administrée.* »

– post-conception → après fécondation
  – « *The first stage of foetal development is implantation, which occurs within 2 weeks* ***post-conception****.* »
  – « ***Après fécondation*** *in vitro (FIV), les embryons obtenus sont congelés et seront réimplantés ultérieurement dans l'utérus de la patiente, à distance de la fin des traitements.* »

– blue-dye → colorer en bleu, colorant bleu
  – « *Motion 4—axillary sampling (****blue-dye*** *guided) (For : Mr. Douglas Macmillan.* »
  – « *Une palpation était également pratiquée et était prélevé tout ganglion suspect même s'il n'était ni radioactif ni* ***coloré en bleu****.* »
  – « *Un traceur radioactif ou un* ***colorant bleu*** *est injecté dans la zone de la tumeur.* »

– randomly → manière randomiser
  – « *In this study 52 women were* ***randomly*** *assigned to oestradiol 0.05 mg/day (n = 26) or placebo dermal patches (n = 26) for 12 weeks.* »
  – « *Chez les femmes traitées pour cancer du sein, la voie percutanée (0,1 mg/j) a été la première testée de* ***manière randomisée*** *en double aveugle contre placebo avec crossover chez 110 patientes [27].* »

Extrait 6.1 – Exemples de traductions fertiles trouvées anglais → français

Anglais → allemand

– hypercalcaemia → zu viel calcium in das blut
  – « *Hypercalcaemia (excessive calcium in the blood)* »
  – « *Kann auch* ***zu viel Calcium ins Blut*** *kommen im Sinne einer Überdosierung durch tägliche Calcium-Tabletten Einnahme als Vorsorge ?* » *'Est-il possible qu'il y ait trop de calcium dans le sang et que cette overdose soit provoquée par une prise quotidienne, à titre préventif, de comprimés de calcium ?'*

– tumour-free → frei von tumor
  – « *Nava emphasizes that oncoplastic procedures often involve wide resections which increase the chance of* ***tumour-free*** *margins.* »
  – « *Resektionsrand ist* ***frei von Tumor****gewebe* » *'La marge de résection est exempte de tissu tumoral'*

– non-invasive → nicht invasiv
  – « *Because the cancer cells have not developed the ability to spread, you may hear DCIS described as a pre-cancerous, intraductal or* ***non-invasive*** *cancer.* »
  – « *Bei fast allen Tumortypen liegt auch eine* ***nicht invasive*** *(duktale oder lobuläre) Tumorkomponente vor, aus der sie hervorgegangen sind und die für die Größe der Operation mitentscheidend ist.* » *'Pour presque tous les types de tumeurs, il existe également une composante (ductale ou lobulaire) non invasive, dont la tumeur est issue, et qui déterminera l'importance de l'opération chirurgicale à effectuer.'*

– post-mastectomy → nach der entfernung der brust
  – « *Disease relapse (local or distant recurrences) occurred in 29 women, in whom 25 were* ***post-mastectomy****.* »
  – « ***Nach der Entfernung*** *der Brust gibt es unterschiedliche Techniken diese wiederherzustellen.* » *'Après une ablation des seins, il existe plusieurs techniques de reconstruction mammaire.'*

Extrait 6.2 – Exemples de traductions fertiles trouvées anglais → allemand

Couverture des ressources bilingues Nous avons également noté qu'un tiers des silences étaient dus à l'absence de traduction dans nos ressources bilingues. Une solution pour enrichir le dictionnaire bilingue pourrait être d'extraire des traductions de termes de données parallèles issues d'autres domaines. Une fois ces exemples de traductions extraits, nous pourrions employer des méthodes empiriques pour apprendre des alignements entre morphèmes ou entre chaînes de caractères (Claveau, 2009; Claveau et Kijak, 2011). Les équivalences traductionnelles entre morphèmes pourraient également être apprises à partir des paires de cognats extraites des corpus comparables.

Traductions fertiles

Définition des traductions candidates Un premier problème concerne la définition d'une traduction candidate. Nous pensons qu'une phase de regroupement des traductions proposées serait nécessaire. Actuellement, nous considérons que deux traductions sont identiques si elles correspondent à la même suite de paires `(lemme, partie du discours)`. Autant ceci a du sens lorsque la traduction inclut des mots outils comme des prépositions (*temps* $\underline{de}$ *progression* se distingue bien de *temps* $\underline{sans}$ *progression*) ; autant il faudrait pouvoir neutraliser la variation au niveau des déterminants : *nach der operation 'après l'opération'* et *nach ein operation 'après une opération'* devraient être ramenées à une même forme neutre *nach* DÉTERMINANT *operation*. La question des adverbes d'intensité est plus difficile à trancher car ces derniers impliquent un changement de sens. Par exemple, nous traduisons *prematurely* par *manière plus précoce* et *manière très précoce*. Faut-il les regrouper en une forme neutre du type *manière* ADVERBE *précoce* ? La question reste en suspens.

Insertion des mots outils Un deuxième type de problème concerne la gestion des mots outils que l'on autorise entre chacun des composants traduits ; i.e. lorsqu'on traduit *breast-cancer* en deux mots *cancer* et *sein* et que lors, de la phase de sélection, on considère comme traduction candidate toute suite de mots commençant par *cancer* suivi de zéro à trois mots outils et finissant par *sein*. Si l'inclusion des mots outils permet de ramener des traductions correctes (*cancer* $\underline{du}$ *sein*), elle crée aussi des traductions fausses voire des contre-sens (*cancer* $\underline{en\ dehors\ du}$ *sein*).

Ordre des composants L'ordre des composants peut causer des erreurs dans le cadre de la traduction classique mais c'est surtout avec les traductions fertiles qu'il est problématique. Par exemple, pour *breast-cancer*, le système va proposer des traductions comme *sein et cancer, sein mais sans cancer, sein après un cancer*. La méthode de génération actuelle utilise en quelque sorte une approche par force brute : nous générons toutes les permutations pour être sûrs d'avoir généré le bon ordre de composants et seule l'attestation d'une des permutations dans les textes cibles nous permet de la considérer comme traduction potentielle. Cette approche pourrait être améliorée. Une première solution pourrait être d'apprendre un modèle de langue et d'associer une probabilité à chaque traduction générée. Nous pourrions également spécifier des patrons de traduction (ex. : NOM_1+$\text{NOM}_2 \rightarrow \text{NOM}_2$ PRÉPOSITION DÉTERMINANT$^?$ NOM_1). Comme pour les familles morphologiques, ces patrons ne remplaceraient pas la méthode de génération actuelle mais auraient plus de poids et les traductions produites via ces patrons seraient favorisées.

Plusieurs perspectives d'amélioration viennent d'être proposées dans ce chapitre et nous

voyons que se profile peu à peu l'idée de favoriser certaines traductions au détriment d'autres, parce que, par exemple, elles auraient été produites avec une ressource plus fiable (ex. : patron plutôt que génération par la force brute) ou parce que nous aurions des informations sur la probabilité de rencontrer le terme cible dans la langue cible (modèle de langue). Nous savons aussi, grâce à nos expériences présentées dans le chapitre 2, que les traducteurs n'apprécient pas d'avoir trop de traductions candidates. Il est donc nécessaire de pouvoir ordonner les traductions candidates de façon à faire remonter la meilleure traduction en première position. Cette problématique est l'objet du chapitre suivant.

Chapitre 7

Formalisation et évaluation de l'ordonnancement de traductions candidates

Sommaire

Introduction

Ce chapitre décrit des travaux exploratoires visant à expérimenter plusieurs méthodes d'ordonnancement des traductions candidates. Nous commençons par détailler les critères d'ordonnancement choisis (7.1) puis nous indiquons la façon dont nous les combinons (7.2). Les résultats sont donnés en section 7.3. Les limites de notre travail ainsi que les perspectives de recherches sont discutées en section 7.4

7.1 Critères d'ordonnancement

Nous avons testé quatre critères d'ordonnancement :

1. La similarité entre les contextes du terme source et les contextes du terme cible (section 7.1.1)
2. La fréquence du terme cible (section 7.1.2)
3. La probabilité de traduction des parties du discours (section 7.1.3)
4. La fiabilité des modes de traduction utilisés pour traduire les composants du terme source (section 7.1.4)

7.1.1 Similarité des contextes

Ce critère d'ordonnancement, que nous notons C, correspond au score de similarité obtenu avec la méthode distributionnelle directe (section 1.2.1) et se base donc sur la même hypothèse : plus deux termes tendent à apparaître dans des contextes similaires, plus il est possible qu'ils aient un sens proche et qu'ils soient des traductions l'un de l'autre.

L'implantation de la méthode est identique à celle décrite dans la section 1.3.1 : la taille des contextes est de 5 mots à droite et à gauche de la tête du vecteur et la normalisation du nombre de co-occurrences est faite avec le taux de vraisemblance (cf. annexe p. 192). La traduction des vecteurs est faite à l'aide du dictionnaire généraliste bilingue (cf. section 5.5.1). Une traduction fertile est une unité polylexicale : son vecteur de contexte correspond à un vecteur moyen calculé à partir des vecteurs de chacun des mots lexicaux qui la composent.

La similarité entre le vecteur du terme source s et le vecteur de sa traduction candidate t est calculée avec le jaccard pondéré :

$$C(s,t) = \frac{\Sigma_{m \in s \cap t} min(TV(s,m_i), TV(t,m_i))}{\Sigma_{m_i \in s \cup t} max(TV(s,m_i), TV(t,m_i))} \tag{7.1}$$

où $TV(x,m_i)$ est le nombre de co-occurrences normalisé (taux de vraisemblance) entre le terme x et le mot de contexte m_i.

7.1.2 Fréquence du terme cible

Avec ce critère (noté F), nous faisons l'hypothèse que plus le terme cible est fréquent, plus il est possible qu'il appartienne à la thématique du corpus et donc qu'il soit une bonne traduction.

La fréquence du terme cible t est donnée par :

$$F(t) = \frac{nbocc(t)}{N} \tag{7.2}$$

où $nbocc(t)$ est le nombre d'occurrences de t dans le corpus cible et N le nombre total de mots dans le corpus cible.

7.1.3 Probabilité de traduction des parties du discours

Ici, nous voulons capturer le fait que, par exemple, il est plus probable qu'un nom soit traduit par un nom ou par une suite NOM PRÉPOSITION NOM plutôt que par un adverbe (du moins pour la

traduction de l'anglais vers le français). Ces probabilités de traduction entre parties du discours, notées P, ont été acquises à partir du corpus parallèle EMEA (Tiedemann, 2009). Ce corpus est constitué de textes parallèles appartenant à l'Agence Européenne des Médicaments. Les textes appartiennent au domaine médical. Ces textes sont alignés au niveau phrastique et disponibles en ligne au format TMX [1]. Les alignements sous-phrastiques ont été réalisés avec le logiciel d'alignement de LINGUA ET MACHINA qui correspond à une implantation de l'algorithme ANYMALIGN (Lardilleux, 2010).

Nous avons segmenté en mots, lemmatisé et étiqueté les textes avec l'analyseur XELDA puis extrait les alignements sous-phrastiques. Nous avons obtenu une table d'alignements A dans laquelle chaque alignement $a \in A = \{lem_s, pos_s, lem_t, pos_t, p(s|t), p(t|s)\}$ où lem_s, respectivement lem_t, sont le(s) lemme(s) du segment sous-phrastique source, respectivement cible ; pos_s, respectivement pos_t, sont le(s) partie(s) du discours du segment sous-phrastique source, respectivement cible ; $p(s|t)$, respectivement $p(t|s)$, est la probabilité de traduction du segment cible vers le source, respectivement source vers le cible.

La probabilité qu'une traduction candidate ayant le(s) partie(s) du discours y soit la traduction d'un terme source ayant la partie du discours x correspond à :

$$P(y|x) = \frac{\sum_{a \in A|pos_s=x, pos_t=y} p(t|s)}{\sum_{a \in A|pos_s=x} p(t|s)} \tag{7.3}$$

Pour calculer les probabilités de traduction, nous n'avons retenu que les alignements dans lesquels une unité lexicale source était alignée avec une ou plusieurs unités lexicales cibles. Pour les deux couples de langues, nous avons rencontré des unités lexicales cibles d'au maximum cinq mots (mots outils et mots lexicaux) [2].

À partir d'un corpus anglais-allemand de 363 982 phrases alignées, nous avons acquis des probabilités de traduction pour 108 612 paires de suites de parties du discours. Pour l'anglais-français, nous avons acquis environ 191 854 paires de suites de parties du discours. Le corpus de départ contenait 373 127 phrases alignées. Un extrait du lexique final est donné dans l'annexe B.3.6.

7.1.4 Mode de traduction des composants

Comme certains modes de traduction d'un composant sont plus fiables que d'autres, nous avons défini un critère de fiabilité noté M qui prend en compte la façon dont a été traduit chacun des composants du terme source. Par exemple, on peut supposer qu'un composant traduit par le dictionnaire généraliste aura une traduction plus correcte qu'un composant traduit avec le dictionnaire de cognats.

Nous distinguons dix modes de traduction :

- Le composant correspond à un mot :
  - le mot est traduit directement :
    - via le dictionnaire généraliste (mode **DICO**) ;
    - via le dictionnaire de cognats (mode **COGN**).
  - le mot est traduit indirectement :

1. http ://opus.lingfil.uu.se/EMEA.php
2. Les traductions candidates générées en français font au maximum cinq mots également. En ce qui concerne les traductions candidates en allemand, quatre d'entre elles ont plus de cinq mots, ce qui représente 0,19% de la totalité des traductions générées en allemand.

  - via le dictionnaire généraliste et les familles morphologiques (mode **MORPHO**) ;
  - via le dictionnaire généraliste et le dictionnaire de synonymes (mode **SYNO**) ;
  - via le dictionnaire de cognats et les familles morphologiques (mode **MORPHOCOGN**) ;
  - via le dictionnaire de cognats et le dictionnaire de synonymes (mode **SYNOCOGN**).
- Le composant correspond à un morphème lié, il est traduit via la table de traduction des morphèmes :
  - préfixe traduit par un préfixe (mode **PREF**) ;
  - confixe traduit par un confixe (mode **CONF**) ;
  - suffixe traduit par un suffixe (mode **SUFF**) ;
  - préfixe, confixe ou suffixe traduit par un mot (mode **FERT**).

La traduction d'un terme source peut avoir été générée de plusieurs manières. Par exemple, la traduction $façon\ anormal$ a été générée de quatre manières différentes comme indiqué dans le tableau 7.1 :

	décomposition	traduction	recomposition
1	{ab, normal, ly}	{a:PREF, normal:DICO, façon:FERT}	{façon, anormal}
2	{ab, normal, ly}	{a:PREF, normal:COGN, façon:FERT}	{façon, anormal}
3	{abnormal, ly}	{anormal:DICO, façon:FERT}	{façon, anormal}
4	{abnormal, ly}	{anormal:COGN, façon:FERT}	{façon, anormal}

TABLE 7.1 – Exemple de traduction candidate issue de multiples générations

- La génération 1 a découpé $abnormally$ en ab-, normal et -ly. Le préfixe ab- a été traduit par le préfixe a-, le mot normal a été traduit grâce au dictionnaire généraliste en normal et le suffixe -ly a été traduit par le mot façon.
- La génération 2 a suivi le même processus si ce n'est que la traduction de normal a été obtenue via le dictionnaire de cognats.
- Dans la génération 3, le terme source a été découpé en abnormal et -ly, abnormal a été traduit via le dictionnaire généraliste et -ly a été traduit par façon.
- Dans la quatrième génération, abnormal a été traduit par le dictionnaire de cognats.

L'étape de recomposition donne le même patron de terme cible : façon suivi de anormal. Ce patron est recherché dans le corpus cible et nous obtenons la traduction façon/NOM anormal/ADJECTIF (« *...des gènes protecteurs contre le développement de tumeurs sont réduits au silence de* ***façon anormale***. »).

Quel que soit le nombre de générations ayant permis d'obtenir le terme cible $façon\ anormal$, au final, nous ne comptons qu'une et une seule traduction : $abnormally \rightarrow façon\ anormal$. Ce phénomène de génération multiple est beaucoup plus marqué pour le français (4,27 générations différentes par traduction) que pour l'allemand (2,43).

Lorsque nous calculons le critère M pour un terme cible t, nous prenons en compte tous les modes de traduction utilisés par toutes les générations qui ont permis d'obtenir le terme cible :

$$M(t) = \frac{\sum_{g \in G(t)} \sum_{c \in g} fiabilite(m(c))}{\sum_{g \in G(t)} |c \in g|} \tag{7.4}$$

où $G(t)$ est l'ensemble des générations ayant donné t, chaque $c \in g$ est un des composants de la génération g et $fiabilite(m(c))$ est la fiabilité du mode de traduction m de c.

Dans notre exemple, si $PREF = 0,6$; $DICO = 0,5$; $COGN = 0,6$; et $FERT = 0,4$; alors, le score de fiabilité de $façon\ anormal$ est de $0,5$:

$$M(\text{façon anormal}) = \frac{2 \times 0,6 + 2 \times 0,5 + 2 \times 0,6 + 4 \times 0,4}{10} \tag{7.5}$$

La fiabilité de chaque mode de traduction a été calculée sur notre jeu de données d'entraînement T décrit en section 5.4. Pour un mode de traduction m, sa fiabilité est donnée par :

$$fiabilite(m) = \frac{|\{g : m \in g, \mathcal{A}(t(g)) \in \{\text{EXACT}\}\}|}{|\{g : m \in g\}|} \tag{7.6}$$

où $\{g : m \in g\}$ sont toutes les générations qui ont utilisé le mode de construction m et $\mathcal{A}(t(g))$ est l'annotation du terme cible t donné par la génération g.

Les valeurs obtenues sont données dans le tableau 7.2.

On observe que les composants traduits avec le dictionnaire généraliste sont moins fiables que ceux traduits avec le dictionnaire de cognats, quelle que soit la langue. Ceci confirme une fois de plus l'intérêt d'enrichir le dictionnaire bilingue avec des cognats identifiés dans le corpus comparable.

Les composants traduits de façon indirecte (ressource bilingue + ressource de variation) sont parmi les moins fiables. La combinaison cognats et variantes morphologiques est la plus sûre, suivie par la combinaison dictionnaire généraliste + synonymes, puis la combinaison dictionnaire généraliste + variantes morphologiques. La combinaison cognats + synonymes est le mode de traduction le moins fiable des dix modes de traduction, en particulier pour l'allemand.

Les modes de traduction les plus fiables sont les traductions non fertiles des confixes et des suffixes. Les préfixes sont également très fiables pour l'allemand mais pas pour le français. La faible fiabilité des préfixes pour le français est surprenante, d'autant plus que les expériences décrites en section 6.3.2 indiquaient une bonne précision pour la traduction des mots préfixés. En analysant des sorties du système, nous avons observé qu'en français, la ou les bases lexicales contenues dans les termes avec préfixe sont dans leur grande majorité (71 %) traduites en utilisant des ressources de variation, ce qui peut faire baisser la qualité de la traduction. Ce phénomène est moins marqué en allemand (66 %).

La traduction fertile est le deuxième mode de traduction le moins fiable, particulièrement pour l'allemand.

7.2 Combinaison de critères

7.2.1 Standardisation des valeurs

Nos critères d'ordonnancement ont tous une valeur entre 0 et 1 mais ils présentent des échelles de valeur très différentes. Par exemple, pour le critère de la fréquence, les valeurs sont très basses ($<$ 0,005) alors que pour la fiabilité des modes de traductions, les valeurs varient entre 0,19 et 0,92. Or si nous combinons les valeurs brutes, les critères avec des valeurs très basses seront complètement occultés par les critères avec des valeurs plus hautes.

Nous avons standardisé nos valeurs en suivant la méthode décrite par Gendre (1977, p. 48-50) : les valeurs observées sont remplacées par leur percentile puis le percentile est transformé en score-z à l'aide de la table de la loi normale. Les détails de la standardisation sont donnés dans l'annexe A.4, p. 194.

	EN-FR	EN-DE
DICO	0,49	0,38
COGN	0,57	0,48
MORPHO	0,34	0,29
SYNO	0,43	0,32
MORPHOCOGN	0,41	0,37
SYNOCOGN	0,21	0,05
PREF	0,61	0,79
CONF	0,79	0,67
SUFF	0,63	0,92
FERT	0,37	0,19

TABLE 7.2 – Fiabilité des modes de traduction

7.2.2 Combinaison linéaire

Nos premières expériences ont consisté à combiner chacun des critères :

$$score(t) = \alpha C + \beta F + \gamma P + \delta M \tag{7.7}$$

où α, β, γ et δ sont des coefficients pondérateurs ($0 \leq \alpha, \beta, \gamma, \delta \leq 1$ et $\alpha + \beta + \gamma + \delta = 1$).

Nous avons réalisé deux expériences. Dans la première, tous les coefficients pondérateurs se valent : nous considérons que les critères sont autant informatifs les uns que les autres. Dans la seconde expérience, nous avons attribué un poids différent à chaque critère : ici, nous considérons que certains critères sont plus informatifs que d'autres, ils doivent donc avoir un poids plus fort. Les poids ont été appris automatiquement sur le jeu de données d'entraînement T (cf. section 5.4) en utilisant l'algorithme 6 : nous avons simplement testé plusieurs jeux de poids possibles et retenu celui qui donnait la meilleure précision sur le Top1.

Les poids obtenus sont consultables dans le tableau 7.3. Fréquence du terme cible et partie du discours ont des poids équivalents. Le critère ayant reçu le meilleur poids est celui des modes de traduction.

La similarité des contextes est le critère ayant reçu le plus petit poids. Si ce critère est le moins informatif, c'est probablement parce que les termes sources et cibles sont généralement peu fréquents : entre 73 % et 81 % ont 5 occurrences ou moins dans le corpus (cf. tableaux 7.11 et 7.12). Ceci montre l'intérêt d'avoir recours à la traduction compositionnelle plutôt qu'à l'approche distributionnelle pour les termes complexes.

Comme l'atteste l'écart moyen calculé pour chaque coefficient [3], les valeurs des coefficients sont stables quel que soit le couple de langues. Ceci semble indiquer que l'importance à donner à chaque critère d'ordonnancement est indépendante des langues en jeu dans la traduction. Il serait intéressant de vérifier cette hypothèse sur d'autres couples de langues. Si l'hypothèse s'avérait exacte, il serait alors possible d'utiliser des données parallèles dans des langues bien dotées pour apprendre les coefficients pondérateurs et les appliquer à des langues peu dotées.

3. Moyenne des valeurs absolues des écarts à la moyenne : $\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$.

Il serait également intéressant de voir si les coefficients sont stables pour un même couple de langues mais sur des domaines de spécialité différents.

Critère	EN-FR	EN-DE	écart moyen
Contexte C (coeff. α)	0,12	0,15	0,015
Fréquence F (coeff. β)	0,24	0,22	0,010
Partie du discours P (coeff. γ)	0,25	0,26	0,005
Modes de traduction M (coeff. δ)	0,39	0,37	0,010

TABLE 7.3 – Poids accordés à chaque critère d'ordonnancement

7.2.3 Apprentissage d'un modèle d'ordonnancement

Dans un deuxième temps, nous avons expérimenté des algorithmes de *learning-to-rank*. Pour cela, nous nous sommes basés sur la librairie RankLib[4] qui met à disposition des implantations de divers algorithmes de *learning-to-rank*. Parmi les algorithmes implantés, nous avons sélectionné ceux qui appartiennent à la famille des algorithmes *list-wise* : ADARANK (Li et Xu, 2007), COORDINATE ASCENT (Metzler et Croft, 2000) et LAMBDA MART (Wu *et al.*, 2010).

ADARANK et LAMBDA MART sont basés sur la technique de *boosting*. Le boosting consiste à combiner les résultats donnés par plusieurs modèles habituellement plus simples et moins performants (pris isolément) qu'un modèle général. Ces modèles sont appris un par un, le poids de chaque exemple du jeu de données étant réévalué en fonction des erreurs du modèle appris précédemment (les exemples mal classifiés voient leur poids augmenter ; les exemples bien classifiés voient leur poids diminuer). COORDINATE ASCENT, quant à lui, apprend un modèle linéaire. Il doit son nom à la technique employée pour optimiser le choix des paramètres.

Les variables fournies aux algorithmes sont les quatre critères C, F, P et M. Les valeurs ont été préalablement standardisées en suivant la méthode décrite en section 7.2.1. Les paramètres proposés par la librairie RankLib offrent la possibilité de spécifier une mesure à optimiser pour les algorithmes *list-wise*. Nous avons choisi d'optimiser la précision sur le Top1. Tous les autres paramètres ont été laissés avec leur valeur par défaut. L'apprentissage du modèle a été fait sur le jeu de données T.

7.3 Évaluation

7.3.1 Réference et mesures d'évaluation

Les méthodes d'ordonnancement ont été évaluées sur le jeu de données E (décrit en section 5.4) qui correspond aux termes sources de la référence *a priori* pour lesquels le système a pu générer une traduction. Les données d'entraînement (utilisées pour l'apprentissage des scores de fiabilité, des poids de la combinaison linéaire et des modèles d'ordonnancement) correspondent au jeu de données T, c'est-à-dire les termes sources n'appartenant pas à la référence *a priori* et pour lesquels le système a pu générer une traduction. Les ensembles de données E et T sont disjoints.

