Université de Nantes

ÉCOLE DOCTORALE STIM

« SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DES MATÉRIAUX »

Année 2010

Alignement multilingue en corpus comparables spécialisés

Caractérisation terminologique multilingue

THÈSE DE DOCTORAT

Discipline : Informatique

Spécialité : Traitement Automatique du Langage Naturel

Présentée
et soutenue publiquement par

Emmanuel PROCHASSON

*Le 17 décembre 2009 à l'*UFR *Sciences & Techniques, Université de Nantes,*
devant le jury ci-dessous

Président	:	Pr. Kamel SMAÏLI	Université Nancy 2
Rapporteurs	:	Éric GAUSSIER, Professeur	Université Joseph Fourier
		Yves LEPAGE, Professeur	Université de Caen
Examinateurs	:	Béatrice DAILLE, Professeur	Université de Nantes
		Emmanuel MORIN, Professeur	Université de Nantes
		Kamel SMAÏLI, Professeur	Université Nancy 2

Directeur de thèse : Pr. Béatrice DAILLE
Encadrant de thèse : Pr. Emmanuel MORIN

Laboratoire: **L**ABORATOIRE D'**I**NFORMATIQUE DE **N**ANTES **A**TLANTIQUE.
CNRS UMR 6241. 2, rue de la Houssinière, BP 92 208 – 44 322 Nantes, CX 3.

ALIGNEMENT MULTILINGUE EN CORPUS COMPARABLES SPÉCIALISÉS

CARACTÉRISATION TERMINOLOGIQUE MULTILINGUE

Multilingual alignment from specialised comparable corpora

Multilingual terminology characterisation

Emmanuel PROCHASSON

⋈

favet neptunus eunti

Université de Nantes

Emmanuel PROCHASSON

Alignement multilingue en corpus comparables spécialisés

Caractérisation terminologique multilingue

IV+X+122 p.

Ce document a été préparé avec LaTeX2e et la classe `these-LINA` version v. 0.1 de l'association de jeunes chercheurs en informatique LOGIN, Université de Nantes.
Cette classe est conforme aux recommandations du ministère de l'éducation nationale, de l'enseignement supérieur et de la recherche (circulaire n° 05-094 du 29 mars 2005), de l'Université de Nantes, de l'école doctorale « Sciences et Technologies de l'Information et Mathématiques » (ED-STIM), et respecte la norme AFNOR NF Z41-006 (octobre 1983) *Présentation des thèses et documents assimilés* de l'association française de normalisation (AFNOR).

Computers are good at following instructions, but not at reading your mind.

— Donald E. Knuth, the TEXbook (page 9).

Il se prénommait Rinri, ce qui signifie Moral, […] mais l'onomastique japonaise est coutumière des hapax.

— Amélie Nothomb, Biographie de la faim.

Résumé

Les corpus comparables rassemblent des documents multilingues n'étant pas en relation de traduction mais partageant des traits communs. Notre travail porte sur l'extraction de lexique bilingue à partir de ces corpus, c'est-à-dire la reconnaissance et l'alignement d'un vocabulaire commun multilingue disponible dans le corpus. Nous nous concentrons sur les corpus comparables spécialisés, c'est-à-dire des corpus constitués de documents révélateurs de la terminologie utilisée dans les langues de spécialité. Nous travaillons sur des corpus médicaux, l'un deux couvre la thématique du diabète et de l'alimentation, en français, anglais et japonais ; l'autre couvre la thématique du cancer du sein, en anglais et en français. Nous proposons et évaluons différentes améliorations du processus d'alignement, en particulier dans le cas délicat de la langue japonaise. Nous prolongeons ce manuscrit par une réflexion sur la nature des corpus comparables et la notion de comparabilité.