4. http://people.cs.umass.edu/ vdang/ranklib.html

Algorithme 6 Trouver les meilleurs poids

Require: $training\_data$
  $step \leftarrow 0.01$
  $best\_precision \leftarrow 0$
  $best\_weight\_set \leftarrow \emptyset$
  for $\alpha = 0\,;\, \alpha <= 1\,;\, \alpha = \alpha + step$ **do**
    for $\beta = 0\,;\, \beta <= 1\,;\, \beta = \beta + step$ **do**
      for $\gamma = 0\,;\, \gamma <= 1\,;\, \gamma = \gamma + step$ **do**
        for $\delta = 0\,;\, \delta <= 1\,;\, \delta = \delta + step$ **do**
          if not $\alpha + \beta + \gamma + \delta == 1$ **then**
            continue
          end if
          $exact \leftarrow 0$
          $total \leftarrow 0$
          for all $source\_term, translations$ **in** $training\_data$ **do**
            $ranked\_translations \leftarrow$ rank($translations$, $\{\alpha, \beta, \gamma, \delta\}$)
            if $ranked\_translations$[0] is EXACT **then**
              $exact+ = 1$
            end if
            $total+ = 1$
          end for
          $precision = exact/total$
          if $precision > best\_precision$ **then**
            $best\_precision \leftarrow precision$
            $best\_weight\_set \leftarrow \{\alpha, \beta, \gamma, \delta\}$
          end if
        end for
      end for
    end for
  end for
  return $best\_weight\_set$

La mesure d'évaluation est la précision sur le $TopN$, soit la fraction de termes sources qui ont au moins une traduction correcte parmi les N premières traductions candidates :

$$TopN = \frac{1}{|S|} \sum_{j=1}^{|S|} \alpha(T_{j^N}, R_j) \tag{7.8}$$

$$\alpha(T_{j^N}, R_j) = \begin{cases} 1 & \text{si } T_{j^N} \cap R_j \neq \emptyset \\ 0 & \text{sinon} \end{cases}$$

où :
– S est l'ensemble des termes sources
– T_{j^N} est l'ensemble des N premières traductions candidates pour le terme source j
– R_j est l'ensemble des traductions correctes pour le terme source j

La définition d'une traduction "correcte" peut varier : soit ce sont uniquement les traductions données par l'UMLS (dans ce cas, la précision est notée P) ; soit ce sont les traductions de l'UMLS ou les traductions annotées EXACT par les traducteurs (précision notée P_E) ; soit ce sont les traductions de l'UMLS ou les traductions annotées EXACT ou ACCEPTABLE (précision notée P_{EA}).

7.3.2 Bases de comparaison

Nous avons utilisé six bases de comparaison. La base de comparaison basse (ALÉATOIRE) correspond à la précision obtenue lorsque les traductions sont ordonnées aléatoirement (la précision indiquée est une moyenne sur 100 ordonnancements aléatoires). La base de comparaison haute (MEILLEURE PRÉCISION POSSIBLE) correspond à la précision qui serait obtenue si la meilleure traduction était toujours placée au rang 1. Les quatres autres bases de comparaison C, F, P et M correspondent à chacun des quatres critères utilisés séparément. Afin d'appréhender la difficulté de la tâche, nous indiquons dans la figure 7.1 le nombre de traductions par terme source pour chaque couple de langues.

Pour les traductions anglais → français, on observe que presque les deux-tiers (64 %) des termes sources n'ont qu'une seule traduction candidate. 13 % d'entre eux ont deux traductions candidates et 23 % d'entre eux ont plus de deux traductions candidates. Le nombre maximum de traductions candidates est 13.

Pour les traductions anglais → allemand, la tâche est un peu plus difficile : seulement une petite moitié (46 %) des termes sources ont une seule traduction candidate, un quart d'entre eux (24 %) ont deux traductions candidates et 29 % ont plus de deux traductions candidates. Le nombre maximum de traductions candidates est 28.

7.3.3 Résultats obtenus

Nous donnons, pour chaque couple de langues, la précision obtenue sur le Top 1 à 5. Nous indiquons également le "Rang précision maximum", c'est-à-dire le rang à partir duquel la meilleure précision possible a été atteinte.

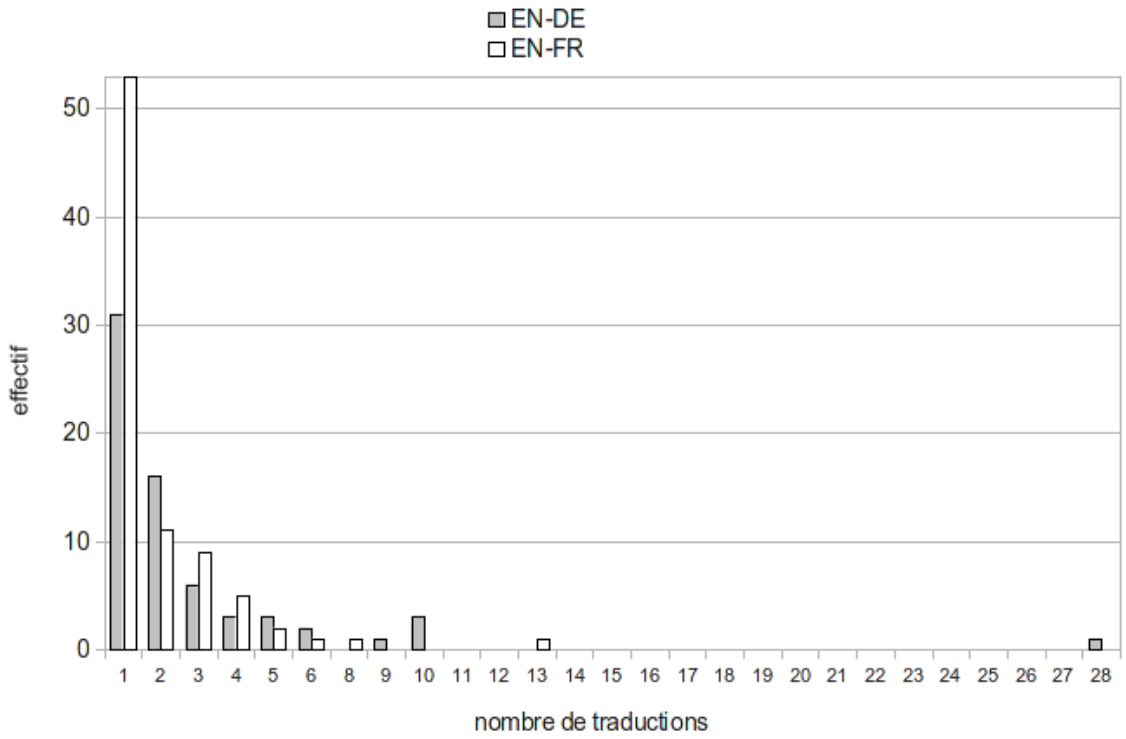


FIGURE 7.1 – Nombre de traductions par terme source

7.3.3.1 Ordonnancement des traductions anglais → français

Les résultats sont donnés dans les tableaux 7.4 (P), 7.5 (P_E) et 7.6 (P_{EA}). Les méthodes sont ordonnées par leur précision sur le Top1 décroissante puis le Top2, puis le Top3.

On observe que, quelle que soit la définition que l'on a d'une traduction correcte (UMLS, EXACT, ACCEPTABLE...), les méthodes qui se détachent sont la COMBINAISON NON PONDÉRÉE et la COMBINAISON PONDÉRÉE suivies de COORDINATE ASCENT, LAMBDA MART et le critère M. Nous notons aussi que toutes les méthodes testées font systématiquement mieux que le classement aléatoire. C'est particulièrement visible avec les traductions de l'UMLS seules : la moins bonne des méthodes (critère C isolé) fait 9 points de plus que l'aléatoire sur le Top1 et les meilleures méthodes font jusqu'à 10,5 points de plus que l'aléatoire (toujours sur le Top1). Évidemment, plus on descend dans les Top, plus l'écart se resserre avec l'ordonnnancement aléatoire. Il en est de même pour la souplesse que l'on donne à la définition d'une traduction correcte : plus on est souple, moins l'ordonnancement a d'intérêt.

L'ordonnancement semble être surtout pertinent lorsque l'on cherche à obtenir une traduction UMLS ou EXACT sur le Top1 ou Top2. Comme les listes des traductions candidates sont généralement courtes (seulement 23 % avec plus de deux traductions candidates l'anglais → français ; 29 % pour l'allemand), on atteint assez vite la meilleure précision possible. Pour les traductions de l'UMLS uniquement (P), la meilleure précision possible est obtenue au Top3. Ceci

est encore plus marqué lorsque l'on prend en compte des traductions EXACT ou ACCEPTABLE. Pour les traductions UMLS et EXACT (P_E), la meilleure précision possible est obtenue au Top2. Pour les traductions UMLS, EXACT et ACCEPTABLE (P_{EA}), la meilleure précision possible est obtenue au Top1.

L'apprentissage de poids pour la combinaison linéaire semble ne pas présenter d'intérêt : les résultats sont quasi-équivalents (+ 1,2 à -1,2 point). Parmi, les méthodes de *learning-to-rank*, ADARANK se démarque comme étant moins performante que les deux autres. D'une façon générale, les méthodes s'appuyant sur des données d'entraînement ne se démarquent pas comme étant meilleures que la simple combinaison non pondérée des quatres critères.

Si on compare chacun des critères utilisés isolément, le critère M (fiabilité des modes de traduction) est celui qui donne les meilleurs résultats (il s'est d'ailleurs vu attribuer le poids le plus important lors de l'apprentissage : 0,39). La similarité des contextes (critère C) est clairement le moins bon critère, il avait d'ailleurs reçu le plus petit poids (0,12). Malgré tout, il reste meilleur que l'ordonnancement aléatoire : il permet d'ailleurs d'obtenir la meilleure précision au rang 4 au lieu du rang 13 pour l'aléatoire. Pour finir, le critère de fréquence F semble être un peu plus intéressant que celui de la probabilité de traduction des parties du discours (critère P).

7.3.3.2 Ordonnancement des traductions anglais → allemand

Les résultats sont donnés dans les tableaux 7.7 (P), 7.8 (P_E) et 7.9 (P_{EA}). Les méthodes ont également été ordonnées en fonction leur précision sur le Top1 puis Top2 puis Top3.

Pour l'allemand, nous avons une belle marge de progression puisqu'il y a beaucoup plus de traductions candidates par terme source. Toutes les méthodes testées se comportent mieux que l'ordonnancement aléatoire. La moins bonne méthode obtient de 5 à 1,2 points de plus que l'ordonnancement aléatoire sur le Top1 selon que l'on prend en compte les traductions de l'UMLS, EXACT ou ACCEPTABLE.

Ici, la combinaison pondérée fait systématiquement mieux que la combinaison non pondérée (4,5 à 1,5 points de plus sur le Top1 selon le type de traduction "correcte" choisi). Comme pour l'ordonnancement anglais → français, COORDINATE ASCENT, LAMBDA MART et la COMBINAISON NON PONDÉRÉE font partie des méthodes les plus intéressantes. Néanmoins, avec les traductions de l'UMLS, COORDINATE ASCENT, LAMBDA MART et la COMBINAISON NON PONDÉRÉE ne sont réellement intéressants qu'à partir du Top2 (P=0,682). Au Top1, elles sont devancées par des méthodes qui sont généralement moins bonnes par ailleurs (critère P et ADARANK). Le critère M fondé sur les modes de traduction obtient aussi de bons résultats à partir du Top1 (P) et du Top2 (P_E et P_{EA}).

Enfin, les critères P et C obtiennent les moins bons résultats. Même si P a une bonne précision sur le Top1 (notamment avec les traductions UMLS), il faut attendre le rang 17 pour atteindre la précision maximum. Avec le critère C, qui est le moins bon de tous, il faut attendre le rang 28.

7.4 Discussion

Nous concluons ce chapitre par une discussion générale sur les résultats obtenus. Nous présentons tout d'abord un bilan de nos expérimentations puis nous proposons plusieurs perspectives de recherche.

	Précision (UMLS)					Rang précision
	Top1	Top2	Top3	Top4	Top5	maximum
MEILLEURE PRÉCISION POSSIBLE	,795	,795	,795	,795	,795	1
COMBINAISON PONDÉRÉE	**,759**	**,783**	**,795**	**,795**	**,795**	**3**
COMBINAISON NON PONDÉRÉE	,747	**,783**	**,795**	**,795**	**,795**	**3**
COORDINATE ASCENT	,747	**,783**	**,795**	**,795**	**,795**	**3**
LAMBDA MART	,747	**,783**	**,795**	**,795**	**,795**	**3**
M	,747	,771	**,795**	**,795**	**,795**	**3**
F	,723	,783	**,795**	**,795**	**,795**	**3**
ADARANK	,687	,759	,783	**,795**	**,795**	4
P	,687	,759	,783	**,795**	**,795**	4
C	,663	,747	,783	**,795**	**,795**	4
ALÉATOIRE	,654	,733	,772	,783	,787	13

TABLE 7.4 – Résultats ordonnancement anglais → français (P)

	Précision (UMLS ou EXACT)					Rang précision
	Top1	Top2	Top3	Top4	Top5	maximum
MEILLEURE PRÉCISION POSSIBLE	,94	,94	,94	,94	,94	1
COMBINAISON NON PONDÉRÉE	**,928**	**,94**	**,94**	**,94**	**,94**	**2**
COMBINAISON PONDÉRÉE	**,928**	**,94**	**,94**	**,94**	**,94**	**2**
COORDINATE ASCENT	**,928**	**,94**	**,94**	**,94**	**,94**	**2**
LAMBDA MART	**,928**	**,94**	**,94**	**,94**	**,94**	**2**
M	**,928**	**,94**	**,94**	**,94**	**,94**	**2**
F	,916	,928	**,94**	**,94**	**,94**	3
ADARANK	,892	,904	,928	**,94**	**,94**	4
P	,892	,904	,928	**,94**	**,94**	4
C	,88	,904	,928	**,94**	**,94**	4
ALÉATOIRE	,836	,898	,928	,931	,932	13

TABLE 7.5 – Résultats ordonnancement anglais → français (P_E)

	Précision (UMLS ou EXACT ou ACCEPTABLE)					Rang précision maximum
	Top1	Top2	Top3	Top4	Top5	
MEILLEURE PRÉCISION POSSIBLE	,952	,952	,952	,952	,952	1
COMBINAISON NON PONDÉRÉE	**,952**	**,952**	**,952**	**,952**	**,952**	**1**
COORDINATE ASCENT	**,952**	**,952**	**,952**	**,952**	**,952**	**1**
COMBINAISON PONDÉRÉE	,94	**,952**	**,952**	**,952**	**,952**	2
LAMBDA MART	,94	**,952**	**,952**	**,952**	**,952**	2
M	,94	**,952**	**,952**	**,952**	**,952**	2
ADARANK	,94	,94	**,952**	**,952**	**,952**	3
F	,94	,94	**,952**	**,952**	**,952**	3
P	,94	,94	**,952**	**,952**	**,952**	3
C	,928	,94	**,952**	**,952**	**,952**	3
ALÉATOIRE	,917	,942	,949	,95	,951	8

TABLE 7.6 – Résultats ordonnancement anglais → français (P_{EA})

	Précision (UMLS)					Rang précision maximum
	Top1	Top2	Top3	Top4	Top5	
MEILLEURE PRÉCISION POSSIBLE	,697	,697	,697	,697	,697	1
COMBINAISON PONDÉRÉE	**,621**	,667	**,697**	**,697**	**,697**	**3**
M	**,621**	,652	**,697**	**,697**	**,697**	**3**
ADARANK	**,621**	,652	,652	,667	,667	17
P	**,621**	,652	,652	,667	,667	17
LAMBDA MART	,606	**,682**	**,697**	**,697**	**,697**	**3**
C	,606	,667	,667	,682	,682	28
F	,591	,667	**,697**	**,697**	**,697**	**3**
COMBINAISON NON PONDÉRÉE	,576	**,682**	**,697**	**,697**	**,697**	**3**
COORDINATE ASCENT	,576	**,682**	**,697**	**,697**	**,697**	**3**
ALÉATOIRE	,526	,627	,654	,665	,672	28

TABLE 7.7 – Résultats ordonnancement anglais → allemand (P)

	Précision (UMLS ou EXACT)					Rang précision maximum
	Top1	Top2	Top3	Top4	Top5	
MEILLEURE PRÉCISION POSSIBLE	,879	,879	,879	,879	,879	1
COMBINAISON PONDÉRÉE	**,848**	**,879**	**,879**	**,879**	**,879**	**2**
LAMBDA MART	**,848**	,864	,864	,864	**,879**	**5**
COMBINAISON NON PONDÉRÉE	,833	,864	**,879**	**,879**	**,879**	**3**
COORDINATE ASCENT	,833	,864	**,879**	**,879**	**,879**	**3**
F	,833	,848	**,879**	**,879**	**,879**	**3**
ADARANK	,833	,848	,848	,848	,848	17
P	,833	,848	,848	,848	,848	17
M	,818	,864	**,879**	**,879**	**,879**	**3**
C	,803	,864	,864	,864	,864	28
ALÉATOIRE	,77	,832	,846	,853	,857	28

TABLE 7.8 – Résultats ordonnancement anglais → allemand (P_E)

	Précision (UMLS ou EXACT ou ACCEPTABLE)					Rang précision maximum
	Top1	Top2	Top3	Top4	Top5	
MEILLEURE PRÉCISION POSSIBLE	,894	,894	,894	,894	,894	1
COMBINAISON PONDÉRÉE	**,879**	**,894**	**,894**	**,894**	**,894**	**2**
LAMBDA MART	**,879**	,879	,879	,879	**,894**	5
COMBINAISON NON PONDÉRÉE	,864	**,894**	**,894**	**,894**	**,894**	**2**
COORDINATE ASCENT	,864	**,894**	**,894**	**,894**	**,894**	**2**
M	,864	,879	**,894**	**,894**	**,894**	3
ADARANK	,864	,879	,879	,879	,879	17
P	,864	,879	,879	,879	,879	17
F	,848	,864	**,894**	**,894**	**,894**	3
C	,818	,879	,879	,879	,879	28
ALÉATOIRE	,806	,859	,872	,877	,88	28

TABLE 7.9 – Résultats ordonnancement anglais → allemand (P_{EA})

7.4.1 Bilan

Intérêt de la combinaison des critères

Globalement, nous observons que quatre méthodes d'ordonnancement se distinguent quel que soit le couple de langues : COMBINAISON NON PONDÉRÉE, COMBINAISON PONDÉRÉE, COORDINATE ASCENT et LAMBDA MART. Ces quatre méthodes sont suivies de près par le score M, qui a d'ailleurs obtenu le plus fort poids lors de l'apprentissage (0,39 pour l'anglais → français et 0,37 pour l'anglais → allemand).

On note aussi que les techniques basées sur un apprentissage ne font pas clairement mieux que la simple combinaison des quatre critères. On observe une différence surtout pour l'allemand entre la COMBINAISON PONDÉRÉE et LAMBDA MART qui obtiennent entre 4,5 et 1,5 points de plus que la combinaison simple (non pondérée).

Par contre, les expériences montrent clairement l'intérêt de combiner divers critères, fût-ce de façon simple : quelle que soit la référence utilisée ou le couple de langues, la combinaison des critères donne de meilleurs ou d'aussi bons résultats que le meilleur critère pris isolément. Une exception est le cas du critère M qui est meilleur sur le Top1 pour l'anglais → allemand (traductions UMLS) mais dès le Top2, la combinaison non pondérée reprend le dessus. Globalement, le critère M est le meilleur de tous les critères pris isolément.

Résultats mitigés des méthodes basées sur l'apprentissage

Nous pensons que l'apport moindre des méthodes basées sur un apprentissage s'explique par le fait que nos données d'entraînement et nos données d'évaluation sont assez différentes. Le tableau 7.10 récapitule le nombre de traductions candidates par terme source, en fonction des couples de langues et selon le jeu de données : entraînement (T) et évaluation (E). Le jeu de données d'entraînement comporte plus de traductions candidates par terme source que le jeu de données d'évaluation (3,04 contre 1,92 pour l'anglais → français ; 3,11 contre 2,83 pour l'anglais → allemand).

Il en est de même pour la fréquence des termes sources et des traductions candidates (tableaux 7.11 et 7.12). Beaucoup plus de termes sources/traductions apparaissent dans les basses fréquences en ce qui concerne le jeu de données d'entraînement. Notons que, d'une façon générale, termes sources et traductions candidates sont peu fréquents. Pour l'anglais → français, entre 63 % et 81 % des termes ont un nombre d'occurrences inférieur ou égal à 5. Pour l'anglais → allemand, ce sont entre 56 % et 80 % des termes sources qui apparaissent 5 fois ou moins dans le corpus. Ce nombre important de termes associés à un petit nombre d'occurrences explique les faibles performances du critère C (l'approche distributionnelle nécessite que les termes à traduire soient fréquents). C'est l'allemand qui a le plus de termes avec des petites fréquences et c'est aussi la langue sur laquelle le critère C a donné les moins bons résultats.

Enfin, les données d'entraînement et les données d'évaluation sont aussi différentes en ce qui concerne la répartition des annotations (tableau 7.13). D'une façon générale, les données d'entraînement comportent beaucoup plus d'exemples de traductions annotées FAUX ou PROCHE qu'il n'y en a dans les données d'évaluation. Corollairement, il y a moins d'exemples de traductions annotées EXACT dans le jeu d'apprentissage. La proportion de traductions ACCEPTABLE reste équivalente.

Toutes ces raisons expliquent en partie que les méthodes basées sur l'apprentissage n'aient pas obtenu des résultats nettement supérieurs à ceux obtenus par des méthodes

a priori plus simples. Nous avions initialement choisi d'utiliser la référence *a priori* comme données d'évaluation dans le but de nous conformer aux usages du domaine : l'UMLS est considérée comme la ressource de référence pour la traduction de termes médicaux.

En conclusion, le bilan global est mitigé. D'un côté, nous avons pu montrer l'intérêt de combiner plusieurs critères d'ordonnancement. De l'autre, nous nous attendions à ce que les modèles d'ordonnancement basés sur l'apprentissage donnent de meilleurs résultats.

7.4.2 Perspectives de recherche

Modèles indépendants des langues et des domaines

Une première perspective de recherche serait d'utiliser la totalité des traductions annotées pour évaluer l'apport de l'apprentissage : nous utiliserions alors la validation croisée pour créer plusieurs jeux d'entraînement/évaluation à partir de la référence *a priori*. Toutefois, nous considérons qu'il n'est finalement pas très réaliste d'effectuer l'apprentissage sur des traductions issues du corpus comparable et annotées à la main. En conditions réelles, lorsque l'on a recours aux corpus comparables, c'est justement parce que l'on manque de données parallèles pour le couple de langues et/ou le domaine de spécialité concerné.

Une première piste de recherche sera plutôt d'étudier dans quelle mesure des données parallèles issues d'autres domaines voire d'autres langues peuvent être utilisées comme données d'apprentissage. Par exemple, pour apprendre les parties du discours, nous avons utilisé un corpus dans le domaine médical mais il serait intéressant de voir si ces probabilités de traduction peuvent être apprises sur un corpus d'un autre domaine sans que cela impacte trop la précision. De même, la stabilité des coefficients pondérateurs laisse penser que le poids à accorder à chaque critère est indépendant des langues (section 7.2.2).

Mesures d'évaluation plus adaptées

La mesure d'évaluation employée est aussi discutable. Nous avons utilisé la précision sur le TopN car c'est la mesure d'évaluation la plus fréquemment rencontrée pour l'évaluation de l'extraction de lexiques à partir de corpus comparables. Toutefois, d'autres mesures pourraient être envisagées, surtout pour évaluer des algorithmes d'*ordonnancement*. En effet, le TopN ne rend pas compte de la capacité à bien ordonner les traductions de la plus à la moins utile pour le traducteur. Par exemple, même si l'algorithme place une traduction correcte en première position, il est toujours utile de trouver d'autres traductions correctes juste après plutôt qu'au rang 10 par exemple.

Au chapitre 1, nous avons présenté plusieurs mesures qui pourraient apporter un autre angle de vue sur les résultats de l'ordonnancement. Le Mean Reciprocal Rank indique le rang moyen auquel se situe la première traduction correcte. La MAP indique la précision moyenne obtenue lorsque toutes les traductions correctes ont été sélectionnées. Nous pourrions aussi utiliser le Tau de Kendall (1983) pour évaluer la corrélation entre le rang que le système a attribué à chaque traduction candidate et le rang attribué par des juges.

Identification des traductions correctes

Jusqu'ici, nous avons uniquement cherché à ordonner les traductions candidates pour pouvoir sélectionner la ou les meilleures. Or, cette démarche est vaine si toutes les traductions générées sont incorrectes. Une autre façon de procéder serait d'utiliser les critères de qualité pour directement classifier une traduction comme EXACT, ACCEPTABLE, PROCHE ou FAUX.

Combinaison de différentes méthodes

Nous pensons que les modèles de combinaison de critères pourraient être exploités pour ordonner des traductions obtenues via diverses méthodes : méthode distributionnelle, compositionnelle, empirique, cognats. Ceci permettrait donc de combiner les méthodes afin de tirer le meilleur parti de chacune d'entre elles : nous augmenterions ainsi la taille du lexique extrait sans trop pénaliser sa précision.

En contexte industriel, il est fréquent d'être confronté au besoin de combiner des lexiques ayant plusieurs origines. Les techniques de *learning-to-rank* pourraient être employées pour aggréger plusieurs lexiques. Lorsque les lexiques proposent des traductions différentes pour un même terme source, ces traductions pourraient être ordonnées en fonction de l'origine de la traduction. Par exemple, plus de poids serait donné aux traductions émanant de lexiques constitués manuellement. Les traductions émanant de corpus parallèles auraient plus de poids que celles émanant de corpus comparables.

Recherche de nouveaux indices d'ordonnancement

Pour conclure cette discussion, si l'apport de l'apprentissage automatique reste à prouver pour l'ordonnancement de traductions candidates, la combinaison de critères a montré des résultats encourageants (pour l'anglais → français, la combinaison pondérée obtient 75,9 % de précision sur le Top1 contre 74,7 % pour le meilleur critère pris isolément ; pour l'anglais → allemand, la combinaison non pondérée obtient 68,2 % de précision sur le Top2 contre 65,2 % pour le meilleur critère pris isolément). S'ouvre alors la possibilité d'explorer toute une gamme d'indices de qualité de la traduction : fréquence du terme source, ratio des fréquences du terme source et terme cible, nombre de générations ayant donné le terme cible, partie du discours du terme source, partie du discours du terme cible, nombre d'occurrences dans les textes scientifiques, vulgarisés, etc.

Dans ce chapitre, nous avons fait état de travaux exploratoires visant à améliorer l'ordonnancement des traductions générées par notre système. Ces travaux correspondaient au dernier axe de notre plan d'amélioration de la traduction compositionnelle énoncé en fin de chapitre 3.

Les dernières pages de ce mémoire sont consacrées à un bilan général de notre travail de thèse ainsi qu'à une remise en perspective de nos travaux dans un cadre applicatif.

	EN-FR		EN-DE	
	T	E	T	E
termes sources avec 1 traduction candidate	39 %	64 %	44 %	46 %
termes sources avec 2 traductions candidates	23 %	13 %	21 %	24 %
termes sources avec 3 traductions candidates	14 %	11 %	11 %	9 %
termes sources avec 4 traductions candidates	7 %	6 %	8 %	4 %
termes sources avec 5 traductions candidates et plus	18 %	6 %	17 %	16 %
nb. traduction / terme source	3,04	1,92	3,11	2,83

TABLE 7.10 – Nombre de traductions candidates par terme sources

	Termes sources		Traductions candidates	
	T	E	T	E
# occ. minimum	1	1	1	1
# occ. moyen	5,33	11,42	14,22	10,72
# occ. médian	2	3	2	3
# occ. maximum	116	260	2196	230
% termes avec # occ. $= 1$	47 %	24 %	45 %	34 %
% termes avec # occ. $<= 5$	81 %	66 %	73 %	63 %
% termes avec # occ. $<= 10$	88 %	80 %	81 %	80 %

TABLE 7.11 – Nombre d'occurrences des termes sources et traductions candidates (anglais → français)

	Termes sources		Traductions candidates	
	T	E	T	E
# occ. minimum	1	1	1	1
# occ. moyen	5,25	19,36	14,22	6,74
# occ. médian	2	4	1	2
# occ. maximum	83	260	1248	196
% termes avec # occ. $= 1$	47 %	21 %	51 %	47 %
% termes avec # occ. $<= 5$	77 %	56 %	80 %	80 %
% termes avec # occ. $<= 10$	88 %	68 %	88 %	88 %

TABLE 7.12 – Nombre d'occurrences des termes sources et traductions candidates (anglais → allemand)

	EN-FR		EN-DE	
	T	E	T	E
EXACT	30 %	67 %	24 %	61 %
ACCEPTABLE	10 %	16 %	7 %	6 %
PROCHE	18 %	6 %	14 %	6 %
FAUX	41 %	11 %	55 %	27 %

TABLE 7.13 – Annotations attribuées aux traductions candidates

Conclusion et perspectives

Bilan

Ce travail de thèse avait pour but d'explorer les modes d'exploitation des corpus comparables pour la traduction assistée par ordinateur. Nous poursuivions deux objectifs : d'une part, nous souhaitions observer les lexiques extraits dans une perspective applicative et analyser la façon dont ils étaient appréhendés par les traducteurs afin de dégager des pistes de recherche utiles pour la TAO ; d'autre part, nous voulions mettre en œuvre les pistes de recherche dégagées et tenter de faire progresser les techniques actuelles d'extraction de lexiques à partir de corpus comparables.

Perspective applicative

Pour répondre à notre premier objectif, nous avons commencé par implanter la méthode de référence en extraction de lexiques à partir de corpus comparables. Cette méthode, qui est basée sur l'hypothèse distributionnelle, consiste à aligner des termes qui tendent à apparaître dans des contextes similaires. Les termes ont été préalablement identifiés dans les corpus source et cible à l'aide d'un extracteur de termes. Nous basant sur les besoins exprimés par divers travaux en traductologie, nous avons enrichi le lexique bilingue d'informations périphériques extraites du corpus et d'Internet (collocations, fréquence, lien vers la fiche Wikipédia...) puis nous avons développé une interface permettant de consulter les alignements de termes, leurs concordances ainsi que leurs fiches terminologiques.

Cet outil nous a permis de mettre en œuvre une évaluation applicative des lexiques bilingues. Notre protocole d'évaluation a consisté à comparer la qualité de traductions humaines produites avec et sans les lexiques extraits de corpus comparables : nous avons considéré que la différence de qualité observée permettait de rendre compte de l'utilité des lexiques. L'évaluation a porté sur deux thématiques : CANCER DU SEIN (domaine médical) et SCIENCES DE L'EAU (domaine de l'environnement). Le bilan de l'évaluation a été mitigé. Concernant la thématique SCIENCES DE L'EAU, il s'est avéré que les lexiques extraits avaient peu de vocabulaire en commun avec les lexiques à traduire (14 % pour la partie source), ce qui ne nous a pas permis de tirer des conclusions à partir de ces données. Toutefois, cette déconvenue a mis en avant le fait qu'il est nécessaire de recourir à des thématiques à granularité fine pour espérer bénéficier des corpus comparables. Le lexique CANCER DU SEIN étant plus couvrant (67 % pour la partie cible, 94 % pour la partie source), nous avons pu observer que le lexique extrait du corpus comparable améliorait effectivement la qualité des traductions par rapport à une situation où le traducteur n'aurait que des ressources généralistes à sa disposition. Toutefois, dans un domaine très bien documenté tel que le domaine médical, les corpus comparables présentent peu d'intérêt par rapport à une situation où le traducteur utiliserait les ressources disponibles

sur Internet.

Cette expérience nous a permis de recueillir les impressions des traducteurs face à ce nouveau type de lexiques qu'ils n'ont pas l'habitude de manipuler. Nous avons appris que l'ambiguïté des lexiques (i.e. le fait qu'un terme source est associé à plusieurs traductions candidates) est gênante et que les traducteurs préfèreraient une liste très réduite de traductions candidates quitte à ce que moins de termes sources soient traduits.

Partant de ce constat, nous avons fait le choix de nous orienter vers des techniques de génération de traduction qui avaient été depuis peu appliquées avec succès à l'extraction de lexiques à partir de corpus comparables (Morin et Daille, 2010). Ces techniques consistent non pas à aligner des termes préalablement extraits des corpus mais plutôt à générer des traductions de termes sources qui peuvent ensuite être filtrées grâce aux corpus comparables. Nous avons également fait état de méthodes de génération "empiriques", c'est-à-dire basées sur l'apprentissage de relations de traductions à partir d'exemples issus de terminologies bilingues. Cependant, ces méthodes ont été écartées car elles nécessitent qu'il existe déjà un minimum de ressources spécialisées bilingues. Nous nous sommes alors orientés vers des méthodes de génération de traduction basées sur le principe de compositionnalité.

Contributions à la traduction compositionnelle

Nos contributions à la traduction compositionnelle ont porté sur la traduction d'unités monolexicales morphologiquement complexes. L'état-de-l'art de ces méthodes nous a permis de dégager trois pistes de recherches :

- Une première piste a concerné la génération de traductions fertiles, c'est-à-dire de paires de traductions dans lesquelles le terme cible possède plus de mots lexicaux que le terme source. Ces traductions, qui s'apparentent souvent à des variantes paraphrastiques du terme cible "canonique" s'inscrivent avantageusement dans notre cadre de recherche : elles permettent d'identifier des équivalences traductionnelles dont la structure est éloignée du terme source et s'adaptent donc bien aux spécificités des corpus comparables qui contiennent des formulations naturelles, non influencées par la langue source. Par exemple, *cytogenetic* peut être traduit par *génétique des cellules*[5]. Les traductions fertiles permettent également de proposer des variantes vulgarisées du terme canonique, ce qui est utile lorsque le traducteur traduit des textes vulgarisés. Par exemple, *oophorectomy* peut être traduit par *ovariectomie* ou *ablation des ovaires*.
- Notre deuxième piste de recherche avait pour but de produire un système de traduction relativement indépendant de la structure morphologique des termes à traduire. En effet, les méthodes existantes sont chacune adaptées à un ou deux types de construction morphologique et ne proposent pas de solution générique. Le système que nous avons implanté peut traduire des termes préfixés, des composés populaires et savants ainsi que quelques cas de suffixation. Par exemple, le terme *abnormaly* est décomposé en `ab-+normal+-ly` et traduit par *(de) façon anormale* ; le terme *over-represented* est découpé en `over+re-+presented` et traduit par *sur-représentation* ; le terme *phyto-estrogens* est découpé en `-phyto-+-estro-+-gen-` et traduit par *Pflanzenöstrogen*.
- Notre troisième et dernier axe de recherche a abordé la problématique de l'ordonnancement des traductions générées : nous avons exploré plusieurs critères d'ordonnancement ainsi que des méthodes de combinaison de ces critères. Nous avons introduit l'usage d'algorithmes issus de la recherche d'information qui modélisent

5. Contexte anglais : ≪ *A molecular* ***cytogenetic*** *approach to studying platinum resistance* ≫ ; contexte français : ≪ *Parmi les anomalies moléculaires ou* ***génétiques des cellules*** *tumorales* ≫

directement le problème de l'ordonnancement (approches *list-wise* en *learning-to-rank*).

Nous avons évalué notre méthode de traduction sur un corpus médical correspondant à la thématique du cancer du sein. Deux couples de langues ont été étudiés : anglais → français et anglais → allemand. Les évaluations menées ont montré que la génération de traductions fertiles augmentait la qualité du lexique uniquement pour la traduction de l'anglais vers le français. Nous avons émis l'hypothèse que, dans le cas de la traduction de l'anglais vers l'allemand, les traductions fertiles étaient moins intéressantes à cause du type morphologique de l'allemand qui est une langue germanique ayant tendance à former les mots par agglutination de morphèmes plutôt qu'en créant des locutions ou des syntagmes.

En ce qui concerne l'indépendance à la structure morphologique du terme source, les expériences ont également montré qu'il était plus avantageux que l'algorithme prenne en entrée plusieurs cas de constructions morphologiques plutôt qu'il soit dédié à un certain type de structure. La généricité augmente la couverture du lexique et nous n'avons pas observé de baisse trop importante de la précision.

Enfin, concernant l'ordonnancement de traductions candidates, nous avons montré que la combinaison de plusieurs critères donnait des résultats supérieurs à un ordonnancement aléatoire ou à l'usage d'un critère isolé et ce, même sans recourir à des modèles appris à partir d'exemples de traduction.

Limites et perspectives de recherche

Domaine de connaissance choisi

À notre sens, la principale limite de nos travaux concerne le domaine auquel appartiennent les corpus sur lesquels nous avons travaillé. En effet, le domaine médical est un domaine particulièrement bien doté en ressources parallèles. Il est donc peu pertinent d'avoir mené sur ce type de domaine des recherches destinées à acquérir des ressources linguistiques pour les domaines peu dotés. Néanmoins, le fait de travailler dans le domaine médical nous a donné accès à des ressources de référence (que ce soit l'UMLS, des dictionnaires médicaux en ligne ou les traductions de référence utilisées pour l'évaluation applicative). Ces références nous ont permis de rapidement et facilement évaluer nos méthodes. Dans un domaine peu doté, nous aurions dû avoir recours à un expert pour évaluer la pertinence des traductions produites.

Le domaine médical a également été un domaine idéal pour tester une méthode basée - entre autres - sur l'analyse des composés savants. Même si les racines néoclassiques se retrouvent dans de nombreux domaines techniques, on sait que ces dernières sont particulièrement fréquentes dans le domaine médical (Namer et Baud, 2007). Notre première perspective sera donc d'expérimenter nos méthodes sur un domaine peu doté, par exemple sur une thématique liée à l'écologie.

Langues et structures morphologiques

Une seconde limite de notre travail concerne le petit nombre et la petite variété morphologique des couples de langues testés. Nous avons vu que la traduction de l'anglais vers le français se prêtait bien à la génération de traductions fertiles. Il serait intéressant de tester plus avant l'influence du sens de traduction et du type morphologique des langues en jeu sur la génération de ce type de traductions. Nous songeons en particulier à tester notre méthode sur des langues sources agglutinantes comme le turc ou le finnois.

Nous avons essayé d'améliorer la généricité des techniques actuelles de traduction compositionnelle. Toutefois, nous sommes restés focalisés sur quatre processus de construction morphologique fréquents dans les langues d'Europe de l'Ouest. De plus, nous avons envisagé le phénomène de fertilité uniquement sous l'angle de la traduction d'unités monolexicales vers des unités polylexicales. La généricité du système peut donc encore être augmentée : gestion des infixes et des circumfixes, traduction d'unités polylexicales vers monolexicales, etc.

Évaluation applicative

Nous avons évalué notre méthode de génération de traduction par comparaison à une référence. Nous souhaiterions pouvoir mener une évaluation applicative de cette méthode, en particulier comparer la qualité des traductions obtenues avec un lexique uniquement construit grâce à la méthode distributionnelle vs. un lexique contenant des traductions issues de la méthode distributionnelle et de la méthode compositionnelle.

En lien avec la problématique du domaine de connaissance choisi, nous aimerions pouvoir renouveler notre expérience d'évaluation applicative dans de meilleures conditions. Tout d'abord, nous choisirions des textes appartenant à une thématique liée à un domaine peu doté. Ensuite, nous nous assurerions que le corpus source offre une bonne couverture du vocabulaire des textes à traduire et que corpus source et cible sont suffisamment comparables.

Contextualisation et outillage

Pour conclure ce mémoire, nous souhaitons revenir sur les quatre axes de recherches soulevés en fin de chapitre 2 que sont : (i) la collecte automatique de corpus comparables ; (ii) la recherche d'équivalences traductionnelles ; (iii) la contextualisation des termes et (iv) le développements d'outils de TAO faits pour les corpus comparables.

D'un point de vue industriel, nous pensons que les deux premières thématiques de recherche ont été suffisamment explorées et qu'il existe aujourd'hui des techniques raisonnablement mûres pour être transférées à l'industrie. Toutefois, comme semblent l'indiquer nos expériences, les lexiques issus de corpus comparables ne peuvent pas répondre au même cas d'usage qu'un lexique issu de corpus parallèles. En effet, lexiques issus de corpus comparables seront toujours de petite taille et ambigus. Nous pensons que les corpus comparables ont surtout un rôle à jouer non pas lors de la phase de traduction proprement dite mais plutôt en amont de la traduction, lors de la phase de recherche documentaire décrite par Durieux (2010) : ils doivent permettre au traducteur de se familiariser avec un domaine nouveau, d'en comprendre les notions et d'en observer les usages. Ils peuvent également être d'une grande aide aux terminologues puisqu'ils permettent d'identifier et d'observer les termes dans des contextes d'usages non influencés par une langue source.

Pour exploiter tout le potentiel des corpus comparables, il conviendrait alors d'approfondir les recherches visant à contextualiser les termes et à développer des outils d'exploration adaptés. Il faudrait par exemple pouvoir sélectionner les contextes d'un terme qui sont les plus informatifs pour le traducteur, que ce soit au niveau sémantique (contextes définitoires) ou au niveau linguistique (collocations, sous-catégorisation). Il serait aussi intéressant de pouvoir "aligner" les contextes, c'est-à-dire de mettre en regard, à la manière des concordanciers de corpus parallèles, un contexte source et un contexte cible qui partagent une large de part de vocabulaire. Enfin, les corpus comparables manquent cruellement d'outils permettant leur exploration. Il serait bon d'intégrer aux outils de TAO des fonctionnalités avancées telles que

des moteurs de recherche gérant des requêtes complexes [6] ou des possibilités de visualisation des proximités sémantiques entre termes.

6. Par exemple, le moteur NETSPEAK (Riehmann *et al.*, 2012) dispose d'un langage de requête avancé : la requête "{ a b c }" s'apparie avec toutes les permutations de a, b et c et donne leurs fréquences sur Internet – http://www.netspeak.cc.

Annexes

Annexe A

Mesures

Sommaire

A.1 Normalisation des vecteurs

Dans les formules à suivre, M est la tête du vecteur, m est l'un de ses cooccurrents, $nbOcc(M, m)$ est le nombre de fois où M et m cooccurrent, $nbOcc(M)$ est le nombre d'occurrences de M, $nbOcc(m)$ est le nombre d'occurrences de m.

A.1.1 Taux de vraisemblance

source : Morin *et al.* (2004)

Le taux de vraisemblance se calcule sur la base de la table de contingence suivante :

	m	$\neg m$	TOTAL
M	a	b	e
$\neg M$	c	d	h
TOTAL	f	g	N

TABLE A.1 – Table de contingence des coocurrences de mots observées sur le corpus

où a est le nombre fois où M et m cooccurrent ($nbOcc(M, m)$) , b le nombre de fois où M co-occurre avec un autre mot que m, etc. N est le nombre total de co-occurrences calculées sur le corpus.

Le taux de vraisemblance TV entre M et m s'obtient par :

$$\begin{aligned} TV(M,m) = {} & a\log(a) + b\log(b) + c\log(c) + N\log(N) \\ & - e\log(e) - f\log(f) - g\log(g) - h\log(h) \end{aligned} \tag{A.1}$$

A.1.2 Discounted log-ods

source : Laroche et Langlais (2010).

Le calcul du *discounted log-ods* se base également sur la table de contigence A.1 :

$$odds\_ratio = log\frac{(a+\frac{1}{2})(d+\frac{1}{2})}{(b+\frac{1}{2})(c+\frac{1}{2})} \tag{A.2}$$

A.1.3 Information mutuelle

source : Morin *et al.* (2004, p. 106)

$$IM(M,m) = \log\frac{nbOcc(M,m)}{nbOcc(M)\cdot nbOcc(m)} \tag{A.3}$$

A.1.4 TFIDF

source : Fung (1998, p. 8-11)

$$TFIDF(M, m) = TF \cdot IDF \tag{A.4}$$

$$TF = nbOcc(M, m) \tag{A.5}$$

$$IDF(m) = \log \frac{nbMax}{nbOcc(m)} + 1 \tag{A.6}$$

où $nbMax$ la plus haute fréquence rencontrée sur le corpus.

A.2 Similarité de deux vecteurs

Dans les formules à suivre, V_S est le vecteur source de tête M_S , V_C est le vecteur cible de tête M_C, n est le nombre d'entrées du lexique bilingue, m_i est une entrée du lexique bilingue, $P(M_S, m_i)$ est le nombre de co-ocurrences normalisé entre m_i et la tête du vecteur source M_S, $P(M_C, m_i)$ est est le nombre de co-ocurrences normalisé entre m_i et la tête du vecteur cible M_C.

A.2.1 Cosine

source : Fung (1998, p. 12)

$$Cosine(V_S, V_C) = \frac{\sum_{i=1}^{n} P(M_S, m_i) \cdot P(M_C, m_i)}{\sqrt{\sum_{i=1}^{n} P(M_S, m_i)^2 \cdot \sum_{i=1}^{n} P(M_C, m_i)^2}} \tag{A.7}$$

A.2.2 Jaccard pondéré

source : Prochasson (2010, p. 62)

$$J(V_S, V_C) = \frac{\sum_{i=1}^{n} min(P(M_S, m_i), P(M_C, m_i))}{\sum_{i=1}^{n} max(P(M_S, m_i), P(M_C, m_i))} \tag{A.8}$$

A.2.3 Distance euclidienne

source : Fung (1997, p. 196)

$$DE(V_S, V_C) = \sum_{i=1}^{n} |P(M_S, m_i) - P(M_C, m_i)| \tag{A.9}$$

A.2.4 Distance euclidienne normalisée

source : Hazem et Morin (2012, p. 128)

$$DEN(V_S, V_C) = \sqrt{\sum_{i=1}^{n} (\frac{P(M_S, m_i)}{||V_S||} - \frac{P(M_C, m_i)}{||V_C||})} \tag{A.10}$$

A.3 Comparabilité de deux corpus

source : Li et Gaussier (2010, p. 645-646)

$$M(C_e, C_f) = \frac{\sum_{w \in C_e^v \cap D_e^v} \sigma(w, C_f^v) + \sum_{w \in C_f^v \cap D_f^v} \sigma(w, C_e^v)}{|C_e^v \cap D_e^v| + |C_f^v \cap D_f^v|} \tag{A.11}$$

$$\sigma(w, C^v) = \begin{cases} 1 & \text{si } T_w \cap C^v \neq \emptyset \\ 0 & \text{sinon} \end{cases}$$

où :

- C_e, C_f sont les corpus source, resp. cible.
- C_e^v, C_f^v est le vocabulaire du corpus source, resp. cible (mots lexicaux uniquement).
- D_e^v, D_f^v sont les entrées source, resp. cible, d'un dictionnaire bilingue.
- $\sigma(w, C^v)$ est une fonction renvoyant 1 si au moins une traduction du mot w se trouve dans le vocabulaire du corpus C.

A.4 Standardisation des valeurs

source : Gendre (1977, p. 48-50)

A.4.1 Obtention du percentile d'une valeur

Le percentile d'une valeur v est obtenu avec la formule :

$$percentile(v) = \frac{lower(v) + 0.5\ eff(v)}{N} \tag{A.12}$$

où $lower(s)$ est le nombre d'individus avec une valeur plus petite que v, $eff(v)$ est le nombre d'individus avec la valeur v et N est le nombre total d'individus dans la population.

A.4.2 Obtention du score-z associé au percentile

Le score-z est obtenu via la table standard normale (tableau A.2). Dans la table standard normale T la cellule T_{ij} correspond au percentile associé au score-z $i + j$. Par exemple, le percentile $0,5438$ correspond au score-z 0,11 (ligne 0,1 ; colonne 0,01).

Pour les percentiles inférieurs à 0,5 nous obtenons le score-z ainsi :

$$zscore(percentile) = -T(1 - percentile) \tag{A.13}$$

Par exemple, le percentile 0,4562 correspond au score-z $-0,11$.

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990

TABLE A.2 – Table standard normale

source : http ://en.wikipedia.org/wiki/Standard_normal_table#Cumulative_table

A.5 Mesures d'évaluation

A.5.1 Couverture

source : Langlais *et al.* (2008)

$$C = \frac{|ST|}{|S|} \tag{A.14}$$

$$ST = \{s : |\mathcal{T}(s)| > 1\}$$

où :
- S est l'ensemble des termes sources
- $\mathcal{T}(s)$ est l'ensemble des traductions générées par le système pour le terme source s

A.5.2 TopN / Precision au rang N

source : Langlais *et al.* (2008)

$$P_N = \frac{|SR_N|}{|ST|} \tag{A.15}$$
$$SR_N = \{s : \mathcal{T}_N(s) \cap \mathcal{R}(s) \neq \emptyset\}$$
$$ST = \{s : |\mathcal{T}(s)| > 1\}$$

où :
- $\mathcal{T}(s)$ est l'ensemble des traductions générées par le système pour le terme source s
- $\mathcal{T}_N(s)$ est l'ensemble des N premières traductions de s
- $\mathcal{R}(s)$ est l'ensemble des traductions de référence de s

A.5.3 Rappel au rang N

source : Langlais *et al.* (2008)

$$R_N = C \times P_N = \frac{|SR_N|}{|S|} \tag{A.16}$$
$$SR_N = \{s : \mathcal{T}_N(s) \cap \mathcal{R}(s) \neq \emptyset\}$$

où :
- S est l'ensemble des termes sources
- $\mathcal{T}_N(s)$ est l'ensemble des N premières traductions de s
- $\mathcal{R}(s)$ est l'ensemble des traductions de référence de s

A.5.4 F1-mesure au rang N

$$F1_N = 2 \times \frac{P_N \times R_N}{P_N + R_N} \tag{A.17}$$

où :
- P_N est la précision au rang N
- R_N est le rappel au rang N

A.5.5 MRR : Mean Reciprocal Rank

source : Yu et Tsujii (2009)

$$MRR = \frac{1}{|S|} \sum_{i=1}^{|S|} \frac{1}{rang_i} \tag{A.18}$$

où S est l'ensemble des termes sources avec au moins une traduction candidate et $rang_i$ le rang de la première traduction candidate correcte du terme source i.