Mots-clés : corpus comparables, langue de spécialité, alignement multilingue

Abstract

Comparable corpora are sets of documents written in different languages, which are not translations of each other but share common features, such as the topic or the discourse type. Our work concerns bilingual lexicon extraction from such corpora, in other word, the process of finding translation pairs among the common multilingual vocabulary available in comparable corpora. We focus on specialised comparable corpora, for they are likely to reveal the terminology proper to specialised language. We work on corpora made of medical documents: one of them covers the topic of diabetes and feeding, in French, English and Japanese; the other one covers the topic of breast cancer, in French and English. We propose several improvements for the classical alignment process, especially concerning the delicate case of the Japanese language, distant from French and English. We conclude this thesis with thoughts concerning the nature of comparable corpora and the question of comparability.

Keywords: comparable corpora, specialised language, multilingual aligment

Remerciements

J’adresse mes plus sincères remerciements à mes encadrants qui ont consacré énormément de temps à me soutenir. Un grand merci à ma directrice de thèse, Béatrice Daille, qui, malgré un emploi du temps bien rempli a pris le temps de mettre en avant les défauts de mon travail (et quelques fois les qualités), quitte à traverser l’Oural pour ça. Merci à Emmanuel Morin, pour sa disponibilité et le temps que j’ai dû lui faire perdre en me présentant régulièrement à l’improviste dans son bureau pour lui faire part de mes inquiétudes scientifiques, auxquelles il a toujours pris soin de répondre. Merci à eux pour leurs compétences scientifiques et pédagogiques, et pour m’avoir donner l’opportunité de collaborer avec eux au sein de l’équipe TALN.

Merci aux rapporteurs de ce travail, Yves Lepage et Éric Gaussier pour avoir eu le courage de le disséquer pour en révéler les défauts et, heureusement, les qualités. Merci également à Kamel Smaïli pour avoir accepté de présider mon jury de thèse et pour ses commentaires constructifs.

Mes remerciements vont aussi à Kyo Kageura et Akiko Aizawa, qui m’ont accueilli à Tokyo pendant trois mois (et m’ont permi de fouler à la fois la moquette du gratte-ciel du NII et le campus de l’Université de Tokyo), souvenirs qui resteront longtemps gravés dans ma mémoire. Cette période de collaboration a été le point de départ de nombreuses réflexions et je souhaite de tout coeur pouvoir renouveler cette expérience, d’un point de vue scientifique comme personnel.

Merci au Conseil Général de Loire-Atlantique pour avoir financé ce travail, mais aussi pour avoir été à mon écoute pour améliorer le dispositif de financement, pour avoir été aussi attentif aux problématiques des doctorants et de la recherche scientifique en général.

Sommaire

— Corps du document —

Introduction

Motivations

Les corpus comparables sont l'objet d'une attention particulière depuis le milieu des années 1990. Ils sont constitués de documents dans des langues différentes n'étant pas en relation de traduction, à l'inverse des corpus parallèles. Ils sont donc plus faciles à construire que ces derniers car moins contraints : les corpus parallèles nécessitent un travail coûteux de traduction humaine qui limite généralement leur disponibilité aux textes légaux (dépôt de brevet, actes de parlements…) ou à des traductions de, ou vers l'anglais. À l'inverse, les corpus comparables sont généralement plus disponibles, en tout cas plus faciles à constituer dans de nombreuses langues et pour des thématiques variées. Ce travail porte sur l'extraction lexicale bilingue à partir de corpus comparables, il s'agit du processus consistant à trouver un vocabulaire commun dans les différentes sous-parties des corpus, puis à l'aligner dans le but d'obtenir des paires de traductions. L'objectif du processus d'extraction et d'alignement de traductions est de permettre de compléter automatiquement des ressources lexicales multilingues. Ces ressources sont précieuses par exemple pour la traduction statistique automatique, mais aussi pour assister le travail des lexicographes et des terminologues en réalisant une partie du travail préliminaire de reconnaissance. La méthode que nous utilisons pour aligner des paires de traductions s'appuie sur la caractérisation du lexique dans un cadre monolingue.