A.5.6 MAP : Mean Average Precision

source : Manning *et al.* (2008)

$$MAP(Q) = \frac{1}{Q} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} Precision(R_{j^k}) \tag{A.19}$$

où :
- Q est l'ensemble des requêtes
- m_j est le nombre de documents à ramener pour la requête Q_j
- R_{jk} est l'ensemble des documents retournés par le système pour la requête Q_j avant le document k.

A.5.7 NDCG : Normalised Discounted Cumulative Gain

source : Manning *et al.* (2008).

La NDCG est basée sur la DCG (Discounted Cumulative Gain). La DCG se calcule pour un rang donné (k) :

$$DCG@k = rel_1 + \sum_{i=2}^{k} \frac{rel_i}{log_2(i)} \tag{A.20}$$

où rel_i est le score de pertinence du document i.

La NDCG est :

$$NDCG@k = \frac{DCG@k}{IDCG@k} \tag{A.21}$$

où $IDCG@k$ est la DCG maximum qu'il est possible d'atteindre, cette dernière est obtenue en calculant la DCG sur la liste des documents correctement ordonnés.

Pour finir, on donne la NDCG moyenne pour toutes les requêtes Q :

$$NDCG(Q,k) = \frac{1}{Q} \sum_{j=1}^{|Q|} \frac{DCG@k}{IDCG@k} \tag{A.22}$$

A.6 Accord inter-annotateur

A.6.1 Calcul du $Kappa$

source : Carletta (1996)

Le $Kappa$ est calculé à partir de $P(a)$, l'accord observé entre annotateurs, et $P(e)$, la probabilité d'un accord aléatoire ($1/N$ où Nest le nombre de catégories utilisées pour l'annotation) :

$$Kappa = \frac{P(a) - P(e)}{1 - P(e)} \tag{A.23}$$

A.6.2 Interprétation du $Kappa$

Le $Kappa$ s'interprète ainsi (Landis et Koch, 1977) :

$Kappa$	interpretation
< 0	désaccord
0,0 - 0,20	accord très faible
0,21 - 0,40	accord faible
0,41 - 0,60	accord modéré
0,61 - 0,80	accord fort
0,81 - 1,00	accord presque parfait

Annexe B

Données

Sommaire

B.1 Corpus comparables

B.1.1 Sciences de l'eau

Le corpus SCIENCES DE L'EAU a été constitué à partir des résumés d'articles scientifiques de la la revue anglophone *Water Science Technology*[1] et à partir d'articles scientifiques de la revue francophone *Sciences de l'eau*[2]. Sa comparabilité est de 0,77.

	EN	FR
nb de mots	2 202 523	2 382 033
nb de documents	11 914	500

TABLE B.1 – Taille du corpus SCIENCES DE L'EAU (nb. de mots et nb. de documents)

B.1.1.1 Anglais

```
 The pseudo first-order rate constant (k obs)   for PCE
    dechlorination by 0.06 wt\% Cu/Si was 0.028 h -1, which was
    2.8 times higher than that by Si(0) alone.
However, the k obs for PCE dechlorination decreased to 0.0016 h
   -1 when the loading of Cu(II) increased to 3 wt%. The EPMA
   results showed that the distribution of loading of 0.06 wt% Cu
   on the Si(0) surface constant was observed before the total
   coverage of the active sites on the reductive metal by the
   catalytic metal layer. The surface coverage of Cu to Si(0) can
   theoretically calculate by estimation of the lowest energy
   fcc(111) crystallographic orientation.
```

source : http ://www.ncbi.nlm.nih.gov/pubmed/20595750

Extrait B.1 – Paragraphe du corpus Sciences de l'eau anglais

1. http ://www.iwaponline.com/wst/
2. http ://www.rse.inrs.ca/

B.1.1.2 Français

```
L'Interprétation des résultats de près de 150 essais de pompage
    fournit les valeurs suivantes (DASSARGUES étal., 1987) 1.10-9
    $<$ K $<$ 2.10-7 m/sec - pour les limons -pour le conglomérat
    résiduel 1.10-5 $<$ K $<$ 8.10-3 m/sec - pour les craies
    supérieures (fracturées) 2.1 $<$H$<$ K $<$ 5.10-3 m/sec -
    pour les craies inférieures (compactes) 1.10-5 $<$ K $<$
    5.10-4 m/sec La porosité efficace (CASTANY, 1982) ou de
    drainage (DE MARSILY, 1981), déterminée par les essais de
    pompage et par observation des fluctuations de la surface
    piézométrique en fonction de l'infiltration efficace, serait
    de l'ordre de 5 % ; elle dépasserait toutefois 10 % dans de
    nombreuses zones où la craie est plus fracturée et altérée.
```

http ://www.rse.inrs.ca/art/volume4/v4n1_39.pdf

Extrait B.2 – Paragraphe du corpus Sciences de l'eau français

B.1.2 Cancer du sein

Dans sa partie scientifique, le corpus CANCER DU SEIN a été constitué à partir de publications scientifiques collectées sur le portail *Elsevier*[3] et sur *Google Scholar*[4]. Dans sa partie vulgarisée, le corpus a été construit à partir d'articles de sites Internet de vulgarisation à destination des patient-e-s et de leur proches.

	EN	FR	DE
scientifique	198 244	267 180	197 187
vulgarisé	218 336	184 504	201 760
TOTAL	416 580	451 684	398 947

TABLE B.2 – Composition et taille du corpus en nombre d'occurrences

	EN	FR	DE
scientifique	70	78	103
vulgarisé	272	217	162
TOTAL	342	295	265

TABLE B.3 – Composition et taille des corpus en nombre de documents

3. http://www.elsevier.com/
4. http://scholar.google.com/

	EN-FR	EN-DE
scientifique	0,71	0,42
vulgarisé	0,69	0,46
TOUT	0,74	0,45

TABLE B.4 – Comparabilité des corpus

B.1.2.1 Allemand

Discours scientifique

In diesem Jahr werden rund 50000 bis 60000 Frauen an Brustkrebs erkranken. Zirka ein Drittel der betroffenen Frauen wird daran versterben. Damit ist das Mamma-Karzinom die häufigste Krebserkrankung bei Frauen in der westlichen Welt - Tendenz steigend. Die deutsche Antwort darauf heißt Interdisziplinäres-Zertifiziertes-Brustzentrum. Seit Einführung der Zertifizierung vor drei Jahren werden rund die Hälfte aller Brustkrebspatientinnen dort behandelt und die Erfolge geben dieser Therapieeinrichtung recht.

http ://www.uni-frauenklinik-tuebingen.de/bereiche/brustzentrum/aktuelles/pressearchiv.html

Extrait B.3 – Paragraphe du corpus Cancer du sein scientifique allemand

Discours vulgarisé

Wichtig ist, dass die behandelnden Ärzte Bescheid wissen, was der Patient selbst einnimmt oder anwendet. Der Gedanke "vielleicht␣hilft␣es␣nicht,␣es␣schadet␣aber␣auch␣nicht" sei falsch, sagt Kümmel. Schon ein Glas Grapefruitsaft oder Johanniskraut kann die Wirkung einer Chemotherapie aushebeln. Mulitvitaminkuren sind zum Beispiel bei Bestrahlungen kontraproduktiv. Brustkrebspatienten mit hormonempfindlichen Brustkrebs sollten keine hochdosierten Pflanzenextrakte einnehmen, die Phytoöstrogene enthalten, zum Beispiel Soja oder Ginseng. Die Finger lassen sollten Krebspatienten auch von Mitteln, die Heilung versprechen. Beispiele sind bittere Aprikosenkerne, in denen Amygdalin steckt, oder Galavit, ein Stoff, der im Internet als Krebstherapeutikum vermarktet wird. Auch bei Noni-, Goji- und Mangostinsäften gibt es bislang keinen Nachweis, dass sie gegen Krebs wirken.
Medizinische Datenflut

http ://www.netdoktor.de/Magazin/Komplementaermedizin-bei-Kreb-11741.html

Extrait B.4 – Paragraphe du corpus Cancer du sein vulgarisé allemand

B.1.2.2 Anglais

Discours scientifique

Summary scores were calculated according to the EORTC manual.22 Summary statistics were calculated **for** baseline scores, according to categories of **type** of surgery (Mx, WLE), time since surgery, CT, ET and age. Subscale scores were compared between groups of patients defined according to age and the **set** of clinical variables tested using t-tests and analysis of variance (ANOVA) as appropriate. In order to investigate the effect of confounding on observed associations with QOL scores, multiple regression was used, and the significance of variables tested using the F-**test** from an ANOVA that included all relevant independent variables (surgery, time since surgery, CT, ET and age). Hence the p-values from the F-**test** indicate the significance of each variable after allowing **for** the effects of the other variables listed. Variables which remain significant in the multiple regression can be said to have independent statistically significant effects on the outcome in question.

http ://www.sciencedirect.com/science/article/pii/S0960977606002165

Extrait B.5 – Paragraphe du corpus Cancer du sein scientifique anglais

Discours vulgarisé

```
  Breast awareness
Most breast cancers are detected by women who report unusual
   changes to their GP. This highlights the importance of being
   breast aware.
Breast awareness means knowing what your breasts look and feel
   like normally, so you can be aware of any changes and check
   them out with your doctor as soon as possible. If cancer is
   diagnosed, prompt treatment offers the best chance of a
   successful outcome.

So the message from Breakthrough Breast Cancer is simple:
Show your breasts some TLC. Touch. Look. Check. Be breast aware
   visit www.touchlookcheck.org.uk
```

http ://www.plumslingerie.com/breastawareness-article-from-breakthroughorguk-z-18.html

Extrait B.6 – Paragraphe du corpus Cancer du sein vulgarisé anglais

B.1.2.3 Français

Discours scientifique

```
L'évaluation histopathologique est importante, puisqu'elle peut
   permettre aux cliniciens de modifier le schéma thérapeutique.
   Ainsi, des patientes en réponse complète histologique (RCH) et
   ayant subi une mammectomie pourront ne pas recevoir de
   radiothérapie. De plus, une chimiothérapie adjuvante sera plus
   volontiers proposée à une patiente qui n'est pas en RCH qu'à
   une patiente qui est en RCH.
Classifications histopathologiques de la réponse à la
   chimiothérapie
Plusieurs équipes de pathologistes ont proposé des
   classifications histopathologiques de la réponse à la
   chimiothérapie néoadjuvante. Il faut cependant souligner qu'il
   n'existe pas à l'heure actuelle de consensus permettant de
   privilégier une classification plutôt qu'une autre. Cela
   complique notablement les comparaisons des taux de réponses
   complètes histopathologiques (RCH) obtenues dans différentes é
   tudes. Le tableau 1( Tableau 1 ) résume les différentes
   classifications qui ont été décrites dans la littérature.
```

http ://www.cancerdusein.org/cds/index.php ?option=com_content&task=view&id=74&Itemid=169

Extrait B.7 – Paragraphe du corpus Cancer du sein scientifique français

	Nb. documents	Nb. mots
textes scientifiques	3 résumés d'articles	499 mots
textes vulgarisés	1 page web	425 mots

TABLE B.5 – Taille des textes à traduire, thématique SCIENCES DE L'EAU

Discours vulgarisé

L'adolescente
Quelles sont les précautions à prendre si l'on découvre une petite boule chez une adolescente ? Bien évidemment, il convient de consulter un médecin. Celui-ci procédera tout d'abord à l'interrogation de la patiente, c'est-à-dire, à la recherche d'apparition de la boule, à ses variations de volume par rapport aux règles, à sa sensibilité... puis à la recherche des antécédents personnels **et** familiaux, des facteurs de risques... etc. Le médecin procédera ensuite à l'examen clinique des seins par leur palpation méthodique zone par zone. Lorsque la boule décelée n'excède guère plus de 2 cm **et** que le bilan général de l'interrogatoire **et** de l'examen clinique est normal, le médecin pourra avancer le diagnostic de fibroadénome du sein de l'adolescente. Il s'agit d'une sorte de petit fibrome, comme pour l'utérus.

http ://www.cancerdusein.org/cds/index.php ?option=com_content&task=view&id=74&Itemid=169

Extrait B.8 – Paragraphe du corpus Cancer du sein vulgarisé français

B.2 Textes à traduire et traductions de référence

B.2.1 Sciences de l'eau

Les textes à traduire scientifiques et leurs traductions correspondent à des résumés d'articles scientiques de la revue *Sciences de l'eau*[5]. Les textes à traduire vulgarisés et leurs traductions correspondent à des pages du site Internet de la société *Lenntech*[6]

5. http://www.rse.inrs.ca/
6. http://www.lenntech.com/

B.2.1.1 Discours scientifique

Texte à traduire

```
Health Significance of Bacterial Regrowth in Drinking Water: Free
   Opinion
The presence of heterotrophic bacteria in drinking water (tap,
   point-of use treated or bottled) poses a difficult problem
   because we do not clearly know if they are really innocuous.
 Two points of views are presented: they could be totally
     unimportant whatever their number or they can be
     opportunistic pathogens and even frank pathogens if they are
     allowed to multiply in large numbers.
These bacteria are not of faecal origin and are not indicators of
   faecal pollution even if occasionally some can be described as
   coliforms.
```

http ://www.erudit.org/revue/rseau/1995/v8/n3/705225ar.html

Extrait B.9 – Paragraphe d'un texte à traduire, thématique Sciences de l'eau, discours scientifique

Traduction de référence

```
Effets sur la santé de la recroissance bactérienne dans les eaux
   de consommation / Health significance of bacterial regrowth in
   drinking water
La présence de bactéries hétérotrophes dans les eaux de
   consommation (robinet filtrée sur unités domestiques ou
   embouteillées) constitue un problème difficile à résoudre car
   on connaît très mal leurs effets sur la santé humaine.
Deux points de vue s'affrontent: l'une perçoit ces bactéries
   comme des bactéries sans aucune importance quel que soit leur
   nombre, l'autre suppose que certaines d'entre elles sont
   potentiellement pathogènes et que l'on ne doit pas leur
   permettre de se multiplier indûment dans l'eau de consommation.
Ces bactéries hétérotrophes sont présentes partout et elles
   trouvent dans l'eau de consommation une niche écologique qui
   permet parfois leur croissance en grand nombre (e.g., nodules
   du réseau, tuyauterie des maisons, chauffe-eau, eaux
   embouteillées, filtre à charbon actif, etc.).
```

http ://www.erudit.org/revue/rseau/1995/v8/n3/705225ar.html

Extrait B.10 – Paragraphe d'une traduction de référence, thématique Sciences de l'eau, discours scientifique

B.2.1.2 Discours vulgarisé

Texte à traduire

```
Examples from active carbon in different processes:

    * Ground water purification
    * The de-chlorination of process water
    * Water purification for swimming pools
    * The polishing of treated effluent

Process description:

Water is pumped in a column which contains active carbon, this
   water leaves the column through a draining system. The
   activity of an active carbon column depends on the temperature
   and the nature of the substances. Water goes through the
   column constantly, which gives an accumulation of substances
   in the filter. For that reason the filter needs to be replace
   periodically. A used filter can be regenerated in different
   ways, granular carbon can be regenerated easily by oxidizing
   the organic matter. The efficiency of the active carbon
   decreases by 5 - 10% 1). A small part of the active carbon is
   destroyed during the regeneration process and must be
   replaced. If you work with different columns in series, you
   can assure that you will not have a total exhaustion of your
   purification system.
```

http ://www.lenntech.fr/adsorption.htm

Extrait B.11 – Paragraphe d'un texte à traduire, thématique Sciences de l'eau, discours vulgarisé

Traduction de référence

```
Exemples d'application :
    * Traitement des eaux souterraines polluées
    * Traitement contre les micropolluants, adsorption des traces
       de certains métaux lourds
    * Rétention de chlore
    * Filtration fine pour piscines
    * Filtration finale pour le rejet d'effluents

Les performances des filtres à charbon actif dépendent de la
   température, ainsi que du composé à adsorber.

Pour les applications de traitement d'eau dans les procédés des
   industries alimentaires, la filtration au charbon actif est
   souvent accompagnée d'une désinfection UV.

Description du processus

L'eau est pompée dans une colonne qui contient du charbon actif,
   cette eau quitte la colonne à travers un système drainant.
   L'activité de la colonne de charbon actif dépend de la
   température et de la nature des substances. L'eau passe à
   travers la colonne continuellement, ce qui entraîne une
   accumulation des substances sur le filtre. Pour cette raison,
   le filtre a besoin d'être remplacé périodiquement. Un filtre
   utilisé peut être régénéré de différentes façons, le charbon
   granulaire peut être régénéré facilement en oxydant la matière
   organique. L'efficacité du charbon actif diminue alors de 5 à
   10%. Une petite partie du charbon actif est détruite pendant
   le processus de régénération et doit être remplacée. Si vous
   travaillez avec différentes colonnes en série, vous pouvez
   vous assurer que vous n'aurez pas un épuisement total de votre
   système de purification.
```

http ://www.lenntech.com/library/adsorption/adsorption.htm

Extrait B.12 – Paragraphe d'une traduction de référence, thématique Sciences de l'eau, discours vulgarisé

B.2.2 Cancer du sein

B.2.2.1 Discours scientifique

Les textes à traduire scientifiques et leurs traductions correspondent à des résumés d'articles scientiques issus du portail *Elsevier*[7]. Les textes à traduire vulgarisés et leurs traductions correspondent à des pages du site Internet de la *Société canadienne du cancer*

7. http://www.elsevier.com/

	Nb. documents	Nb. mots
textes scientifiques	3 résumés d'articles	508 mots
textes vulgarisés	1 page web	613 mots

TABLE B.6 – Taille des textes à traduire, thématique SCIENCES DE L'EAU

du sein[8]

Texte à traduire

```
Furthermore, these studies have shown that these pathways
   direct and in turn are influenced by the tissue structure.
   Tissue structure is directed by the cooperative interactions
   of the cell-cell and cell-ECM pathways and can be modified
   by stromal factors. Not surprisingly then, loss of tissue
   structure and alterations in ECM components are associated
   with the appearance and dissemination of breast tumors, and
   malignancy is associated with perturbations in cell
   adhesion, changes in adhesion molecules, and a stromal
   reaction.
```

http ://pmcc.web-t.cisti.nrc.ca/articlerender.cgi ?artid=1705474

Extrait B.13 – Paragraphe d'un texte à traduire, thématique Cancer du sein, discours scientifique

8. http://www.cbcf.org/

Traduction de référence

De plus, ces études ont montré que ces voies règlent la structure tissulaire **et** qu'en retour, elles sont influencées par elle. La structure tissulaire est réglée par des interactions coopératives entre les voies de signalement entre cellules **et** entre les cellules **et** la MEC; elle peut é galement être modifiée par des facteurs du stroma. Il n'est donc pas surprenant qu'une perte de structure tissulaire **et** que des modifications de constituants de la MEC soient associées à l'apparence **et** la dissémination des tumeurs mammaires, **et** que la malignité soit associée à des perturbations de l'adhérence cellulaire, à des changements dans les molécules d'adhérence **et** à une réaction du stoma.

http ://www.nrcresearchpress.com/doi/abs/10.1139/o96-089

Extrait B.14 – Paragraphe d'un texte à traduire, thématique Cancer du sein, discours scientifique

B.2.2.2 Discours vulgarisé

Texte à traduire

Breast cancer starts in the cells of the breast. The breast tissue covers an area larger than just the breast. It extends up to the collarbone and from the armpit across to the breastbone in the centre of the chest. The breasts sit on the chest muscles that cover the ribs. Each breast is made of glands, ducts (thin tubes) and fatty tissue. Lobules are groups of glands that can produce milk. Milk flows from the lobules through a network of ducts to the nipple. The nipple is in the centre of a darker area of skin called the areola. Fatty tissue fills the spaces between the lobules and ducts and protects them.

A woman's breasts may feel different at different **times** of her menstrual cycle, sometimes becoming lumpy just before her period. Breast tissue also changes with age. Breast tissue in younger women is mostly made of glands and milk ducts, but older women's breasts are made up mostly of fatty tissue.

Extrait B.15 – Paragraphe d'un texte à traduire, thématique Cancer du sein, discours vulgarisé

Traduction de référence

```
Le cancer du sein se forme dans les cellules du sein. Le tissu
    mammaire ne comprend pas seulement le sein, mais aussi la
    partie du corps comprise entre la clavicule, l'aisselle et la
    lame du sternum, au milieu de la poitrine. Les seins reposent
    sur les muscles de la poitrine qui recouvrent les côtes.
    Chaque sein est constitué de glandes mammaires, de canaux
    galactophores (petits conduits) et de tissu adipeux. Les
    glandes mammaires, groupées en lobules, produisent le lait
    maternel, qui circule depuis les lobules jusqu'au mamelon par
    un réseau de canaux. Le mamelon se trouve au centre d'une
    région cutanée plus foncée, appelée aréole. Le tissu adipeux
    occupe l'espace entre les lobules et les canaux, et les
    protège.

À différents moments de son cycle menstruel, la femme pourra é
    prouver des sensations différentes au niveau de ses seins;
    ceux-ci deviendront parfois grumeleux juste avant les règles.
    Le tissu mammaire subit également des changements au cours de
    la vie. Chez les femmes plus jeunes, le tissu mammaire est
    principalement constitué de glandes et de canaux galactophores
    alors que chez les femmes plus âgées, le tissu adipeux
    prédomine.
```

Extrait B.16 – Paragraphe d'une traduction de référence, thématique Cancer du sein, discours vulgarisé

B.3 Ressources linguistiques

B.3.1 Dictionnaire bilingue généraliste

Le dictionnaire bilingue généraliste est intégré à l'analyseur linguistique XELDA (version 2.8.1)[9]. Il n'est pas directement accessible. Le nombre d'entrées indiqué est celui fourni par la société éditrice du logiciel.

	EN → FR	EN → DE
Dictionnaire bilingue généraliste	37 655 → 59 495	37 655 → 69 285

TABLE B.7 – Taille des dictionnaires généralistes (nb. d'entrées)

9. http://www.temis.com

B.3.2 Dictionnaire de synonymes

Le dictionnaire bilingue généraliste est intégré à l'analyseur linguistique XELDA (version 2.8.1) [10]. Il n'est pas directement accessible. Le nombre d'entrées indiqué est celui fourni par la société éditrice du logiciel.

	EN	FR	DE
Synonymes	5 064 → 7 596	2 387 → 3 169	4 209 → 4 883

TABLE B.8 – Taille des dictionnaires de synonymes (nb. d'entrées)

B.3.3 Tables de traduction des morphèmes

La table de traduction des morphèmes a été constituée manuellement à partir de ressources en lignes [11] et d'un dictionnaire encyclopédique (Drosdowski, 2006).

Légende

:**p** préfixe

:**c** confixe

:**s** suffixe

:**w** mot

EN	DE	FR
ab:p	a:p, ab:p, abs:p, gegenteil:w, nicht:w, umgekehrt:w	a:p, anti:p, contraire:w, inverse:w
anti:p	anti:p, anti:w, gegen:w, gegenteil:w	anti:p, anti:w, contr:p, contraire:w, contre:p, contre:w
auto:p	aut:p, auto:p, eigen:w, persönlich:w, selbst:p, selbst:w, unmittelbar:w	auto:p, propre:w, soimême:w
bi:p	bi:p, doppelheit:w, zwei:w, zweiheit:w	bi:p, deux:w, di:p, double:w
co:p	co:p, com:p, con:p, ko:p, kom:p, kon:p, mit:w, zusammen:w	co:p, ensemble:w, parallèle:w
contra:p	contra:p, gegen:w, gegenteil:w, kontra:p	contr:p, contra:p, contre:p, contre:w, contro:p
counter:p	contra:p, gegen:w, gegenteil:w, kontra:p	contr:p, contra:p, contre:p, contre:w, contro:p
cyto:p	cyto:p, zelle:w, zyto:p	cellule:w, cyto:p
de:p	de:p, entfernen:w, nicht:w, ohne:w, privat:w, stoppen:w	arrêter:w, de:p, dé:p, dés:p, enlever:w, privé:w, sans:w, ôter:w
di:p	bi:p, di:p, doppelheit:w, zwei:w, zweiheit:w	bi:p, deux:w, di:p, double:w
dis:p	dis:p, entfernen:w, in:p, nicht:w, ohne:w, privat:w, stoppen:w, un:p	arrêter:w, de:p, dé:p, dés:p, enlever:w, privé:w, sans:w, ôter:w
dys:p	dys:p, krank:w, schlecht:w, störung:w	dys:p, dé:p, mal:w, mauvais:w

10. http://www.temis.com
11. http ://medical-dictionary.thefreedictionary.com/, http ://georges.dolisi.free.fr/Terminologie/Menu/terminologie_medicale_menu.htm

epi:p	auf:w, epi:p, top:w, über:w	dessus:w, supérieur:w, sur:w, épi:p
extra:p	ausserhalb:w, außerhalb:w, extern:w, extra:p, extra:w, mehr:w	dehors:w, externe:w, extra:p, extra:w, extérieur:w, hors:w
fore:p	anterioren:w, vor:p, vor:w, vorherige:w, vorwärts:w	antérieur:w, avant:w, devant:w, pro:p, pré:p
hyper:p	hyper:p, top:w, zu:w, übermässig:w, übermäßig:w	excessif:w, hyper:p, outrance:w, supérieur:w, sur:p, trop:w
hypo:p	boden:w, hypo:p, niedrig:w, unten:w, unter:p, unter:w	dessous:w, endessous:w, hypo:p, inférieur:w, sous:p, sous:w, sub:p
in:p	gegenteil:w, in:w, innen:w, nicht:w, ohne:w, umgekehrt:w, un:p	contraire:w, in:p, inverse:w, ir:p, non:w, pas:w, sans:w
inter:p	dazwischen:w, inter:p, mitten:w, unter:w, zwischen:w	entr:p, entre:p, entre:w, inter:p
intra:p	in:w, innen:w, innerhalb:w, intern:w, intra:p	dans:w, dedans:w, entre:w, interne:w, intra:p, intra:w, intérieur:w
ir:p	gegenteil:w, nicht:w, ohne:w, umgekehrt:w, un:p, un:w	contraire:w, in:p, inverse:w, ir:p, non:w, pas:w, sans:w
meta:p	met:p, meta:p, meta:w, oben:w, top:w, über:w, überlegen:w	audessus:w, dessus:w, meta:p, supérieur:w, sur:w
micro:p	klein:w, micro:p, mikro:p, winzig:w	micro:p, minuscule:w, petit:w
mid:p	dazwischen:w, halb:w, hälfte:w, mid:p, mitt:p, mitte:w, mitter:w, teilweise:w, während:w	durant:w, en_partie:w, mi:p, moitié:w, semi:p
mis:p	falsch:w, krank:w, miss:p, schlecht:w	erroné:w, faux:w, mal:w, mauvais:w, mé:p
mono:p	allein:w, ein:w, einzeln:w, einzig:w, einzigartig:w, mon:p, mono:p	mono:p, seul:w, un:w, uni:p, unique:w
multi:p	mehr:w, mehrere:w, multi:p, verschiedene:w, viel:w, viele:w	multi:p, multiple:w, plusieurs:w, poly:p
neo:p	erneuert:w, jung:w, neo:p, neu:w, neun:w	neuf:w, nouveau:w, néo:p
non:p	anti:w, gegenteil:w, nicht:p, nicht:w, non:w, umgekehrt:w, un:p	contr:p, contre:p, ne_pas:w, non:p, non:w, pas:w
out:p	aus:p, ausgehen:w, ausserhalb:w, außerhalb:w, extern:w, extra:w, hinaus:p, out:w, äussere:w, äußere:w, über:p	dehors:w, externe:w, extra:p, extra:w, extérieur:w, hors:w, sortir:w
over:p	oberhalb:w, zu_viel:w, über:p, über:w, überlegen:w, übermässig:w, übermäßig:w	audessus:w, dessus:w, excessif:w, hyper:p, outrance:w, super:p, super:w, supérieur:w, sur:p, sur:w, trop:w
para:p	durch:w, nahe:w, neben:w, par:p, para:p, via:w	au_moyen_de:w, côté:w, par:p, para:p, à_côté_de:w, à_travers:w
peri:p	herum:w, peri:p	autour:w, péri:p
poly:p	mehr:w, mehrere:w, poly:p, verschiedene:w, viel:w, viele:w	multi:p, multiple:w, plusieurs:w, poly:p
post:p	hinter:w, hintere:w, nach:p, nach:w, post:p	après:w, post:p, post:w, postérieur:w
pre:p	anterioren:w, pre:p, vor:p, vor:w	antérieur:w, avant:w, devant:w, pro:p, pré:p
pro:p	für:w, günstig:w, pro:p, pro:w, vor:w, zugunsten:w	antérieur:w, avant:w, devant:w, en_faveur_de:w, favorable:w, pour:w, pro:p, pré:p

re:p	erneut_starten:w, etwas_wieder:w, neu:w, re:p, red:p, wieder:w	encore:w, r:p, re:p, recommencer:w, refaire:w, ré:p, à_nouveau:w
self:p	auto:p, eigen:w, eigenen:w, persönlich:w, selbst:p, selbst:w	auto:p, propre:w, soimême:w
semi:p	halb:p, halb:w, hälfte:w, in_teil:w, semi:p, teilweise:w	durant:w, en_partie:w, mi:p, moitié:w, semi:p
sub:p	hypo:p, sub:p, unter:p, unter:w, unterlegen:w	dessous:w, endessous:w, hypo:p, inférieur:w, sous:p, sous:w, sub:p
super:p	für:w, hinaus:w, hoch:p, oberhalb:w, super:p, super:w, zu_viel:w, über:p, über:w, überlegen:w, übermässig:w, übermäßig:w	audessus:w, dessus:w, excessif:w, outrance:w, super:p, super:w, supérieur:w, sur:p, sur:w, trop:w
tel:p	abstand:w, entfernt:w, fern:w, tel:p, tele:p	distant:w, tel:p, tél:p, télé:p
tele:p	abstand:w, entfernt:w, fern:w, tel:p, tele:p	distant:w, tel:p, tél:p, télé:p
trans:p	durch:w, hindurch:w, quer_durch:w, trans:p	:w, trans:p, à_travers:w
tri:p	drei:w, tri:p, trio:w	tri:p, trio:w, trois:w, à_trois:w
ultra:p	jenseits:w, ultra:p, über:w, über_hinaus:w, überlegen:w	dessus:w, supérieur:w, sur:w, ultra:p
un:p	gegenteil:w, nicht:w, ohne:w, umgekehrt:w, un:p	contraire:w, in:p, inverse:w, ir:p, non:w, pas:w, sans:w
under:p	hypo:p, unter:p, unter:w, unterlegen:w	dessous:w, endessous:w, inférieur:w, sous:p, sous:w
uni:p	allein:w, ein:p, einheitlich:w, einmal:w, einzig:w, einzigartig:w, uni:p	mono:p, seul:w, un:w, uni:p, unique:w
adeno:c	adeno:c, druse:w, gangliom:c, ganglion:c, geschwulst:w, glandula:c	adéno:c, ganglion:w, glande:w, glandulaire:w
adreno:c	adrenal:c, nebennieren:w	adréno:c, surrénal:w
aemia:c	blut:w, ämie:c	hémie:c, sang:w, émie:c, émique:c
afro:c	afrikan:w, afrikanisch:w, afro:c	africain:w, afro:c
algia:c	algie:c, schmerz:w, schmerzzustand:w	algie:c, algique:c, douleur:w
amino:c	amin:c, amine:w, amino:c	amine:w, amino:c
andr:c	andro:c, mann:w, männlich:w	andr:c, andro:c, homme:w, masculin:w, mâle:w
andro:c	andro:c, mann:w, männlich:w	andr:c, andro:c, homme:w, masculin:w, mâle:w
angi:c	angio:c, blutgefäß:w	angi:c, angio:c, vaisseau:w
angio:c	angio:c, blutgefäß:w, schiff:c	angi:c, angio:c, vaisseau:w
anthra:c	anthra:c, kohle:w	anthra:c
aroma:c	aroma:c, aromat:c, aromat:w	aroma:c, arôme:w
arthr:c	arthr:c, arthri:c, arthro:c, bandscheiben:w, knochen:w	arthr:c, arthro:c, articulation:w
astro:c	astero:c, astro:c, astron:c, stern:w	astr:c, astre:w, astro:c, étoile:w
audio:c	audi:c, audio:c, gehör:w, hörbar:w	audio:c, son:w, écouter:w
bio:c	bio:c, biologie:w, biologisch:w	bio:c, biologie:w, biologique:w
blast:c	blast:c, blasto:c, embryo:w, embryonal:w, keim:w	blast:c, blastique:c, blastique:w, blasto:c, embryon:w, embryonnaire:w, germe:w, immature:w

blastic:c	blastom:c, blastom:w, keim:w	blast:c, blastique:c, blastique:w, blasto:c, blastome:w, embryon:w, embryonnaire:w, germe:w, immature:w
brachy:c	brachy:c, brachys:c, kurz:w	brachy:c, court:w, raccourcissement:w
bromo:c	brom:c, brome:w, bromo:c, duft:w, riechend:w	brom:c, brome:w, bromide:w, bromo:c, odeur:w, puanteur:w
broncho:c	bronch:c, bronchial:w, bronchitis:w, broncho:c, bronchus:w	bronch:c, bronche:w, bronchio:c, broncho:c
calc:c	calcium:c, calcium:w, kalk:c, verkalkung:w	calc:c, calcium:w, calco:c
carbo:c	karb:c, karbo:c, kohl:w, kohlenstoff:w	carb:c, carbo:c, carbone:w, charbon:w
carcino:c	carcinome:w, karzino:c, karzinom:c, karzinom:w, kreb:w, krebs:w, onco:c, tumor:w	cancer:w, cancér:c, cancéro:c, carcin:c, carcino:c, onco:c
cardio:c	herz:w, kard:c, kardi:c, kardio:c	cardi:c, cardiaque:w, cardio:c, coeur:w
centro:c	zentr:c, zentro:c, zentrum:w	centr:c, centre:w, centrie:c, centrique:c, centrisme:c, centro:c
cerebro:c	gehirn:w, kohlenstoff:w, zerebr:c, zerebralen:w, zerebro:c	cerveau:w, cérébr:c, cérébral:w, cérébro:c
charactero:c	charakter:c, charaktero:c, persönlichkeit:w	charactér:c, charactéro:c
chem:c	chem:c, chemi:c, chemikalie:w, chemikalien:w, chemische:w, chemo:c	chimie:w, chimio:c, chimique:w, chémo:c
chemi:c	chem:c, chemi:c, chemikalie:w, chemikalien:w, chemische:w, chemo:c	chimie:w, chimio:c, chimique:w, chémo:c
chemo:c	chem:c, chemi:c, chemikalie:w, chemikalien:w, chemische:w, chemo:c	chimie:w, chimio:c, chimique:w, chémo:c
chloro:c	chlor:c, chlore:w, chloro:c	chlor:c, chlorine:w, chloro:c, vert:w
chondr:c	chondr:c, chondri:c, chondriale:c, chondrio:c, chondrium:c, chondro:c, knorpel:w	cartilage:w, chondr:c, chondrie:c, chondro:c, grain:w, granulaire:w
chondri:c	chondr:c, chondri:c, chondriale:c, chondrio:c, chondrium:c, chondro:c, knorpel:w	cartilage:w, chondr:c, chondrie:c, chondro:c, grain:w, granulaire:w
chondrion:c	chondr:c, chondri:c, chondriale:c, chondrio:c, chondrium:c, chondro:c, knorpel:w	cartilage:w, chondr:c, chondrie:c, chondro:c, grain:w, granulaire:w
chondro:c	chondr:c, chondri:c, chondriale:c, chondrio:c, chondrium:c, chondro:c, knorpel:w	cartilage:w, chondr:c, chondrie:c, chondro:c, grain:w, granulaire:w
choreo:c	choreo:c, tanz:w	choré:c, choréo:c, danse:w
chorio:c	chori:c, chorio:c, chorion:w, choroidal:w, chorr:c, membran:w	chor:c, chori:c, chorio:c, chorion:w, choro:c, choroïde:w
chromato:c	chrom:c, chromato:c, farbe:w	chrom:c, chromat:c, chromato:c, chromo:c, couleur:w
chromo:c	chrom:c, chromato:c, farbe:w	chrom:c, chromat:c, chromato:c, chromo:c, couleur:w
chrono:c	chron:c, chrono:c, zeit:w	chron:c, chrono:c, temps:w
claustro:c	klaustr:c, klaustro:c, käfig:w, raum:w	claustr:c, claustro:c, cloîtrer:w, clôture:w
clinico:c	klinik:c, klinik:w, klinisch:w	clinico:c, clinique:w

colo:c	darm:w, kolo:c	coli:c, colo:c, colon:w, intestin:w
cryo:c	eis:w, frost:w, kry:c, kryo:c, kälte:w	cry:c, cryo:c, froid:w
cyclo:c	rund:w, zykl:c, zyklo:c	cercle:w, cycl:c, cycle:c, cyclique:w, cyclo:c, rond:w, roue:w, récurrent:w
cyte:c	cyto:c, zell:w, zyt:c, zyto:c	cellule:w, cyt:c, cyte:c, cyto:c
cyto:c	cyto:c, zell:w, zyt:c, zyto:c	cellule:w, cyt:c, cyte:c, cyto:c
demo:c	bevölkerung:w, dem:c, demo:c, staat:w, volk:w	démo:c, peuple:w, population:w
densito:c	densit:c, densito:c, dichte:c, dichte:w	dense:w, densit:c, densito:c, densité:w
dermo:c	derm:c, derm:w, derma:c, derme:c, dermo:c, haut:w	dermat:c, dermato:c, dermo:c, peau:w
di:c	di:c, dia:c, doppelt:w, paar:w, zwei:w	bi:c, deux:w, di:c, double:w
dosi:c	dosi:c, dosis:c, dosis:w, messen:w, rationieren:w	dose:w, dosi:c
eco:c	eco:c, umwelt:w, öko:c, ökologie:w	éco:c, écologie:w, écologique:w
ectasia:c	ausdehnung:w, ectasia:c, erweiterung:w	dilatation:w, ectasie:c
ectomy:c	ablation:w, abschnitt:w, ektomie:c, entfernung:w, resektion:w	ablation:w, coupe:w, ectomie:c, section:w
electro:c	electr:c, electro:c, elektr:c, elektrisch:w, elektro:c, strom:w	électr:c, électricité:w, électrique:w, électro:c
embryo:c	embr:c, embryo:c, embryo:w, leibesfrucht:w, ungeborene:w, ur:c, ur:w	embry:c, embryo:c, embryon:w, foetus:w
endocrino:c	absondern:w, endokrin:c, endokrin:w, endokrinen:w, sekret:w, sekretion:w	endocrino:c
epi:c	auf:w, epi:c, über:w	dessus:w, supérieur:w, sur:w, épi:c
epidemio:c	epidem:c, epidemi:c, epidemie:w, epidemio:c	épidémio:c
erythro:c	erythr:c, erythro:c, rot:w	rouge:w, érythro:c
estro:c	estr:c, estro:c, weibliches_geschlechtshormon:w, östr:c, östro:c, östrogen:w	oestro:c, oestrogène:w
fibro:c	faser:w, fiber:w, fibr:c, fibro:c	fibr:c, fibre:w, fibrille:w, fibro:c, fibrosis:w, filament:w, lobe:w
fluoro:c	fluor:c, fluor:w, fluoreszenz:w, fluori:c, fluoro:c	fluor:c, fluorine:w, fluoro:c, flurorescence:w
fluro:c	fluor:c, fluor:w, fluoreszenz:w, fluori:c, fluoro:c	fluro:c
gastro:c	bauch:w, gaster:c, gastr:c, gastro:c, magen:w	estomac:w, gastr:c, gastrique:w, gastro:c, ventre:w
gen:c	ausgangspunkt:w, auslöser:w, gen:c, hervorbringend:w, hervorgebracht:w, verursachend:w, verursacht:w	gène:c
gene:c	ausgangspunkt:w, auslöser:w, gen:c, hervorbringend:w, hervorgebracht:w, verursachend:w, verursacht:w	gène:c, producteur:w, produire:w
geneous:c	ausgangspunkt:w, auslöser:w, gen:c, hervorbringend:w, hervorgebracht:w, verursachend:w, verursacht:w	gène:c, produit:w, résultant:w

genic:c	gen:c, hervorbringend:w, hervorgebracht:w, verursachend:w, verursacht:w	génique:c
genicity:c	genick:c, hals:w, hervorbringend:w, hervorgebracht:w, kopfgelenk:w, verursachend:w, verursacht:w	génicité:c
geno:c	genitalien:w, geschlecht:c, geschlecht:w	gén:c, génit:c, génital:w, génito:c, géno:c
gens:c	ausgangspunkt:w, auslöser:w, gen:c, hervorbringend:w, hervorgebracht:w, verursachend:w, verursacht:w	gènes:c
glyco:c	glyk:c, glyko:c, zucker:w	gluc:c, gluco:c, glucose:w, glyc:c, glyco:c, sucre:w
gonado:c	geschlechtsdrüse:w, gonad:c, gonade:w, gonado:c, keimdrüse:w	gonade:w, gonado:c
grapher:c	aufnehmen:w, graf:c, graph:c, schreiben:w	enregistrer:w, grapheur:c, écrire:w, écriture:w
graphic:c	aufgernommen:w, graphisch:c, schriftlich:w	enregistrer:w, graphique:c, graphique:w, écrire:w, écriture:w
graphy:c	aufnahme:w, graphie:c, schreiben:w	enregistrer:w, graphie:c, écrire:w, écriture:w
gynaeco:c	frau:w, gynäk:c, gynäko:c	femme:w, gyn:c, gyno:c, gynéco:c
gynec:c	frau:w, gynäk:c, gynäko:c	
gyneco:c	frau:w, gynäk:c, gynäko:c	
haemato:c	blut:w, hämat:c, hämato:c	hém:c, hémat:c, hémato:c, hémo:c, sang:w
hemato:c	blut:w, hämat:c, hämato:c	hém:c, hémat:c, hémato:c, hémo:c, sang:w
hepato:c	hepat:c, hepato:c, leber:w	foie:w, hépar:c, héparo:c, hépat:c, hépato:c
hetero:c	hetero:c, unterschiedlich:w	autre:w, hétéro:c
histio:c	gewebe:w, histo:c	hist:c, histi:c, histio:c, histo:c, tissu:w
histo:c	gewebe:w, histo:c	hist:c, histi:c, histio:c, histo:c, tissu:w
homo:c	entsprechend:w, gleich:w, gleichartig:w, homo:c	homo:c, homéo:c, ressemblant:w, semblable:w
hydro:c	hydr:c, hydro:c, wasser:w	eau:w, hyd:c, hydr:c, hydro:c, liquide:w
hypno:c	hypn:c, hypno:c, schlaf:w	hypn:c, hypno:c, hypnose:w, sommeil:w
hypo:c	hyp:c, hypo:c, unter:c, unter:w	bas:w, dessous:w, endessous:w, faible:w, hypo:c, inférieur:w, sous:c, sous:w, sub:c
hyster:c	gebärmutter:w, hyster:c, hysteri:c, hysterio:c	hystér:c, hystéro:c, utérus:w
immuno:c	immun:c, immun:w, immunität:w, immuno:c	dispense:w, exemption:w, immun:c, immunité:w, immuno:c, remise:w
iso:c	gleich:w, iso:c	iso:c, régulier:w, symétrique:w, uniforme:w, égal:w
kinesio:c	bewegung:w, kinesi:c, kinesio:c	kin:c, kinési:c, kinésio:c, mouvement:w
leio:c	glatt:w, leio:c	lio:c, lisse:w, lisso:c, léio:c
loco:c	lokal:w, loko:c	local:w, loco:c

logic:c	logik:w, logiken:w, logisch:c, logish:w	logique:c, logique:w
logist:c	fachmann:w, loge:c, logik:w, logiken:w, logish:w, wissenschaftler:w	logiste:c
logy:c	lehre:w, logie:c, logik:w, logiken:w, logish:w, wissenschaft:w	logie:c
lymph:c	lymph:c, lymph:w, lympho:c	lymph:c, lymphe:w, lympho:c
lympho:c	lymph:c, lymph:w, lympho:c	lymph:c, lymphe:w, lympho:c
macro:c	gross:w, lang:w, makr:c, makro:c	gros:w, incluant:w, long:w, macrie:c, macro:c
mamma:c	brust:w, mamm:c, mamma:c, mammo:c	mamelle:w, mamm:c, mamma:c, mammaire:w, mammo:c, sein:w
mammo:c	brust:w, mamm:c, mamma:c, mammo:c	mamelle:w, mamm:c, mamma:c, mammaire:w, mammo:c, sein:w
mast:c	brust:w, mast:c, mastie:c, masto:c	mamelle:w, mammaire:w, mast:c, masto:c, mastoïd:c, mastoïdo:c, sein:w
mastia:c	brust:w, mast:c, mastie:c, masto:c	mastie:c
medio:c	medio:c, mitte:w	milieu:w, médi:c, média:c, médio:c
meric:c	mer:c, mere:c, segment:w	mère:c, mér:c, mérisme:c, méro:c, partie:w, élément:w
meta:c	absiedelungen:w, meta:c, stamm:w	audessus:w, dessus:w, meta:c, méta:c, supérieur:w, sur:w
methodo:c	method:c, methode:w, weise:w	méthodo:c
metric:c	mass:w, messen:w, metrisch:c	mesure:w, mesurer:w, métrique:c
metry:c	mass:w, messen:w, metrie:c, metrik:c	calcul:w, calculer:w, mesure:w, mesurer:w, métrie:c
micro:c	fein:w, gering:w, klein:w, mikr:c, mikro:c, millionstel:w	micro:c, minuscule:w, petit:w
milli:c	milli:c, tausendstel:w	milli:c, millième:w
mito:c	faden:w, fadenförmige:w, mito:c	filament:w, mito:c, mitose:w
mmastia:c	brust:w, mast:c, mastie:c, masto:c	mastie:c
mono:c	allein:w, einzeln:w, einzig:w, mon:c, mono:c	mono:c, seul:w, un:w, uni:c, unique:w
morpho:c	form:w, morpho:c	forme:w, morph:c, morphie:c, morphique:c, morphisme:c, morpho:c
musculo:c	musculo:c, muskel:c, muskel:w, muskulär:w	muscle:w, muscul:c, musculo:c, myo:c
mycin:c	hyaluronsäure:w, mycin:c	champignon:w, mucéto:c, myc:c, myce:c, mycine:c, myco:c, mycète:c, mycét:c
myco:c	myko:c, pilz:w	
myo:c	muskel:w, my:c, myo:c	muscle:w, my:c, myo:c
neo:c	erneuert:w, jung:w, neo:c, neu:w	jeune:w, nouveau:w, né:c, néo:c
neuro:c	nerven:c, nerven:w, neur:c	nerf:w, neur:c, neural:c, neurie:c, neuro:c
nucleo:c	kern:w, nucleo:c, zellkern:w	noyau:w, nucleus:w, nuclé:c, nucléo:c
oculo:c	augen:w, okular:c, okulär:c	ocul:c, oculo:c, oeil:w, ophtalm:c, ophtalmo:c
oestro:c	estr:c, estro:c, weibliches_geschlechtshormon:w, östr:c, östro:c, östrogen:w	oestr:c, oestro:c

oligo:c	0:w, olig:c, oligo:c, spuren:c	insuffisant:w, olig:c, oligo:c, petit:w, peu:w
onco:c	anschwellung:w, carcinom:w, karcinom:w, karzin:c, karzino:c, krebs:w, onc:c, onco:c	courbure:w, grosseur:w, onc:c, onch:c, oncho:c, onco:c, tumeur:w, volume:w
oophor:c	eierstock:w, oophor:c, oophori:c, ovar:c, ovarek:c, ovari:c, ovariek:c, ovarien:w, ovarium:w	oo:c, oophor:c, oophoro:c, ovaire:w, ovar:c, ovari:c
ophthalmo:c	augen:w, oculär:c, okular:c, ophtalmo:c	ocul:c, oculo:c, oeil:w, ophtalm:c, ophtalmo:c
organo:c	organ:c, organ:w, organisch:w, organo:c	organ:c, organe:w, organo:c, outil:w
ortho:c	aufrecht:w, ortho:c	correct:w, droit:w, normal:w, ortho:c, régulier:w
osteo:c	knochen:w, osteo:c	os:w, ossi:c, osté:c, ostéo:c
ovar:c	eierstock:w, oophor:c, oophori:c, ovar:c, ovarek:c, ovari:c, ovariek:c, ovarien:w, ovarium:w	oo:c, oophor:c, oophoro:c, ov:c, ovaire:w, ovar:c, ovari:c, ovario:c, ovo:c
ovari:c	eierstock:w, oophor:c, oophori:c, ovar:c, ovarek:c, ovari:c, ovariek:c, ovarien:w, ovarium:w	oo:c, oophor:c, oophoro:c, ovaire:w, ovar:c, ovari:c
patho:c	behandlung:w, krankheit:w, leiden:w, path:c, pathie:c, patho:c	maladie:w, path:c, pathie:c, patho:c, souffrance:w
pathy:c	behandlung:w, krankheit:w, leiden:w, path:c, pathie:c, patho:c	maladie:w, path:c, pathie:c, patho:c, souffrance:w
pharmaco:c	arzneimittel:w, pharmaco:c	pharmac:c, pharmaco:c, remède:w
pheno:c	pheno:c, phenol:w, phäno:c	briller:w, phén:c, phénique:c, phéno:c, phénol:c, phénol:w, phényl:c
phospho:c	leuchten:w, licht:w, phosph:c, phosphat:c, phosphato:c, phospho:c, phosphoreszenz:w	lumineux:w, lumière:w, phosph:c, phosphat:c, phosphato:c, phospho:c, phosphore:w
photo:c	foto:c, foto:w, fotografisch:w, licht:w, photo:c	lumière:w, photo:c
physio:c	nature:w, phys:c, physikalisch:w, physio:c	nature:w, physio:c
phyto:c	pflanze:w, pflanzen:w, phyt:c, phyto:c	bourgeon:w, excroissance:w, phyt:c, phyto:c, plante:w, végétal:w
plasia:c	plasie:c, plastisch:c, plastischer:w, umformbarkeit:w, verformung:w	modeler:w, plase:c, plasie:c, plasique:c
plastia:c	plasie:c, plastisch:c, plastischer:w, umformbarkeit:w, verformung:w	
plasty:c	plasie:c, plastisch:c, plastischer:w, umformbarkeit:w, verformung:w	plastie:c, plastique:c, plastique:w, réparation:w
plasy :c	plasie:c, plastisch:c, plastischer:w, umformbarkeit:w, verformung:w	
progesto:c	gesta:c, progest:c, progesto:c, schwangerschaft:w	progesto:c
proto:c	erster:w, prot:c, proto:c, ur:w, vorderster:w, wichtigster:w	premier:w, prot:c, proto:c, protéin:c, protéino:c
pseudo:c	anschein:w, falsch:w, pseudo:c, schein:c	faux:w, mensonger:w, menteur:w, pseud:c, pseudo:c
psycho:c	pscych:c, psych:c, psyche:w, psychische:w, psycho:c, psychologisch:w, seele:w	esprit:w, mental:w, psych:c, psycho:c, psychologie:w, psychologique:w