Problématique

La caractérisation d'un mot – et *a fortiori* d'un terme – est enregistrée sous la forme d'un vecteur de contexte. Ces vecteurs sont des structures de données contenant des informations sur l'environnement des mots. Ce sont ces informations que nous comparons entre les vecteurs d'une langue source et ceux d'une langue cible pour obtenir des candidats à la traduction. En effet, nous émettons l'hypothèse que les paires de traductions ont des environnements similaires, et donc des vecteurs de contexte similaires. Nous nous intéressons en particulier à la caractérisation terminologique car nous nous concentrons sur des textes en langue de spécialité. Nous travaillons sur deux corpus : le premier rassemble des documents en anglais et en français et traite du *cancer du sein* ; le second contient des documents en anglais, français et japonais et concerne la thématique *diabète et alimentation*, ce qui nous permet de traiter la question délicate de l'alignement du japonais et de langues indo-européennes. Ces corpus sont peu volumineux (quelques centaines de milliers de mots) : c'est une contrainte supplémentaire pour l'alignement que nos propositions doivent traiter en conséquence. Deux tâches nous intéressent particulièrement : la première est de définir quelles informations enregistrer effectivement dans les vecteurs de contexte ; la seconde est de savoir comment exploiter ces informations.

Propositions

Une première approche se concentre sur la première tâche : nous essaierons de rendre les vecteurs de contexte plus discriminants, autrement dit, plus à même de stocker des informations nous permettant de trancher entre des éléments en relation de traduction et des éléments qui ne le sont pas. Nous réalisons

cette tâche en nous appuyant sur une sous-partie connue du vocabulaire spécialisé : les *points d'ancrage*. Ces points d'ancrage seront employés pour leur fiabilité. Ils serviront à « déformer » les vecteurs de contexte des mots en leur accordant un plus grand poids pour l'étape de comparaison.

Une deuxième proposition se penchera sur la seconde tâche, en exploitant des ressources multi-sources pour renforcer la confiance dans un choix de traduction. Nous combinerons les informations apportées par des alignements anglais-japonais et français-japonais pour discriminer les traductions japonaises obtenues.

Une dernière proposition, relative aux deux tâches, étudiera l'importance de la fréquence des mots à traduire pour déterminer *a priori* la meilleure façon de constituer les vecteurs de contexte dans le but d'obtenir un alignement optimal. Cette approche combine indirectement les informations apportées par les dépendances syntagmatiques, aptes à caractériser les mots fréquents, et les dépendances paradigmatiques, plus efficaces pour caractériser les mots de fréquences plus faibles.

Fort des observations et des expériences menées, nous reviendrons en discussion à la notion de *comparabilité* des corpus comparables, pour mieux comprendre comment les appréhender.

Plan

Ce manuscrit est organisé de la façon suivante. Le premier chapitre aborde des généralités sur l'exploitation de corpus multilingues, et définit les notions de corpus parallèles et de corpus comparables. Une partie de ce chapitre est consacrée en particulier à l'extraction de lexique bilingue à partir de corpus comparables et présente différentes approches, en insistant sur l'*approche directe*, support algorithmique de nos recherches. Le deuxième chapitre présente le contexte de ce travail en introduisant les notions de *langue de spécialité* et de *terminologie* ainsi que les ressources linguistiques utilisées, notamment les corpus comparables sur lesquels nous réaliserons nos expériences. Il présente aussi la notion de *translittération* et les observations de ce phénomène linguistique sur l'un de nos corpus. Cette description sera utile dans le cadre de nos expériences, puisque nous utiliserons les translittérations comme exemple de points d'ancrage. Le troisième chapitre traite de l'acquisition sémantique : c'est la tâche qui consiste à classer un vocabulaire en fonction de ses significations en corpus. Nous montrerons que c'est une tâche très proche de la nôtre et nous en inspirerons largement dans le cadre de l'alignement lexical. L'acquisition sémantique permet d'obtenir des relations sémantiques entre mots dans un cadre monolingue ; dans un cadre multilingue, elle permet d'extraire des relations sémantiques multilingues : les paires de traductions. Le chapitre 4 détaille notre implémentation de l'approche directe. Ce chapitre est l'occasion de justifier nos choix d'implémentation et présente une série d'expériences de référence.