radio:c	radio:c, strahl:w, strahlen:w, strahlung:w	radi:c, radiation:w, radio:c, radio:w, rayon:w, rayonnement:w
reflexo:c	reaktion:w, reflex:c, reflex:w, reflexo:c, widerstand:w	retourner:w, réflex:c, réflexe:w, réflexo:c
retin:c	netzhaut:w, retin:c, retino:c	rétin:c, rétine:w, rétino:c
retino:c	netzhaut:w, retin:c, retino:c	rétin:c, rétine:w, rétino:c
retro:c	hinten:w, retro:c, rückgang:w, stange:w, stäbchenförmigen:w, zurück:w	arrière:w, avant:w, derrière:w, rétro:c
rhabdo:c	rhabdo:c, rhabdoider:w, rhadb:c, stäbchenförmige:w	baguette:w, rhabd:c, rhabdo:c
ribo:c	rib:c, ribo:c, ribose:w	ribo:c, ribose:w
scope:c	betrachtung:w, skop:c, untersuchung:w	examiner:w, observer:w, regarder:w, scope:c, scopie:c, scopique:c
scopol:c	besonderer_alcaloid:w, scopol:c, skopol:c	scopol:c
socio:c	gesellschaftlich:w, gesellschafts:w, sozial:w, sozialen:w, sozio:c	social:w, socio:c, société:w
some:c	körper:w, leich:w, som:c, some:c, somen:c	corps:w, soma:c, somat:c, somato:c, some:c, somie:c
sono:c	akustik:w, sona:c, sono:c	son:w, sono:c
soya:c	soja:c, soja:w	soya:c
spectro:c	spektr:c, spektro:c, spektrum:w	spectr:c, spectre:w, spectro:c
stat:c	regulierung:w, stabilisierung:w, stat:c	stat:c
stereo:c	fest:w, hart:w, stereo:c	solide:w, stéréo:c
steroido:c	steroid:c, steroid:w, steroido:c	stérol:w, stérone:c, stéroïd:c, stéroïdo:c
strepto:c	geflochten:w, gewunden:w, strept:c, strepto:c	contourné:w, strep:c, strepto:c
tetra:c	tetr:c, tetra:c, vier:w	quadr:c, quadri:c, quatre:w, tétr:c, tétra:c
thelial:c	brustwarze:w, thel:c, thelal:c, thelialen:c	mamelon:w, thelial:c, thélial:c
thelium:c	brustwarze:w, thel:c, thelal:c, thelialen:c	mamelon:w, thelium:c, thélium:c
thermo:c	hitze:w, temperatur:w, thermo:c	chaleur:w, chaud:w, chauffer:w, therm:c, thermo:c
thrombo:c	blutgerinnsel:w, thromb:c, thrombo:c	caillot:w, thromb:c, thrombo:c
tumor:c	carcinom:w, karzinom:w, tumor:c, tumor:w	gonfler:w, tumeur:w, tumor:c, tumoral:w, tumori:c, tumoro:c
tumori:c	carcinom:w, karzinom:w, tumor:c, tumor:w	gonfler:w, tumeur:w, tumor:c, tumoral:w, tumori:c, tumoro:c
tumour:c	carcinom:w, karzinom:w, tumor:c, tumor:w	gonfler:w, tumeur:w, tumor:c, tumoral:w, tumori:c, tumoro:c
tumouri:c	carcinom:w, karzinom:w, tumor:c, tumor:w	gonfler:w, tumeur:w, tumor:c, tumoral:w, tumori:c, tumoro:c
vasculo:c	blutgefäss:w, blutgefäß:w, gefass:w, gefäß:w, vasculo:c, vaskulo:c, vaskulären:w, vaso:c	vaisseau:w, vas:c, vascul:c, vasculo:c, vaso:c
vaso:c	blutgefäss:w, blutgefäß:w, gefass:w, gefäß:w, vasculo:c, vaskulo:c, vaskulären:w, vaso:c	vaisseau:w, vas:c, vascul:c, vasculo:c, vaso:c

xeno:c	fremd:w, fremde:w, fremder:w, gast:w, xen:c, xeno:c	xén:c, xéno:c, étranger:w
zygous:c	paarig:w, verbunden:w, zygo:c, zygos:c, zygote:c	paire:w, zyg:c, zygo:c, zygote:c, zygotique:c
ability:s	barkeit:s, ilität:s, kapazität:w, keit:s	abilité:s, capacité:w, ibilité:s
able:s	abel:s, bare:s, gemacht_werden_kann:w, können:w	able:s, ible:s, pouvoir:w
hood:s	heit:s, schaft:s, sensein:w, tat:s, tat:w, zustand:w	fait:w, état:w
less:s	a:p, ab:p, abs:p, los:s, nicht:w, ohne:w	a:p, ab:p, aucun:w, privé:w, sans:w
like:s	gleich:w, lich:s, wie:w, ähnlich:w	ressembler:w, semblable:w
ly:s	mit:w, mässig:s, mäßig:s, weg:w, weise:w	avec:w, ement:s, façon:w, manière:w, ment:s
wise:s	richtung:w, weise:s, weise:w, während:w	sens:w

B.3.4 Familles morphologiques

Les familles morphologiques ont été constituées à l'aide de l'algorithme de Porter (1980).

	EN	FR	DE
Familles morphologiques	5 835 → 14 659	7 049 → 17 410	7 348 → 15 818

TABLE B.10 – Taille des dictionnaires de synonymes (nb. d'entrées)

B.3.4.1 Allemand

Évaluation

L'évaluation a porté sur 116 mots regroupés dans 50 familles. Nous avons donc obtenu 6 670 paires de mots classées comme appartenant à la même famille morphologique ou n'appartenant pas à la même famille morphologique. Lorsque nous n'avons pas pu déterminer si une paire de mots appartenait ou pas à une même famille, nous l'avons ôtée de l'évaluation (colonne "indécis").

		annotation manuelle		
		même famille	famille différente	indécis
raciniseur	même famille	72	1	16
	famille différente	0	4 680	1 901

TABLE B.11 – Évaluation des familles morphologiques

angstig	ängstigen:ADJECTIF, ängstigen:PARTICIPE, ängstigen:VERBE
pinkfarb	pinkfarbig:ADJECTIF, pinkfarben:ADJECTIF
Mammographiegerat	Mammographiegeräte:NOM, Mammographiegerät:NOM
sympath	sympathisch:ADJECTIF, sympathisch:NOM
lakon	lakonisch:ADJECTIF, lakonisch:NOM
abruck	abrücken:VERBE, abrücken:ADJECTIF, abrücken:PARTICIPE, abrücken:NOM
bimmeln	bimmeln:VERBE, bimmeln:ADJECTIF
Zeh	Zehe:NOM, Zeh:NOM
Doblin	Döblins:NOM, Döblin:NOM
McQuaid	McQuaids:NOM, McQuaid:NOM
SEDMitglied	SEDMitglied:NOM, SEDMitglieder:NOM
Pattison	Pattisons:NOM, Pattison:NOM
Strass	Strasse:LEX, Strass:NOM, Strasse:NOM
Danton	Dantons:NOM, Danton:NOM
MDax	MDax:LEX, MDax:NOM
masturbi	masturbieren:PARTICIPE, masturbieren:VERBE
bedank	bedanken:PARTICIPE, bedanken:VERBE
FraktionsFuhr	FraktionsFührung:NOM, FraktionsführerNOM
leck	lecken:VERBE, leck:ADJECTIF, lecken:PARTICIPE, lecker:NOM, lecker:ADJECTIF, leck:NOM
verlob	verloben:PARTICIPE, verloben:VERBE, verloben:NOM, verloben:ADJECTIF
spielenStatt	spielenStätte:NOM, spielenStatt:NOM
Pech	Pech:NOM, Pech:LEX
Gnadl	Gnadl:NOM, Gnadls:NOM
ausraub	ausrauben:PARTICIPE, ausrauben:VERBE, ausrauben:ADJECTIF
Werfel	Werfels:NOM, Werfel:NOM

Extrait B.17 – Familles morphologiques allemandes

B.3.4.2 Anglais

Évaluation

L'évaluation a porté sur 140 mots regroupés dans 50 familles. Nous avons donc obtenu 9 730 paires de mots classées comme appartenant à la même famille morphologique ou n'appartenant pas à la même famille morphologique. Lorsque nous n'avons pas pu déterminer si une paire de mots appartenait ou pas à une même famille, nous l'avons ôtée de l'évaluation (colonne "indécis").

		annotation manuelle		
		même famille	famille différente	indécis
raciniseur	même famille	127	18	1
	famille différente	22	9 286	276

TABLE B.12 – Évaluation des familles morphologiques

fiduciari	fiduciary:ADJECTIF, fiduciary:NOM
careen	careen:VERBE, careen:NOM, careen:PARTICIPE
carous	carouse:PARTICIPE, carousal:NOM
anti-oestrogen	anti-oestrogens:ADJECTIF, anti-oestrogens:NOM, anti-oestrogenic:ADJECTIF, anti-oestrogen:ADJECTIF, anti-oestrogen:NOM
reverber	reverberation:NOM, reverberate:VERBE, reverberate:PARTICIPE
deaf	deafness:NOM, deaf:NOM, deaf:ADJECTIF
p21waf1	p21waf1:NOM_PROPRE, p21waf1:NOM
contemporari	contemporary:NOM, contemporary:ADJECTIF
tend	tended:ADJECTIF, tend:PARTICIPE, tend:VERBE
seldom	seldom:ADVERBE, seldom:ADJECTIF
earli	early:ADJECTIF, early:ADVERBE
shoulder	shoulder:VERBE, shoulder:NOM
idealist	idealist:NOM, idealistic:ADJECTIF
televis	televise:VERBE, televised:ADJECTIF, television:NOM, televise:PARTICIPE
Edmund	Edmunds:NOM_PROPRE, Edmund:NOM_PROPRE
play_against	play_against:VERBE, play_against:PARTICIPE
yoke	yoke:VERBE, yoke:NOM
stink	stink:VERBE, stink:PARTICIPE, stinking:ADJECTIF, stink:NOM
dupe	dupe:NOM, dupe:PARTICIPE
pervas	pervasive:ADJECTIF, pervasiveness:NOM
home	home:PARTICIPE, homely:ADJECTIF, home:VERBE, home:NOM
fidget	fidget:VERBE, fidget:NOM
mind-boggl	mind-boggling:NOM, mind-boggling:PARTICIPE
strappi	strappy:NOM, strappy:ADJECTIF
amass	amass:PARTICIPE, amass:VERBE, amass:NOM

Extrait B.18 – Familles morphologiques anglaises

B.3.4.3 Français

Évaluation

L'évaluation a porté sur 126 mots regroupés dans 50 familles. Nous avons donc obtenu 7 875 paires de mots classées comme appartenant à la même famille morphologique ou n'appartenant pas à la même famille morphologique. Lorsque nous n'avons pas pu déterminer si une paire de mots appartenait ou pas à une même famille, nous l'avons ôtée de l'évaluation (colonne "indécis").

		annotation manuelle		
		même famille	famille différente	indécis
raciniseur	même famille	104	6	0
	famille différente	19	7 746	0

TABLE B.13 – Évaluation des familles morphologiques

Nyon	Nyon:NOM_PROPRE, Nyons:NOM_PROPRE
demeur	demeure:NOM, demeuré:ADJECTIF, demeurer:VERBE, demeuré:NOM, demeurer:PARTICIPE
grommel	grommeler:VERBE, grommeler:PARTICIPE
sadiqu	sadique:NOM, sadique:ADJECTIF
Wild	Wilde:NOM_PROPRE, Wild:NOM_PROPRE, Wilder:NOM_PROPRE
compte-tenu	compte-tenu:ADJECTIF, compte-tenu:NOM
Bo	Bo:NOM_PROPRE, Boé:NOM_PROPRE, Boer:NOM_PROPRE
reten	retenir:PARTICIPE, retenir:VERBE, retenir:ADJECTIF
faire_le_diagnostic	faire_le_diagnostic:PARTICIPE, faire_le_diagnostic:VERBE
decoul	découler:PARTICIPE, découler:VERBE
demol	démolir:ADJECTIF, démolir:VERBE, démolir:PARTICIPE
creol	créole:ADJECTIF, créole:NOM
souten	soutenable:ADJECTIF, soutenir:ADJECTIF, soutenir:VERBE, soutenir:PARTICIPE
abolition	abolitionniste:ADJECTIF, abolitionniste:NOM
coherent	cohérent:ADJECTIF, cohérence:NOM
philanthrop	philanthrope:NOM, philanthropie:NOM, philanthropique:ADJECTIF
nevros	névrosé:NOM, névrosé:ADJECTIF, névrose:NOM
barrag	barragiste:NOM, barrage:NOM
reaffirm	réaffirmation:NOM, réaffirmer:PARTICIPE, réaffirmer:VERBE, réaffirmer:ADJECTIF
artisanal	artisanalement:ADVERBE, artisanal:ADJECTIF
rendre_hommage_	rendre_hommage_à:PARTICIPE, rendre_hommage_à:VERBE
judici	judicieux:ADJECTIF, judicieusement:ADVERBE
Schwarz	Schwarzer:NOM_PROPRE, Schwarz:NOM_PROPRE
Moin	Moins:NOM_PROPRE, Moïn:NOM_PROPRE
sabot	saboter:VERBE, sabot:NOM, saboter:PARTICIPE

Extrait B.19 – Familles morphologiques françaises

B.3.5 Dictionnaires de cognats

Les dictionnaires de cognats ont été construits en suivant la méthode de Hauer et Kondrak (2011).

	EN→FR	EN→DE
# exemples positifs	21 202	7 399
# exemples négatifs	21 202	7 399
taux d'erreur	3,49%	6,93%

TABLE B.14 – Identification de cognats : données d'apprentissage et taux d'erreur

	EN↔FR	EN↔DE
Paires de cognats	6 708	6 391

TABLE B.15 – Taille des dictionnaires spécialisés (nb. d'entrées)

B.3.5.1 Anglais ↔allemand

light:NOM	Licht:NOM
car:NOM	Carl:NOM
function:NOM	Funktion:NOM
bear:PARTICIPE	Beat:NOM
phytoestrogen:NOM	PhytoÖstrogen:NOM
home:VERBE	hom:ADJECTIF
actin:NOM_PROPRE	Action:NOM
department:NOM	Department:NOM
strong:ADJECTIF	streng:ADJECTIF
express:ADJECTIF	Expression:NOM
synthesise:PARTICIPE	Synthese:NOM
assay:VERBE	Assay:LEX
quantity:NOM	Quality:NOM
categorise:VERBE	Kategorie:NOM
randomized:ADJECTIF	Randomized:NOM
extreme:NOM	extrem:ADJECTIF
News:NOM_PROPRE	new:ADJECTIF
chi2:ADJECTIF	Chi:NOM
droop:VERBE	Drop:LEX
Garber:NOM_PROPRE	Gerber:NOM
GAIC:NOM_PROPRE	GAIN:NOM
primarily:ADVERBE	primary:ADJECTIF
mastitis:ADJECTIF	Mastitis:NOM
treat:ADJECTIF	treat:ADJECTIF
senior:ADJECTIF	Sensor:NOM

Extrait B.20 – Dictionnaire de cognat anglais ↔ allemand

B.3.5.2 Anglais ↔ français

car:NOM	car:NOM
preferred:ADJECTIF	préférer:PARTICIPE
choose:VERBE	chose:NOM
function:NOM	fonction:NOM
phytoestrogen:NOM	phyto-estrogène:NOM
homogeneity:NOM	homogénéité:NOM
MSKCC:NOM_PROPRE	MSKCC:NOM_PROPRE
home:VERBE	homme:LEX
actin:NOM_PROPRE	action:NOM
department:NOM	département:NOM
dissociate:VERBE	dissocier:PARTICIPE
express:ADJECTIF	exprès:ADJECTIF
synthesise:PARTICIPE	synthèse:NOM
cut:ADJECTIF	cut:LEX
quantity:NOM	quantité:NOM
categorise:VERBE	catégorie:NOM
extreme:NOM	extrême:NOM
chi2:ADJECTIF	Chi2:NOM_PROPRE
embolus:NOM	Emboles:NOM_PROPRE
bombard:VERBE	bomber:VERBE
inject:VERBE	injecté:ADJECTIF
P=0.009:NOM_PROPRE	p=0.003:NOM
mastitis:ADJECTIF	mastite:NOM
senior:ADJECTIF	senior:ADJECTIF
degenerate:VERBE	dégénérer:VERBE

Extrait B.21 – Dictionnaire de cognat anglais ↔ français

B.3.6 Probabilités de traduction de parties du discours

Les probabilités de traduction de parties du discours ont été acquises sur le corpus EMEA (Tiedemann, 2009).

	EN→FR	EN→DE
# phrases alignées	373 127	363 982
# suites de parties du discours alignées	191 854	108 612

TABLE B.16 – Acquisition de probabilités de traduction de partie du discours : taille des données

B.3.6.1 Anglais → allemand

_ADVERBE	NOM	0,2014960293
_ADVERBE	_ADVERBE	0,1946567909
_ADVERBE	ADJECTIF	0,0841178903
_ADVERBE	NOM NOM	0,0477002124
_ADVERBE	VERBE	0,0292847442
_ADVERBE	_PREP	0,0180966766
_ADVERBE	ADJECTIF NOM	0,0164627438
_ADVERBE	_ADVERBE NOM	0,0157798611
_ADVERBE	PARTICIPE	0,0150926129
_ADVERBE	_NUM	0,0146408467
_ADVERBE	_CONJ	0,0109107886
_ADVERBE	NOM NOM NOM	0,0104185454
_ADVERBE	NOM PARTICIPE	0,0102317638
_NUM	_NUM _DET _PREP _PREP _DET	0,000000000315865288145686
_AUX	_PREP _ELIM _CONJ _DET NOM	0,0000000235290034672843
_AUX	ADJECTIF NOM PARTICIPE	0,00000128821293983382
_PUNCT	_DET _PRO _PREP _PUNCT	0,00000000714239298007896
_PRO	_PRO _PREP _AUX	0,00000244990842009058
_PREP	_DET _DET _ADVERBE _DET _DET	0,000000031817650318813 8
ADJECTIF	ADJECTIF _PREP VERBE _PUNCT	0,000000006212831634258 78
_PREP	_AUX _PRO _PUNCT _PUNCT _PUNCT	0,000000020454203776380 3
_PRO	_ADVERBE _CONJ _DET	0,000000049998131022256 7
PARTICIPE	NOM _NUM VERBE	0,00000168907402961649
PROPRE	VERBE _DET NOM _PART	0,000000065416742791788 3
_NUM	_NUM _NUM _PREP _DET	0,000000000282022578701506
_DET	_AUX _PREP _DET _PUNCT _PUNCT	0,000000068565138209766 6

Extrait B.22 – Probabilités de traduction de parties du discours anglais → allemand

B.3.6.2 Anglais → français

_NUM	_NUM	0,5869649849
_PUNCT	_PUNCT	0,4934448322
NOM	NOM	0,4835526372
PROPRE	PROPRE	0,4696379662
_ELIM	NOM	0,3716330294
ADJECTIF	NOM	0,3414517609
_DET	_DET	0,2581942818
ADVERBE	ADVERBE	0,2562987002
VERBE	VERBE	0,2549051439
PARTICIPE	NOM	0,2273221307
_PRO	NOM	0,2237807336
_AUX	VERBE	0,2195352295
VERBE	NOM	0,209632155
VERBE	_NUM VERBE ADJECTIF NOM	0,00000143047001026882
_PRO	_PUNCT NOM _PUNCT	0,00000497600446840555
_DET	_PUNCT _AUX _DET NOM	0,000000971339457971694
ADJECTIF	ADJECTIF VERBE ADJECTIF NOM _ADVERBE	0,00000135277384277605
VERBE	NOM PARTICIPE _CONJ VERBE	0,00000757334158528157
_AUX	_AUX _PREP _DET _PUNCT	0,00000189408477911639
_PREP	_PUNCT _DET _DET _PUNCT	0,00000789759534699127
NOM	_PUNCT NOM _PART _PUNCT	0,0000000296264864581992
_PUNCT	NOM _PREP _PUNCT _PREP	0,0000000324654226367225
_DET	_NUM NOM _CONJ NOM	0,000000342825691048833
_CONJ	_PREP _PUNCT _PRO NOM	0,000000921022892234539

Extrait B.23 – Probabilités de traduction de parties du discours anglais → français

B.4 Termes sources

La liste des termes sources a été construite de façon semi-supervisée et validée manuellement (cf. section 5.2). Les termes sources à traduire sont les termes pour lesquels le dictionnaire bilingue n'a pas pu donner de traduction qui soit attestée dans les textes cibles.

	FR	DE
Termes sources morphologiquement construits	2 025	2 025
Termes sources traduits avec le dictionnaire bilingue généraliste	186	201
Termes sources à traduire (S)	1 839	1 824

TABLE B.17 – Termes sources à traduire

B.4.0.3 Extrait des termes sources à traduire

	S^{DE}	S^{FR}
soft-tissue:ADJECTIF	✔	✔
pair-wise:ADJECTIF	✔	✔
extra-renal:ADJECTIF	✔	✔
epidermal:ADJECTIF	✔	✔
watered-down:ADJECTIF	✔	✔
mammotome:NOM	✔	✔
overgrowth:NOM	✔	✔
dedifferentiated:ADJECTIF	✔	✔
fairly:ADVERBE		
overreact:VERBE	✔	✔
two-site:ADJECTIF	✔	✔
uk-specific:ADJECTIF	✔	✔
fluorometry:NOM	✔	✔
breast-feeding:PARTICIPE	✔	✔
inconvenient:ADJECTIF	✔	✔
erythroblastic:ADJECTIF	✔	✔
bromocriptine:NOM	✔	✔
medroxyprogesterone:NOM	✔	✔
time-to-progression:NOM	✔	✔
anti-p21:ADJECTIF	✔	✔
cost-effective:ADJECTIF	✔	✔
cancer-free:ADJECTIF	✔	✔
in-frame:ADJECTIF	✔	✔
low-power:ADJECTIF	✔	✔
high-risk:ADJECTIF	✔	✔
flu-like:ADJECTIF	✔	✔

Extrait B.24 – Termes sources à traduire

B.5 Données de référence pour l'évaluation de la génération de traduction

B.5.1 Référence *a priori*

La référence *a priori* (R) a été constituée à partir de l'UMLS (Lindberg *et al.*, 1993).