Munis des observations des quatre premiers chapitres, nous introduirons dans le cinquième nos propositions d'améliorations de l'approche directe. Enfin, le sixième chapitre traitera de la statistique des corpus comparables. Il s'agit d'une discussion sur la forme des corpus comparables, la notion de comparabilité et les conséquences à tirer pour leur exploitation.

CHAPITRE 1

Corpus multilingues, extraction lexicale bilingue

Les corpus multilingues sont des corpus composés de documents dans des langues différentes. Nous distinguons et décrivons deux types de corpus multilingues : les corpus *parallèles* et les corpus *comparables*. Les corpus parallèles sont composés de paires de documents étant des traductions mutuelles alors que les corpus comparables sont composés de documents ayant des traits communs (le genre, la période, les thèmes abordés...) sans être des traductions. Nous présentons dans ce chapitre les propriétés et les différents usages des corpus multilingues pour nous concentrer en particulier sur l'extraction lexicale à partir de corpus comparables, objet de notre étude.

1.1 Corpus multilingues

1.1.1 Corpus informatiques

Avant de présenter en détail les corpus multilingues, il nous faut définir la notion de *corpus*. Le *Trésor de la Langue Française*[1] en propose une, concise : « *Recueil de documents concernant une même matière.* ». Cette entrée précise toutefois que, dans le cas de la linguistique, un corpus est un « *Ensemble de documents servant à une analyse linguistique* ». Sinclair propose une définition plus adaptée au Traitement Automatique du Langage Naturel (TALN) :

> « *Un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques et extra-linguistiques explicites pour servir d'échantillon d'emplois déterminés d'une langue* » (Sinclair, 1996, page 4)

Les corpus sont des ressources utilisées pour mettre en évidence des phénomènes langagiers. Une des principales questions qui se pose concerne leur *représentativité*. Dans quelles mesures peut-on généraliser ce qui a été observé dans un corpus ? En d'autres termes, les phénomènes observés au sein du corpus correspondent-ils à une réalité à l'extérieur du corpus ? Cette représentativité peut être très spécifique tant qu'elle est explicite. Nous pouvons par exemple nous intéresser au vocabulaire des brèves journalistiques (en s'appuyant sur un corpus de brèves), sans pouvoir généraliser au vocabulaire d'autres registres, par exemple les textes religieux. Sinclair (1996) parle ainsi d'« *échantillon du langage* ». Les critères à respecter pour obtenir un corpus cohérent et représentatif sont, par exemple, le nombre de documents, la

[1] Dictionnaire de la langue française des XIX^{ème} et XX^{ème} siècle.

période d'écriture des documents, le registre, le style, le média de diffusion, les auteurs, etc. (Pearson, 1998). Il s'agit des critères qui peuvent influencer la façon dont les documents ont été rédigés et leur forme finale. Lors de la constitution d'un corpus, ces critères doivent être sélectionnés soigneusement pour garantir la représentativité du corpus pour les phénomènes étudiés.

Les corpus monolingues sont de plus en plus nombreux et volumineux, notamment en raison de l'augmentation des publications sous forme électronique, ce qui rend leur traitement et leur stockage plus aisés. La forme même des corpus évolue : de réservoir à textes, ils sont devenus des sources d'informations parfois complexes, accompagnées d'annotations et de méta-informations. Les outils pour les manipuler sont eux aussi plus accessibles, transformant les corpus en ressources de plus en plus incontournables pour le linguiste.

Les corpus multilingues sont des corpus monolingues mis en relation. Ils sont composés de plusieurs sous-corpus, un par langue concernée. Dans cette section, nous présentons les propriétés des corpus parallèles puis des corpus comparables, pour introduire leurs utilisations en section 1.2.

1.1.2 Corpus parallèles

Les corpus parallèles sont des ensembles de paires de documents dans des langues différentes étant des traductions mutuelles. Dans un corpus parallèle, nous distinguerons les documents *sources* auxquels seront associés des documents *cibles*, traduction des documents sources. La *Pierre de Rosette* (figure 1.1) est un exemple canonique de corpus parallèle. Trouvée en 1799 en Égypte, elle est gravée d'un même texte dans trois écritures différentes et deux langues (démotique, grec et hiéroglyphe). Elle fût la clé utilisée par Champollion pour percer le mystère des hiéroglyphes en 1822.