	FR	DE
$\|R\|$	126 → 163	90 → 104

TABLE B.18 – Nb. entrées et de traductions dans la référence *a posteriori*

B.5.1.1 Extrait anglais → allemand

EN	DE
discomfort:NOM	unwohlsein, fuehlt sich nicht wohl
meta-analyses:ADJECTIF	metaanalyse
rapamycin:ADJECTIF	sirolimus
progestogen:ADJECTIF	gestagene, progestine
quadrantectomy:ADJECTIF	quadrantektomie
overweight:ADJECTIF	übergewicht
neo-adjuvant:ADJECTIF	neoadjuvante behandlung, neoadjuvante therapie
adriamycin:NOM	adriamycin
antibody:NOM	antikörper
immunotherapy:NOM	immuntherapie
childbirth:NOM	geburt
x-ray:VERBE	röntgenuntersuchung, röntgenaufnahme, röntgenstrahlen, radiografie, roentgen
ultrasonography:ADJECTIF	ultraschall, sonographie
dosimetry:NOM	dosimetrie
glycoprotein:NOM	glykoproteine
intra-abdominal:ADJECTIF	bauchhöhle
brachytherapy:NOM	brachytherapie
workflow:ADJECTIF	workflow, arbeitsfluss, arbeitsablauf
radiosurgery:NOM	radiochirurgie
childhood:NOM	kindheit
hysterectomy:NOM	hysterektomie
polymorphism:NOM	vielgestaltigkeit
androstenedione:NOM	androstendion
heterozygous:ADJECTIF	heterozygot
breastfeed:VERBE	stillen

Extrait B.25 – Référence *a priori* anglais → allemand

B.5.1.2 Extrait anglais → français

EN	FR
oophorectomy:ADJECTIF	ablation d'un ovaire, ovariectomie
radiograph:NOM	radiographie
cytoskeleton:NOM	cytosquelette
self-esteem:NOM	estime de soi
ownership:NOM	propriété, possession
cytoplasm:NOM	cytoplasme
chromatography:NOM	chromatographie
overweight:ADJECTIF	surcharge pondérale, excès de poids
subpopulation:NOM	groupes de population, groupes
hydrocortisone:NOM	hydrocortisone
downregulation:NOM	régulation négative
aftercare:NOM	suivi
re-operation:NOM	réintervention, reprise chirurgicale
housework:NOM	ménage
ataxia-telangiectasia:NOM	ataxietélangiectasie
inpatient:NOM	patients hospitalisés
osteoarthritis:NOM	arthrose
database:NOM	base de données
cardiotoxicity:NOM	toxicité cardiaque
workload:NOM	charge de travail
indomethacin:NOM	indométacine
decision-making:ADJECTIF	prise de décision
biomarker:NOM	biomarqueur
anti-androgen:NOM	antiandrogène
physiotherapist:NOM	kinésithérapeute

Extrait B.26 – Référence *a priori* anglais → français

B.5.2 Référence *a posteriori*

La référence *a posteriori* (P) a été constituée en annotant manuellement les sorties du système.

	FR	DE
$\|P\|$	730 → 2 129	654 → 2 016

TABLE B.19 – Nb. entrées et de traductions dans la référence *a posteriori*

	EN-FR	EN-DE
$Kappa$	0,71	0,77

TABLE B.20 – Accord inter-annotateur sur l'annotation de la référence *a posteriori*

Légende des extraits

E traduction annotée EXACT

A traduction annotée ACCEPTABLE

P traduction annotée PROCHE

F traduction annotée FAUX

B.5.2.1 Extrait anglais → allemand

EN	An.	DE (lemmes + parties du discours)	
hands-on:ADJECTIF	F	auf der hand	_PREP _DET NOM
discontinuation:NOM	P	nicht fortsetzen	_ADVERBE _ELIM
increasingly:ADVERBE	F	sich nicht weiter vermehren	_PRO _ADVERBE _ADVERBE VERBE
greatly:ADVERBE	F	grat	NOM
one-breasted:ADJECTIF	F	einer oder sogar beide brust	_ELIM _CONJ _ADVERBE _DET NOM
non-selective:ADJECTIF	E	nichtselektiv	ADJECTIF
high-risk:ADJECTIF	E	hochrisiko	LEX
well-fitted:ADJECTIF	P	gut zu sie passen	ADJECTIF _PREP _PRO VERBE
workflow:ADJECTIF	E	workflow	NOM
increasingly:ADVERBE	A	zunehmen	ADJECTIF
ultimately:ADVERBE	E	schliesslich	_ADVERBE
metaphase:ADJECTIF	F	metaplasie	_ADVERBE
pretreatment:NOM	F	treatments	ADJECTIF
three-year:ADJECTIF	F	jahr drei	NOM _NUM
self-examination:NOM	P	untersuchung selbst	NOM _ADVERBE
adjuvant-therapy:NOM	E	adjuvante therapie	NOM NOM
inter-patient:ADJECTIF	E	zwischen der patient	_PREP _DET LEX
well-differentiated:ADJECTIF	F	differenzierung der echt	NOM _DET ADJECTIF
co-operative:ADJECTIF	F	mit ein operation	_PREP _DET NOM
three-year:ADJECTIF	F	jahr nur drei	NOM _ADVERBE _NUM
helpless:ADJECTIF	F	nicht jede so hilfsbereit	_ADVERBE _PRO _ADVERBE ADJECTIF
privately:ADVERBE	F	eigen	ADJECTIF
retinoblastoma:NOM	E	retinoblastom	NOM
in-frame:ADJECTIF	F	infrage	_ADVERBE
increasingly:ADVERBE	P	mit ein erhöhen	_PREP _DET ADJECTIF

Extrait B.27 – Référence *a posteriori* anglais → allemand

B.5.2.2 Extrait anglais → français

EN	An.	FR (lemmes + parties du discours)	
fabric-covered:ADJECTIF	F	tissu le recouvrir	NOM _PRO VERBE
pre-menopausal:ADJECTIF	E	avant le ménopause	_PREP _DET NOM
primarily:ADVERBE	E	fondamentalement	ADVERBE
co-operative:ADJECTIF	A	coopérateur	ADJECTIF
angiogenesis:NOM	A	angiogenèser	VERBE
disappearance:NOM	F	disparité	NOM
interphase:ADJECTIF	F	interposer	PARTICIPE
breast-cancer:NOM	P	cancer in_situ de sein	NOM _ADVERBE _PREP NOM
post-operative:ADJECTIF	E	après ce intervention	_PREP _DET NOM
disorderly:ADJECTIF	F	avec le affection	_PREP _DET NOM
double-blind:ADJECTIF	E	double aveugle	NOM ADJECTIF
cross-talk:ADJECTIF	E	crosstalk	NOM
adversely:ADVERBE	F	opposé	ADJECTIF
one-piece:ADJECTIF	F	un morceau	_DET NOM
co-repressors:ADJECTIF	A	corépression	NOM
disorderly:ADJECTIF	E	manière désordonné	NOM ADJECTIF
colourless:ADJECTIF	E	sans colorant	_PREP NOM
part-time:ADJECTIF	F	taille et le temps	NOM _CONJ _DET NOM
last-timed:ADJECTIF	F	passetemps	NOM
comparatively:ADVERBE	F	relative	NOM
cancer-trained:ADJECTIF	F	cancer et de son traitement	NOM _CONJ _PREP _DET NOM
oophorectomy:ADJECTIF	E	ovariectomie	NOM
on-treatment:NOM	E	sur un traitement	_PREP _DET NOM
consecutively:ADVERBE	A	consécutif	ADJECTIF
lymphocyte:NOM	E	lymphocyte	NOM

Extrait B.28 – Référence *a posteriori* anglais → français

B.6 Données pour l'apprentissage et l'évaluation du modèle d'ordonnancement

Les données d'apprentissage (T) correspondent aux sorties du systèmes annotées qui ne sont pas dans la référence *a priori*. Les données d'évaluation (E) correspondent aux sorties du systèmes annotées qui font partie de la référence *a priori*.

	FR	DE
$\|T\|$	647 → 1 970	588 → 1 829
$\|E\|$	83 → 159	66 → 187

TABLE B.21 – Nb. d'entrées et de traductions pour l'apprentissage et l'évaluation du modèle d'ordonnancement

B.6.1 Extraits des données d'apprentissage

Légende des extraits

$a > b$ la traduction a est plus pertinente que la traduction b

a, b la traduction a est aussi pertinente que la traduction b

F : traduction(s) annotée(s) FAUX

A : traduction(s) annotée(s) ACCEPTABLE

P : traduction(s) annotée(s) PROCHE

F : traduction(s) annotée(s) FAUX

B.6.1.1 Extrait anglais → allemand

EN	Traductions DE ordonnées (lemmes)
metastasis-free|ADJECTIF	E : metastasefrei
vasomotor|ADJECTIF	E : vasomotorisch
stick-on|ADJECTIF	A : ankleben (VERBE), aufkleben, ankleben (NOM)
prickly|ADJECTIF	F : stich, trick
non-specific|ADJECTIF	E : unspezifisch > P : nicht spezifisch > F : nicht oder nur wenig spezifisch
first-line|ADJECTIF	E : first line, firstline
factor-a|NOM	F : factor (_ADVERBE), factor (ADJECTIF)
high-risk|ADJECTIF	E : gross risiko, hoch risiko, high risk, hochrisiko > F : hoch das risiko, hoch oder ein risiko, risiko an das hoch, risiko hoch, hoch thrombose risiko, risiko in etwa so hoch, hoch ihr risiko, risiko gross
tumour-bearing|PARTICIPE	E : tumortragen
workstation|NOM	E : workstation

Extrait B.29 – Données pour l'apprentissage du modèle d'ordonnancement anglais → allemand

B.6.1.2 Extrait anglais → français

EN	Traductions FR ordonnées (lemmes)
dinucleotide|NOM	P : nucléotide, nucléotider, nucléotidique
non-african|ADJECTIF	E : nonafricaines
stick-on|ADJECTIF	F : à coller, surmonter, à déterminer
socio-economic|ADJECTIF	E : socioéconomique, social ou économique, économique et social, social et économique
whole-genome|ADJECTIF	E : ensemble de génome
incomplete|ADJECTIF	E : incomplet > A : pas complètement, incomplètement > P : pas totalement > F : pas moins tout_à_fait, non le totalité, non totalement, non un fin
non-randomised|ADJECTIF	E : non randomiser (_ADVERBE ADJECTIF), non randomiser (_ADVERBE PARTICIPE)
prickly|ADJECTIF	P : piquer, piqûre, piquer
non-specific|ADJECTIF	E : non spécifique > F : pas déterminer, pas encore déterminer, pas encore bien déterminer, pas toujours de déterminer, pas de déterminer
cardioprotective|ADJECTIF	E : protection de coeur, protection cardiaque, cardioprotecteur

Extrait B.30 – Données pour l'apprentissage du modèle d'ordonnancement anglais → français

B.6.2 Extrait des données d'évaluation

$a > b$ la traduction a est plus pertinente que la traduction b

a, b la traduction a est aussi pertinente que la traduction b

F : traduction(s) annotée(s) FAUX

A : traduction(s) annotée(s) ACCEPTABLE

P : traduction(s) annotée(s) PROCHE

F : traduction(s) annotée(s) FAUX

B.6.2.1 Extrait anglais → allemand

EN	Traductions DE ordonnées (lemmes)
ultrasound|NOM	F : über die fest
headache|NOM	E : kopfschmerzen
gynaecomastia|NOM	E : gynäkomastie > F : frau durch brust, frau zu die brust, brust der frau, frau nach brust, frau brust, frau nicht an brust, frau bereits an ein brust, frau der brust, frau in der brust, frau an brust, frau gegen brust, brust durch der frau, frau nach ein brust, frau an brust, frau bereits an brust, frau weniger an brust, frau nach ein brust, frau ein brust, frau ohne brust, frau brust, frau mit ein brust, frau mit brust, frau ihr brust, frau mit brust, frau trotz brust, frau die brust, brust bei frau
radiograph|NOM	A : radiographisch, radiography, radiographie
wellbeing|ADJECTIF	F : wellen (NOM), wellen (VERBE)
fibroblast|NOM	E : fibroblasten, fibroblast
mammogram|NOM	E : mammogramm, mammogramme
dosimetry|NOM	E : dosimetrie
overweight|ADJECTIF	A : übergewicht > F : gewicht über
antibody|NOM	E : antikörper (NOM), antikörper (ADJECTIF), antikörper (_ADVERBE) > P : antikörper (LEX)

Extrait B.31 – Données pour l'évaluation du modèle d'ordonnancement anglais → allemand

B.6.2.2 Extrait anglais → français

EN	Traductions FR ordonnées (lemmes)
radiograph|NOM	E : radiographie > A : radiographier, radiographier, radiographique
overweight|ADJECTIF	E : surcharge > F : supérieur avec et sans charge
intramuscular|ADJECTIF	E : dans le muscle, intramusculaire, dans un muscle > A : intramusculaire > F : entre le muscle
orchidectomy|NOM	E : orchidectomie
cytoplasmic|ADJECTIF	E : cytoplasme
g-protein|NOM	E : protéine g
brachytherapy|NOM	E : brachythérapie
fibroblast|NOM	E : fibroblaste
radioactivity|NOM	E : radioactiviter
cardiovascular|ADJECTIF	E : cardiovasculaire, cardiovasculaire > A : cardiovasculairer

Extrait B.32 – Données pour l'évaluation du modèle d'ordonnancement anglais → français

B.6.3 Extrait des sorties du système ordonnées

Les traductions présentées dans l'extrait ont été ordonnées avec la combinaison non pondérée.

Légende des extraits

$a > b$ la traduction a a été placée avant la traduction b par l'algorithme.

B.6.3.1 Extrait anglais → allemand

EN	Traductions DE ordonnées (lemmes)
ultrasound|NOM	über die fest
headache|NOM	kopfschmerzen
gynaecomastia|NOM	gynäkomastie > frau mit brust > frau an brust > frau nicht an brust > frau nach brust > frau die brust > frau brust > frau in der brust > frau ihr brust > frau mit brust > frau weniger an brust > frau ohne brust > frau nach ein brust > frau bereits an ein brust > frau mit ein brust > frau trotz brust > frau zu die brust > frau der brust > frau gegen brust > frau an brust > frau brust > brust durch der frau > brust bei frau > frau ein brust > frau nach ein brust > frau durch brust > frau bereits an brust > brust der frau
radiograph|NOM	radiography > radiographie > radiographisch
wellbeing|ADJECTIF	wellen (NOM) > wellen (VERBE)
fibroblast|NOM	fibroblast > fibroblasten
mammogram|NOM	mammogramm > mammogramme
dosimetry|NOM	dosimetrie
overweight|ADJECTIF	übergewicht > gewicht über
antibody|NOM	antikörper (NOM) > antikörper (LEX) > antikörper (_ADVERBE) > antikörper (ADJECTIF)

Extrait B.33 – Sorties ordonnées par le système (anglais → allemand)

B.6.3.2 Extrait anglais → français

EN	Traductions FR ordonnées (lemmes)
radiograph|NOM	radiographie > radiographique> radiographier (PARTICIPE) > radiographier (VERBE)
overweight|ADJECTIF	surcharge > supérieur avec et sans charge
intramuscular|ADJECTIF	intramusculaire (ADJECTIF) > intramusculaire (NOM) > dans le muscle > dans un muscle > entre le muscle
orchidectomy|NOM	orchidectomie
cytoplasmic|ADJECTIF	cytoplasme
g-protein|NOM	protéine g
brachytherapy|NOM	brachythérapie
fibroblast|NOM	fibroblaste
radioactivity|NOM	radioactiviter
cardiovascular|ADJECTIF	cardiovasculaire (ADJECTIF) > cardiovasculaire (NOM) > cardiovasculairer

Extrait B.34 – Sorties ordonnées par le système (anglais → français)

Annexe C

Interface de consultation des lexiques extraits de corpus comparables

Le prototype est consultable librement en ligne à l'adresse : http://80.82.238.151/Metricc/InterfaceValidation/ [1].

Il permet de consulter les lexiques extraits du corpus CANCER DU SEIN anglais-français et du corpus SCIENCES DE L'EAU ainsi que les fiches terminologiques associées aux termes.

1. Le nom d'utilisateur est "test". Laisser le champ mot de passe vide.

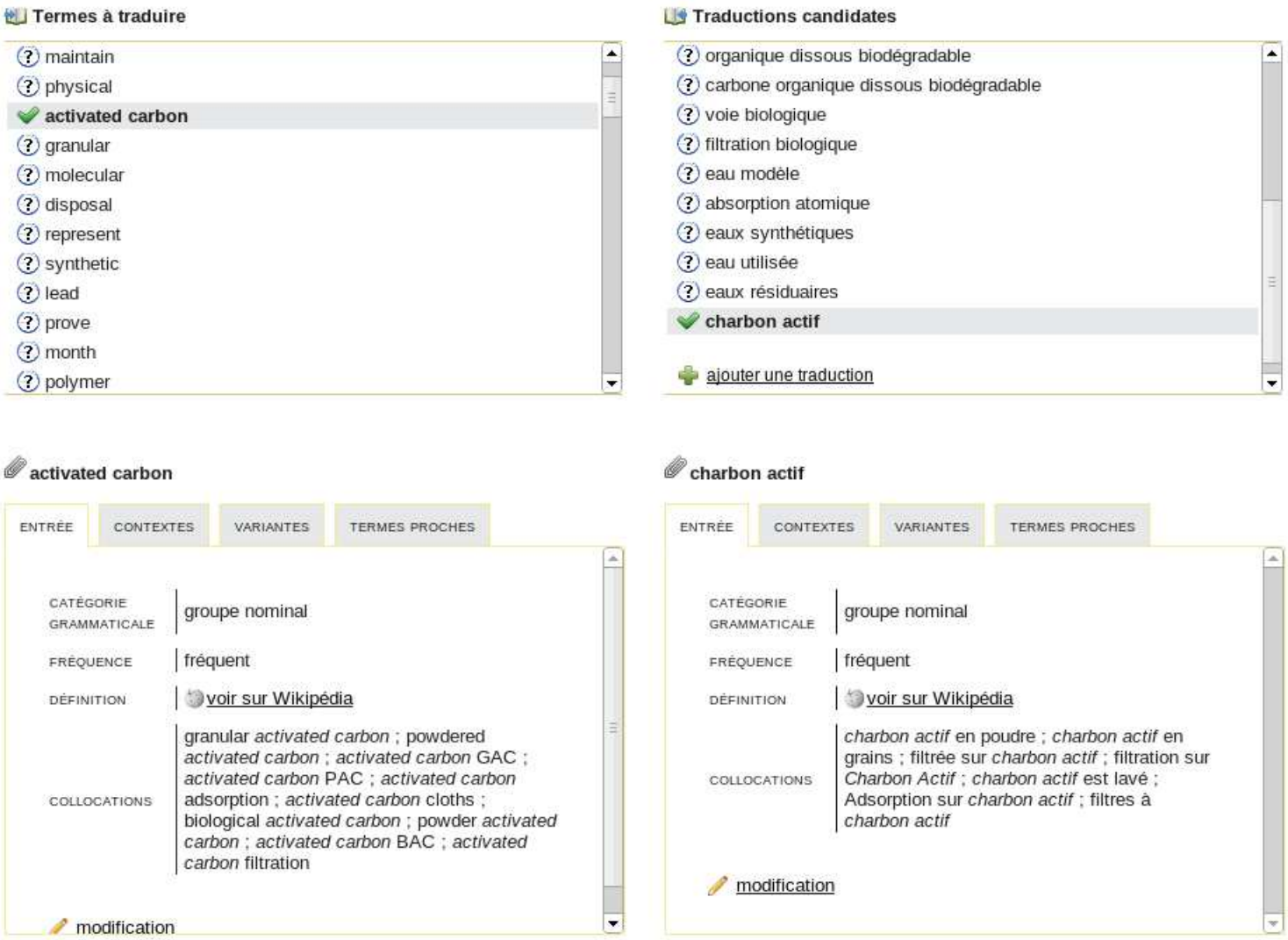


FIGURE C.1 – Prototype d'explorateur de corpus comparables : équivalences traductionnelles et fiches terminologiques

FIGURE C.2 – Prototype d'explorateur de corpus comparables : contextes bilingues

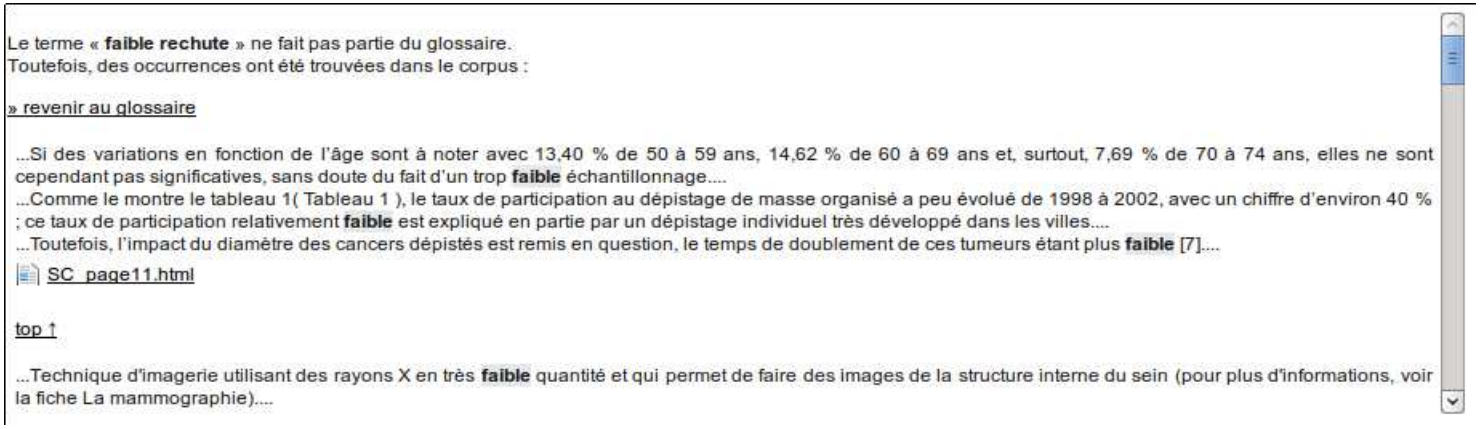


FIGURE C.3 – Prototype d'explorateur de corpus comparables : contextes monolingues

Références

Liste des tableaux

Table des figures

Liste des algorithmes

Liste des extraits

Bibliographie

ALPAC : Languages and machines : computers in translation and linguistics. Publication 1416, Automatic Language Processing Advisory Committee - Division of Behavorial Sciences, National Academy of Sciences, National Research Council, Washington D. C., États-Unis d'Amérique, 1966.

M. BAKER : Corpus-based translation studies : The challenges that lie ahead. *In* H. SOMERS, éditeur : *Terminology, LSP and Translation : Studies in Language Engineering in Honour of Juan C. Sager*. John Benjamins, Amsterdam, Pays-Bas et Philadelphia, États-Unis d'Amérique, 1996.

T. BALDWIN et T. TANAKA : Translation by machine of complex nominals. *In Proceedings of the ACL 2004 Workshop on Multiword expressions : Integrating Processing*, pages 24–31, Barcelona, Spain, 2004.

S. BANERJEE et A. LAVIE : METEOR : an automatic metric for MT evaluation with improved correlation with human judgments. *In Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics*, pages 65–72, Ann Arbor, Michigan, 2005.

L. BAUER : *English word-formation*. Cambridge University Press, Cambridge, United Kingdom, 1983.

H. D. BÉCHADE : *Phonétique et morphologie du français moderne et contemporain*. Presses Universitaires de France, 1992.

P. BENNISON et L. BOWKER : Designing a tool for exploiting bilingual comparable corpora. *In Proceedings of LREC 2000*, Athens, Greece, 2000.

H. BLANCHON et C. BOITET : Pour l'évaluation externe des systèmes de TA par des méthodes fondées sur la tâche. *Traitement Automatique des Langues*, 48(1):33–65, 2007.

D. BOURIGAULT : *LEXTER un Logiciel d'EXtraction de TERminologie. Application à l'extraction des connaissances à partir de textes*. Thése en mathématiques, informatique appliquée aux sciences de l'Homme, École des HautesÉtudes en Sciences Sociales, Paris, 1994.

L. BOWKER et J. PEARSON : *Working with Specialized Language : A Practical Guide to Using Corpora*. Routledge, London/New York, 2002.

P. BROWN, S. DELLA PIETRA, V. DELLA PIETRA et R. MERCER : The mathematics of statistical machine translation : parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

P. BROWN, S. Della PIETRA, F. JELINEK, J. LAFFERTY, R. MERCER et P. ROOSSIN : A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, 1990.

F. BROWN DE COLSTOUN, E. DELPECH et E. MONNERET : Libellex : une plateforme multiservices pour la gestion des contenus multilingues. *In* M. LAFOURCADE et V. PRINCE, éditeurs : *Actes de la 18ème conférences sur le traitement automatique des langues naturelles*, volume 2, page 319, Montpellier, France, 2011.

C. CALLISON-BURCH, F. CAMEROB, P. KOEHN, C. MONZ et J. SCHROEDER : Further Meta-Evaluation of machine translation. *In Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio, 2008.

C. CALLISON-BURCH, C. FORDYCE, P. KOEHN, C. MONZ et J. SCHROEDER : (Meta-) evaluation of machine translation. *In Proceedings of the 2nd workshop on Statistical Machine Translation*, page 136–158, Prague, Czech Republic, 2007.

C. CALLISON-BURCH, P. KOEHN, C. MONZ, K. PETERSON, M. PRZYBOCKI et O. ZAIDAN : Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. *In Proceedings of the Sixth Workshop on Statistical Machine Translation*, Uppsala, Sweden, 2010.

C. CALLISON-BURCH, P. KOEHN, C. MONZ et J. SCHROEDER : Findings of the 2009 workshop on statistical machine translation. *In Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece, mars 2009. Association for Computational Linguistics.

Z. CAO, T. QIN, T.-Y. LIU, M.-F. TSAI et H. LI : Learning to rank : From pairwise approach to listwise approach. Rapport technique MSR-TR-2007-40, Microsoft Research, 2007.

M. CARL, B. DRAGSTED et A. L. JAKOBSEN : On the systematicity of human translation process. *In Tralogy*. Kluwer Academic Publisher, Paris, France, 2011.

J. CARLETTA : Assessing agreement on classification tasks : The kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996.

M. CARPUAT, H. Daumé III, A. FRASER, C. QUIRK, F. BRAUNE, A. CLIFTON, A. IRVINE, J. JAGARLAMUDI, J. MORGAN, M. RAZMARA, A. TAMCHYNA, K. HENRY et R. RUDINGER : Domain adaptation in machine translation : Final report. *In 2012 Johns Hopkins Summer Workshop Final Report*. 2012. URL `http://hal3.name/damt/`. dernière consultation le 01/02/2013.

B. CARTONI : Traduction de règles de construction des mots pour résoudre les problèmes d'incomplétude lexicale en traduction automatique étude de cas. *In Proceedings of RECITAL 2005*, pages 565–574, Dourdan, France, 2005.

B. CARTONI : Les adjectifs relationnels dans les lexiques informatisés : formalisation et exploitation dans un contexte multilingue. *In Actes de la 16e Conférence Traitement Automatique des Langues Naturelles*, Senlis, France, 2009a.

B. CARTONI : Lexical morphology in machine translation : A feasibility study. *In Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 130–138, Athènes, Grèce, 2009b.

C.-C. CHANG et C.-J. LIN : LIBSVM : A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27 :1–27 :27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Y. CHIAO : *Extraction lexicale bilingue à partir de textes médicaux comparables : application à la recherche d'information translangue*. Thèse de doctorat en informatique médicale, Université de Paris 6, Paris, France, 2004.

Y. CHIAO et P. ZWEIGENBAUM : Looking for French-English translations in comparable medical corpora. *Journal of the American Society for Information Science*, 8:150–154, 2002.

V. CLAVEAU : Translation of biomedical terms by inferring rewriting rules. *In* V. PRINCE et M. ROCHE, éditeurs : *Information Retrieval in Biomedicine : Natural Language Processing for Knowledge Integration*, pages 106–123. Medical Information Science Reference, 2009.

V. CLAVEAU et E. KIJAK : Morphological analysis of biomedical terminology with Analogy-Based alignment. *In Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 347–354, Hissar, Bulgaria, 2011.

D. CORBIN : *Morphologie dérivationnelle et structuration du lexique*. Presses Universitaires de Lille, Lille, 1987.

D. COSTAOUEC et F. GUÉRIN : *Syntaxe fonctionnelle. Théorie et exercices*. Presses Universitaires de Rennes, Rennes, 2007.

B. DAILLE : *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. Thèse de doctorat en informatique fondamentale, Université Paris 7, Paris, France, 1994.

B. DAILLE, E. GAUSSIER et J.-M. LANGÉ : Towards automatic extraction of monolingual and bilingual terminology. *In Proceedings of the 15th International Conference on Computational Linguistics*, pages 515–521, Kyoto, Japan, 1994.

B. DAILLE et E. MORIN : French-English terminology extraction from comparable corpora. *In Proceedings, 2nd International Joint Conference on Natural Language Processing*, volume 3651 de *Lecture Notes in Computer Sciences*, pages 707–718, Jeju Island, Korea, 2005. Springer.

J. DARBELNET : Réflexions sur le discours juridique. *Meta : journal des traducteurs / Meta : Translator's Journal*, 24(1):26–34, 1979.

C. DE GROC : Babouk : Focused web crawling for corpus compilation and automatic terminology extraction. *In Proceedings of the IEEE/WIC/ACM International Conferences on Web Intelligence*, pages 497–498, Lyon, France, 2011.

E. DÉJEAN et E. GAUSSIER : Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica, Alignement lexical dans les corpus multilingues*, pages 1–22, 2002.

L. DÉLÉGER : *Exploitation de corpus parallèles et comparables pour la détection de correspondances lexicales : application au domaine médical*. Thèse de doctorat en informatique médicale, Université Pierre et Marie Curie, Paris, 2009.

E. DELPECH et B. DAILLE : Dealing with lexicon acquired from comparable corpora : validation and exchange. *In Proceedings of the 2010 Terminology and Knowledge Engineering Conference (TKE 2010)*, pages 211–223, Dublin, Ireland, 2010.

A. P. DEMPSTER : Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

G. DODDINGTON : Automatic evaluation of machine translation quality using n-gram Co-Occurrence statistics. *In Proceedings of the second international conference on Human Language Technology Research*, pages 128–145, San Diego, California, 2002.

G. DROSDOWSKI : *Das grosse Wörterbuch der deutschen Sprachen in 8 Bänden*. Duden, 2006.

T. DUNNING : Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.

C. DURIEUX : *Fondement didactique de la traduction technique*. La maison du dictionnaire, Paris, France, 2010.

C. ENGUERARD et L. PANTERA : Automatic natural acquisition of a terminology. *Journal of Quantitative Linguistics*, 2(1):27–32, 1995.

J. FIDRMUC : The economics of multilingualism in the EU. CEDI Discussion Paper Series n° 11-04, Centre for Economic Development and Institutions (CEDI), Brunel University, London, UK, 2011.

Y. FREUND et R. E. SCHAPIRE : Large margin classification using the perceptron algorithm. *Machine Learning*, 37(3):277–296, 1999.

I. FRIEDBICHLER et M. FRIEDBICHLER : The potential of domain-specific target-language corpora for the translator's workbench. *In First international conference on Corpus Use and Learning to Translate*, Bertinoro, Italie, 1997.

P. FUNG : Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. *In* D. YAROVSKY et K. CHURCH, éditeurs : *Proceedings of the 3rd Workshop on Very Large Corpora*, pages 173–183, 1995.

P. FUNG : Finding terminology translations from non-parallel corpora. *In Proceedings of the 5th Annual Workshop on Very Large Corpora*, pages 192–202, Hong Kong, 1997.

P. FUNG : A statistical view on bilingual lexicon extraction : From parallel to Non-Parallel corpora. *In Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas*, pages 1–17, Langhorne, PA, USA, 1998.

P. FUNG et P. CHEUNG : Mining Very-Non-Parallel corpora : Parallel sentence and lexicon extraction via bootstrapping and EM. *In Proceedings of EMNLP 2004*, pages 57–63, Barcelona, Spain, 2004.

N. GARERA et D. YAROWSKY : Translating compounds by learning component gloss translation via multiple languages. *In Proceedings of the 3rd International Joint Conference on Natural Language Processing*, volume 1, pages 403–410, Hyderabad, Inde, 2008.

E. GAUSSIER, D. A. HULL et S. AÏT-MOKHTAR : Term alignment in use : MachineAided human translation. *In* J. VÉRONIS, éditeur : *Parallel Text Processing*, pages 253–274. Kluwer Academic Publisher, London, 2000.

L. GAVIOLI et F. ZANETTIN : Comparable corpora and translation : a pedagogic perspective. *In Corpus use and learning to translate*. Bertinoro, Italie, 1997.

F. GENDRE : *L'analyse statistique univariée : Introduction à son utilisation pratique*. Librairie Droz, France, 1977.

L. GOEURIOT : *Découverte et caractérisation des corpus comparables spécialisés*. Thèse en informatique, Université de Nantes, Nantes, 2009.

N. GRABAR : *Terminologie médicale et morphologie. Acquisition de ressources morphologiques et leur utilisation pour le traitement de la variation terminologique*. Thèse de doctorat en informatique médicale, Université de Paris 6, Paris, France, 2004.

G. GREFENSTETTE : The world wide web as a resource for example-based machine translation tasks. *ASLIB'99 Translating and the computer*, 21, 1999.

S. HAGEN, J. FOREMAN-PECK, S. DAVILA-PHILIPON et B. NORDGREN : ELAN : Effects on the european economy of shortages of foreign languages skills in enterprise. Rapport technique, CILT, the National Centre for Languages, England, 2006.

M. HALL, E. FRANK, G. HOLMES, B. PFAHRINGER, P. REUTEMANN et I. H. WITTEN : *The WEKA Data Mining Software : An Update*, volume 11. 2009.

R. HARASTANI, B. DAILLE et E. MORIN : Neoclassical compound alignments from comparable corpora. *In Proceedings of the 13th International Conference on Intelligent Text Processing and Computational Linguistics*, pages 72–82, 2012.

B. HAUER et G. KONDRAK : Clustering semantically equivalent words into cognate sets in multilingual lists. *In Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 865–873, Chiang Mai, Thailand, 2011.

A. HAZEM et E. MORIN : ICA for bilingual lexicon extraction from comparable corpora. *In Proceedings of the 5th Workshop on Building and Using Comparable Corpora*, Istanbul, Turkey, 2012.

J. HUTCHINS : ALPAC : the (in)famous report. *MT News International*, (14):9–12, 1996.

J. HUTCHINS : Machine translation : general overview. *In* Mitkov R., éditeur : *The Oxford Handbook of Computational Linguistics*, pages 501–511. Oxford University Press, New York, USA, 2005.

C. IACOBINI : Composizione con elementi neoclassici. *La formazione delle parole in italiano*, pages 69–96, 2003.

ISO : Terminology work – principles and methods. Rapport technique 704, International Organization for Standardization, 2009.

C. JACQUEMIN : A symbolic and surgical acquisition of terms through variation. *In* Scheler G. WERMTER S., Riloff E., éditeur : *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, pages 425–438. Springer, Heidelberg, Germany, 1996.

H. JI : Mining name translations from comparable corpora by creating bilingual information networks. *In Proceedings of the 2nd Workshop on Building and Using Comparable Corpora : from Parallel to Non-parallel Corpora*, pages 34–37, Suntec, Singapore, 2009.

E. L. KEENAN et L. M. FALTZ : *Boolean semantics for natural language*. Dordrecht, Holland, 1985.

M. KENDALL : A new measure of rank correlation. *Biometrika*, 30(1-2):81–89, 1983.

A. KILGARRIFF, P.V.S. AVINESH et J. POMIKÀLEK : BootCatting comparable corpora. *In Proceedings of the International Conference on Terminology and Artificial Intelligence*, pages 123–126, Paris, France, 2011.

P. KOEHN et K. KNIGHT : Learning a translation lexicon from monolingual corpora. *In Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX)*, pages 9–16, Philadelphia, Pennsylvania, États-Unis d'Amérique, 2002. Association for Computational Linguistics.

P. KOEHN et C. MONZ : Manual and automatic evaluation of machine translation between european languages. *In Proceedings of the Workshop on Statistical Machine Translation*, pages 102–121, New York City, New York, États-Unis d'Amérique, 2006. Association for Computational Linguistics.

J.R. LANDIS et G.G. KOCH : The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–74, 1977.

P. LANGLAIS, F. YVON et P. ZWEIGENBAUM : Analogical translation of medical words in different languages. *In Proceedings of the 6th international conference on Advances in Natural Language Processing*, pages 284–295, 2008.

P. LANGLAIS, F. YVON et P. ZWEIGENBAUM : Improvements in analogical learning : Application to translating multi-terms of the medical domain. *In 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 487–495, Athens, Greece, 2009.

A. LARDILLEUX : *Contribution des basses fréquences à l'alignement sous-phrastique multilingue : une approche différentielle*. Thèse de doctorat, Université de Caen Basse-normandie, Caen, 2010.

A. LAROCHE et P. LANGLAIS : Revisiting context-based projection methods for term-translation spotting in comparable corpora. *In Proceedings of the 23rd International Conference on Computational Linguistics*, pages 617–625, Beijing, China, 2010.

R. LAROSE : Méthodologie de l'évaluation des traductions. *Méta : journal des traducteurs / Meta : Translators' Journal*, 43(2):163–186, 1998.

S. LÉON : *Acquisition automatique de traductions d'unités lexicales complexes à partir du Web*. Thèse en sciences du langage - traitement automatique des langues, Université de Provence - Aix-Marseille I, Marseille, France, 2008.

Y. LEPAGE : *De l'analogie rendant compte de la commutation en linguistique*. Mémoire d'habilitation à diriger des recherches, Université Joseph Fourier, Grenoble I, Grenoble, France, 2003.

V. I. LEVENSHTEIN : Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics*, 10:707–710, 1966.

B. LI et E. GAUSSIER : Improving corpus comparability for bilingual lexicon extraction from comparable corpora. *In 23ème International Conference on Computational Linguistics*, pages 23–27, Beijing, Chine, 2010.

B. LI, E. GAUSSIER, E. MORIN et A. HAZEM : Degré de comparabilité, extraction lexicale bilingue et recherche d'information interlingue. *In Actes de la 18ème conférences sur le traitement automatique des langues naturelles*, volume 1, pages 211–222, Montpellier, France, 2011.

H LI et J. XU : Adarank : A boosting algorithm for information retrieval. *In Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval*, pages 391–398, Amsterdam, The Netherlands, 2007.

D.A.B. LINDBERG, B.L. HUMPHREYS et A.T. MCCRAY : The unified medical language system. *Methods Inf Med*, 32:81–91, 1993.

T.-Y. LIU : WWW 2009 tutorial on learning to rank for information retrieval. Madrid, Spain, 2009. 18th International World Wide Web Conference.

T.-Y. LIU : *Learning to Rank for Information Retrieval*. Springer Verlag, New York City, États-Unis d'Amérique, 2011.

Tie-Yan LIU : Learning to rank : From pairwise approach to listwise approach. *In Pao-Lu Hsu Statistics Conference : Machine Learning*, 2007.

C. LOVIS, R. BAUD, AM. RASSINOUX, PA. MICHEL et JR. SCHERRER : Medical dictionaries for patient encoding systems : a methodology. *Artificial Intelligence in Medecine*, pages 201–214, 1998.

L. MACKEN, E. LEFEVER et V. HOSTE : Language-independent bilingual terminology extraction from a multilingual parallel corpus. *In Proceedings of the 22nd International Conference on Computational Linguistics*, volume 1, pages 529–536, Machester, Royaume-Uni, 2008.

C. D. MANNING, P. RAGHAVAN et H. SCHÜTZE : *Introduction to Information Retrieval*. Cambridge University Press, HTML édition, 2008. URL `http://nlp.stanford.edu/IR-book/html/htmledition/irbook.html`.

A. MARTINET : *La grammaire fonctionnelle du français*. Didier, Paris, 1979.

A. MARTINET : *Syntaxe générale*. Armand Colin, Paris, 1985.

A. M. MC ENERY et R. Z. XIAO : Parallel and comparable corpora : What is happening ? *In* M. Rogers G. ANDERMAN, éditeur : *Incorporating Corpora : The Linguist and the Translator.*, Translating Europe, pages 18–31. Multilingual Matters, Clevedon, UK, 2007.

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

I. MEL'ČUK : *Aspects of the Theory of Morphology*. Mouton de Gruyter, Berlin . New York, 2006.

D. METZLER et W. B. CROFT : Linear feature-based models for information retrieval. *Information Retrieval*, 10:257–274, 2000.

I. MEYER : Extracting knowledge-rich contexts for terminography : A conceptual and methodological framework. *In* D. BOURIGAULT, C. JACQUEMIN et M.-C. L'HOMME, éditeurs : *Recent Advances in Computational Terminology*, pages 279–302. John Benjamins, 2001.

E. MORIN et B. DAILLE : Compositionality and lexical alignment of multi-word terms. *In* P. RAYSON, S. PIAO, S. SHAROFF, S. EVERT et Villada-Moirón B., éditeurs : *Language Resources and Evaluation (LRE)*, volume 44 de *Multiword expression : hard going or plain sailing*, pages 79–95. Springer Netherlands, 2010.

E. MORIN et B. DAILLE : Revising the compositional method for terminology acquisition from comparable corpora. *In International Conference on Computational Linguistics (COLING)*, pages 1797–1810, Mumbai, Inde, 2012.

E. MORIN, B. DAILLE, K. TAKEUCHI et K. KAGEURA : Bilingual terminology mining - using brain, not brawn comparable corpora. *In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 664–671, Prague, Czech Republic, 2007.

E. MORIN, S. DUFOUR-KOWALSKI et B. DAILLE : Extraction de terminologies bilingues à partir de corpus comparables. *In Actes de la 11ème Conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN)*, pages 309–318, Fès, Maroc, 2004.

M. NAGAO : A framework of a mechanical translation between japanese and english by analogy principle. *In* Banerji R. ELITHORN A., éditeur : *Artificial and Human Intelligence*, pages 173–180. Amsterdam : North-Holland, 1984.

F. NAMER : Les mots composés morphologiquement. Note technique faite dans le cadre du projet umlf landisco, Université Nancy 2, 2003.

F. NAMER : Morphosémantique pour l'appariement de termes dans le vocabulaire médical : approche multilingue. *In Actes de la 12ème sur le Traitement Automatique des Langues*, pages 63–72, Dourdan, France, 2005.

F. NAMER et R. BAUD : Defining and relating biomedical terms : Towards a cross-language morphosemantics-based system. *International Journal of Medical Informatics*, 76(2-3):226–33, 2007.

F. OCH et H. NEY : A comparison of alignment models for statistical machine translation. *In Proceedings of the 18th Conference on Computational Linguistics*, volume 2, pages 1086–1090, 2000.

P. G. OTERO et J. R. CAMPOS : An approach to acquire word translations from Non-Parallel texts. *In Lecture Notes in Computer Science*, volume 3808 de *Progress in Artificial Intelligence*, pages 600–610. Springer-Verlag, 2005.

S. OZDOWSKA : *ALIBI, un système d'ALIgnement BIlingue à base de règles de propapagation syntaxique*. Thèse de doctorat en sciences du langage, Université Toulouse II Le Mirail, Toulouse, France, 2006.

K. PAPINENI, S. ROUKOS, T. WARD et W.-J. ZHU : BLEU : a method for automatic evaluation of machine translation. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, 2002.

M. PEREZ : Interview d'Héloïse Portal, responsable francophone de linguee, 2010. URL `http://fr.locita.com/societe/linguee-un-moteur-de-recherche-de-traductions-contextuelles-1175/`. dernière consultation le 01/02/2013.

E. PLANAS : *TELA, Structures et Algorithmes pour la Traduction fondée sur la Mémoire*. Thèse de doctorat en informatique, Université Joseph Fourier, Grenoble I, Grenoble, France, 1998.

E. PLANAS : Similis : un logiciel d'aide à la traduction au service des professionnels. *Traduire*, (206):41–48, 2005.

E. PLANAS : Metricc : Rapport final sur l'évaluation de l'apport des lexiques bilingues pour la traduction. Délivrable ANR nº 28 lot 4.3, Université de Nantes, Nantes, 2011.

E. PLANAS et O. FURUSE : Multi-level similar segment matching algorithm for translation memories and example-based machine translation. *In Proceedings of the 18th International Conference on Computational Linguistics*, pages 621–627, Saarbrücken, Allemagne, 2000.

M. F. PORTER : An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

E. PROCHASSON : *Alignement multilingue en corpus comparables spécialisés : Caractérisation terminologique multilingue*. Thèse en informatique, Université de Nantes, Nantes, 2010.

E. PROCHASSON et E. MORIN : Points d'ancrage pour l'extraction lexicale bilingue à partir de petits corpus comparables spécialisés. *Traitement Automatique des Langues*, 50(1):238–304, 2009.

R. RAPP : Identifying word translations in Non-Parallel texts. *In Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 320–322, Boston, Massachussets, États-Unis d'Amérique, 1995.

R. RAPP : Automatic Identification of Word Translations from Unrelated English and German Corpora. *In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 519–526, College Park, MD, USA, 1999.

S. RAUF et H. SCHWENK : On the use of comparable corpora to improve SMT performance. *In Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 16–23, Athènes, Grèce, 2009.

K. REISS : *Translation criticism, the potentials and limitations : categories and criteria for translation quality assessment*. St. Jerome Pub., Manchester, GB, 1971.

M. RENDERS, H. DÉJEAN et É. GAUSSIER : Assessing automatically extracted bilingual lexicons for CLIR in vertical domains : XRCE participation in the GIRT track of CLEF 2002. volume 2785 de *Advances in Cross-Language Information Retrieval*, pages 363–371. Springer Berlin Heidelberg, 2003.

M. RIEGEL, J.-C. PELLAT et R. RIOUL : *Grammaire méthodique du français*. Presses Universitaires de France (PUF), Paris, France, 2005.

P. RIEHMANN, H. GRUENDL, M. POTTHAST, M. TRENKMANN, B. STEIN et B. FROEHLICH : WORDGRAPH : Keyword-in-context visualization for NETSPEAK's wildcard search. volume 18, pages 1411–1423. IEEE, 2012.

X. ROBITAILLE, X. SASAKI, M. TONOIKE, S. SATO et S. UTSURO : Compiling French-Japanese terminologies from the web. *In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 225–232, Trento, Italy, 2006.

F. SADAT, M. YOSHIKAWA et S. UEMURA : Learning bilingual translations from comparable corpora to Cross-Language information retrieval : Hybrid statistics-based and linguistics-based approach. volume 11, pages 57–64, Sappro, Japan, 2003.

S. SCHULZ, K. MARKÓ, P. DAUMKE, U. HAHN, S. HANSER, P. NOHAMA, R. ANDRADE, E. PACHECO et M. ROMACKER : Semantic atomicity and multilinguality in the medical domain : Design considerations for the MorphoSaurus subword lexicon. *In Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 1684–1687, Genoa, Italy, 2006.

G. SCURTU : Traduire le vocabulaire juridique français en roumain. *Meta : journal des traducteurs / Meta : Translator's Journal*, 53(4):884–898, 2008.

A. SECARĂ : Translation evaluation - a state of the art survey. *In eCoLoRe / MeLLANGE Workshop*, pages 39–44, Leeds, UK, 2005.

SFT : Commission statistiques et étude du marché. SFT enquête tarifs 2009, Syndicat national des traducteur professionnels, 2009.

S. SHAROFF, B. BABYCH, P. RAYSON, P. MUDRAYA et S. PIAO : ASSIST : automated semantic assistance for translators. *In Proceedings to the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 139–142, Trento, Italie, 2006.

M. SNOVER, B. DORR, R. SCHWARTZ, L. MICCIULLA et J. MAKHOUL : A study of translation edit rate with targeted human annotation. *In Proceedings of Association for Machine Translation in the Americas (AMTA 2006)*, pages 224–231, 2006.

A. SOKOLOV, G. WISNIEWSKI et F. YVON : Non-linear n-best list reranking with few features. *In AMTA*, San Diego, Californie, États-Unis d'Amérique, 2012. 10 pages.

H. SOMERS : Machine translation : latest developments. *In* R. MITKOV, éditeur : *The Oxford Handbook of Computational Linguistics*, pages 512–528. Oxford University Press, New York, USA, 2005.

T. TALVENSAARI, A. PIRKOLA, K JÄRVELIN, M. JUHOLA et J. LAURIKAALA : Focused web crawling in the acquisition of comparable corpora. *Information Retrieval*, 11(5):427–445, 2008.

J. TIEDEMANN : News from opus - a collection of multilingual parallel corpora with tools and interfaces. *In* N. NICOLOV, K. BONTCHEVA, G. ANGELOVA et R. MITKOV, éditeurs : *Recent Advances in Natural Language Processing (RANLP 2009)*, volume V, pages 237–248, Amsterdam/Philadelphia, 2009. John Benjamins.

J. TURIAN, L. SHEN et I. D. MELAMED : Evaluation of machine translation and its evaluation. *In Proceedings of MT Summit IX*, page 386—393, New Orleans, USA, 2003.

J. VÉRONIS : From the rosetta stone to the information society. a survey of parallel text processing. *In* J. VÉRONIS, éditeur : *Parallel Text Processing*, pages 1–24. Kluwer Academic Publisher, Londres, Royaume-Uni, 2000.

A. VITERBI : Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, 1967.

Š. VINTAR : Bilingual term recognition revisited the bag-of-equivalents term alignment approach and its evaluation. *Terminology*, 16(2):141–158, 2010.

M. WELLER, A. GOJUN, U. HEID, B. DAILLE et R. HARASTANI : Simple methods for dealing with term variation and term alignment. *In Proceedings of the 9th International Conference on Terminology and Artificial Intelligence*, pages 87–93, Paris, France, 2011.

M. WILLIAMS : The application of argumentation theory to translation quality assessment. *Meta : journal des traducteurs / Meta : Translator's Journal*, 46(2):326–344, 2001.

M. WILLIAMS : *Translation quality assessment : an argumentation-centred approach*. University of Ottawa Press, 2004.

Q. WU, J. C. BURGES, K. SVORE et J. GAO : Adapting boosting for information retrieval measures. *Journal of Information Retrieval*, 13:254–270, 2010.

K. YU et J. TSUJII : Extracting bilingual dictionary from comparable corpora with dependency heterogeneity. *In NAACL-Short '09 Proceedings of Human Language Technologies*, volume Short Papers, pages 121–124, Boulder, Colorado, USA, 2009.

F. ZANETTIN : Bilingual comparable corpora and the training of translators. *Meta : journal des traducteurs / Meta : Translator's Journal*, 43(4):616–630, 1998.