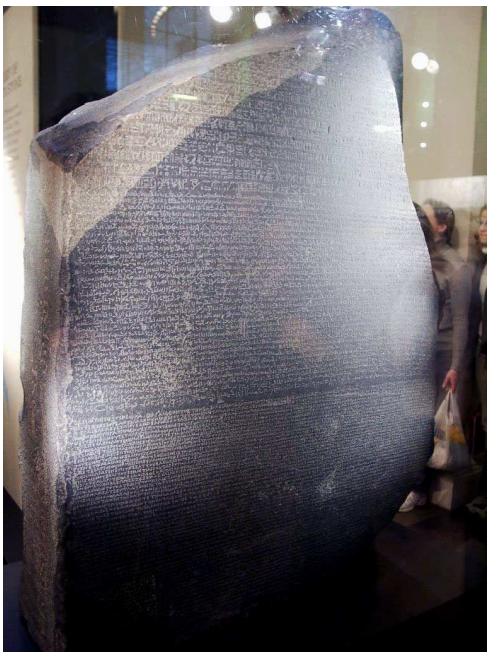


Figure 1.1 – Une reproduction de la Pierre de Rosette, au British Museum.

Cet exemple est particulièrement intéressant puisqu'il présente un corpus de taille très modeste, mais qui fût un levier puissant dans la compréhension d'une langue jusqu'alors inconnue et à la structure inhabituelle. Champollion put résoudre cette énigme en s'appuyant sur plusieurs propriétés des textes :

1. il n'y a pas de traduction manquante entre un document *source* et un document *cible* (Langé et Gaussier (1995) parlent de *quasi-bijection*, page 71) ;
2. les positions des mots en relation de traductions sont comparables entre un document *source* et un document *cible* (Langé et Gaussier (1995) parlent de *quasi-synchronisation*).

Ces propriétés permettent de poser plusieurs hypothèses utilisées pour l'exploitation des corpus parallèles (Fung, 1995a) :

– un mot n'a qu'un seul sens au sein d'un même corpus (pas d'homonyme) ;
– un mot a une traduction unique ;
– les fréquences des mots en relation de traduction sont comparables entre un document *source* et un document *cible*.

Les corpus parallèles sont donc un outil précieux pour observer la transposition de phénomènes langagiers d'une langue à l'autre tant qu'ils sont alignés. L'alignement des corpus parallèles est un enjeu majeur de leur exploitation et est largement discuté dans la littérature (Véronis, 2000). Toutefois, ils présentent plusieurs inconvénients limitant leur usage. Tout d'abord concernant leur disponibilité. La traduction d'un document est le travail d'un humain, elle est donc coûteuse. Les documents en relation de traduction sont rares, tout au moins spécifiques à certains domaines. Citons à titre d'exemple le corpus parallèle *EUROPARL* qui rassemble les actes du Parlement Européen traduits dans 11 langues. Dans le cadre général, il est difficile de disposer de ressources linguistiques parallèles suffisantes pour étudier des phénomènes inter-langues, en particulier lorsqu'ils ne concernent pas l'anglais. Par ailleurs, la rédaction du document cible est influencée par le document source, le traducteur devant respecter le contenu et la structure du document d'origine. Les corpus parallèles permettent donc d'étudier *la façon dont des phénomènes langagiers sont traduits d'une langue à une autre*, mais ne permettent pas d'étudier ces phénomènes comme si les documents avaient été rédigés indépendamment.

1.1.3 Vers des corpus comparables

Face aux faiblesses des corpus parallèles, les recherches se sont d'abord tournées vers les corpus parallèles bruités (Fung, 1995b), c'est-à-dire des corpus composés de documents en relation de traduction mais ne respectant pas toutes les contraintes décrites précédemment (certaines traductions pouvaient manquer ou être décalées dans les documents). Par la suite, les recherches se sont penchées sur des corpus *non-parallèles* avant de définir et d'adopter les *corpus comparables* comme objet d'étude. Une définition en est donnée par Déjean et Gaussier (2002) :

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

Cette définition est intéressante car elle englobe également les corpus parallèles (qui partagent effectivement un vocabulaire commun) et reste vague sur la quantité de vocabulaire partagé nécessaire pour définir que deux sous-corpus forment un corpus comparable. Cette définition amène tout naturellement à la question du degré de comparabilité des documents des sous-parties d'un corpus comparable. La notion de comparabilité est relative à la quantité de traits (qualitatifs et quantitatifs) partagés par les documents du corpus. La figure 1.2 schématise la notion de comparabilité par rapport à la définition des corpus.

Cette figure montre tout d'abord que les documents parallèles sont *parfaitement comparables*. Ils ont, par définition, tout en commun sauf la langue d'écriture des sous-corpus. À l'inverse, des corpus non reliés sont très faiblement comparables. Ils le sont toutefois, car il est probable qu'ils partagent tout

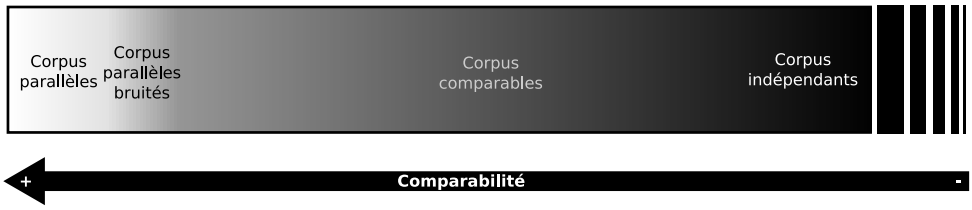


Figure 1.2 – Comparabilité des corpus multilingues.

de même un vocabulaire commun (la propriété inverse serait étonnante), mais le volume de ce vocabulaire peut être considéré comme *négligeable*, en accord avec la définition donnée précédemment.

Toutefois, cette définition est réductrice ; (nous nous y tiendrons néanmoins par la suite, nous intéressant tout particulièrement à la *sous-partie non négligeable* du vocabulaire commun entre les deux langues). En effet, le concept de corpus comparable et leurs utilisations dépendent largement du point de vue de l'expérimentateur et de l'objet de ses recherches ; il serait alors présomptueux d'en proposer une définition universelle. Bowker et Pearson, (2002, page 95) proposent une définition plus large des corpus comparables :

> « *Un corpus comparable est composé d'ensembles de textes, dans des langues différentes, qui ne sont pas des traductions mutuelles*[2] »

Les auteurs précisent que la *comparabilité* implique un nombre de traits communs entre chaque partie du corpus, le seul trait obligatoirement différent étant la langue de chaque partie. Cette définition impose beaucoup moins de contraintes sur les documents constituant le corpus ; notons qu'elle englobe elle aussi les corpus indépendants mais rejette les corpus parallèles.

Nous présentons dans la section suivante différents usages des corpus, parallèles ou comparables. Nous mettons en avant les différentes propriétés attendues de ces corpus en fonction de l'usage pour lequel ils sont destinés.

1.2 Utilisation des corpus multilingues

1.2.1 Analyse contrastive multilingue

Les corpus multilingues, qu'ils soient parallèles ou comparables sont des outils précieux pour l'analyse contrastive multilingue. Aijmer *et al.* (1996) décrivent plusieurs avantages des corpus multilingues.

- Ils révèlent des phénomènes propres à une langue, difficiles à remarquer dans un cadre monolingue (c'est-à-dire, des phénomènes relevés dans un langage qui n'apparaissent pas dans un autre).
- Ils mettent en évidence des différences entre langues (d'un point de vue syntaxique, typologique ou culturel) mais soulignent aussi des phénomènes universels.
- Ils éclairent les différences entre textes traduits et textes sources, mais aussi entre textes « natifs » d'une langue et textes traduits.

Dans ce type de travaux, l'emploi des corpus comparables sera différent de celui des corpus parallèles. Les corpus parallèles pourront servir, par exemple, à mettre en évidence des phénomènes propres au processus de traduction (voir par exemple Pastor *et al.*, 2008) alors que les corpus comparables serviront à comparer des usages *naturels* d'une langue à l'autre. De ce point de vue, les différentes parties

[2] « *Comparable corpora consist of set of texts in different language that are not translations of each other* »

d'un corpus sont utilisées comme des *calques superposables*, pour mettre en lumière les points communs et les différences interlangues.

1.2.2 Lexicographie

Les corpus multilingues sont également utilisés pour créer ou compléter des ressources linguistiques. Ils sont alors un outil précieux pour le lexicographe puisqu'ils permettent de suivre automatiquement l'évolution d'une langue pour mettre à jour de nouvelles traductions. Les corpus multilingues sont particulièrement intéressants dans les domaines de spécialité qui emploient un vocabulaire précis et propre au domaine, susceptible d'évoluer très rapidement au fil du progrès des sciences et des techniques. Par exemple, le *grand dictionnaire Oxford-Hachette français-anglais* a été compilé à partir d'un corpus multilingue de plus de 10 millions de mots (Véronis, 2000 ; Grundy, 1996 ; Knowles, 1996).

Ces ressources permettent par ailleurs d'observer un mot et ses traductions dans leurs contextes. Elles permettent également de rechercher et d'aligner des structures plus complexes que les mots simples, enregistrant des termes complexes et des collocations (voir sections 2.1.1 et 3.2.1). Dans ce contexte, les corpus multilingues peuvent être vus comme des ensembles de *sacs de mots* ou de sacs d'*expressions* ; nous puisons dans ces *sacs* le vocabulaire que nous cherchons à aligner.

1.2.3 Traduction automatique statistique et assistance à la traduction

Brown *et al.* (1990) proposent une approche probabiliste pour la traduction automatique, basée sur l'apprentissage de modèles de langue à partir de corpus parallèles. Ce travail discute différents problèmes relatifs à la traduction, typiquement le problème de la fertilité : un mot peut être traduit par n mots et réciproquement (Tillmann et Ney, 2003). Par exemple *not* en anglais sera généralement traduit par une structure syntaxique *ne...pas*. Par ailleurs, l'ordre des mots peut être différent, en fonction de contraintes grammaticales, c'est le phénomène de distortion. Un exemple emprunté à Brown *et al.* (1990), en figure 1.3, présente un alignement entre l'anglais et le français. Il montre que les mots ne peuvent en général pas s'aligner un à un.

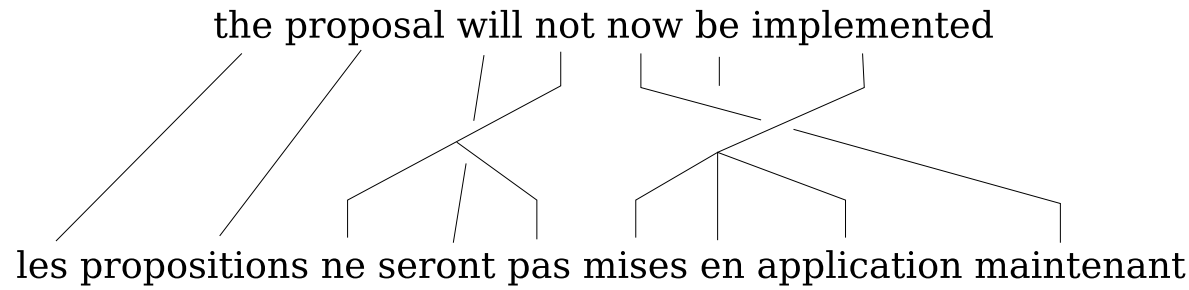


Figure 1.3 – Alignement français-anglais.

Brown *et al.* (1990) utilisent un corpus parallèle (le *Hansard*, actes du parlement Canadien) pour apprendre automatiquement des modèles de langue, en alignant les phrases puis les mots des phrases, comme dans l'exemple précédent. Le corpus parallèle d'apprentissage est utilisé pour calculer une estimation de la fertilité et de la distortion. Ces estimations associées à un modèle probabiliste permettent alors d'évaluer la pertinence d'un candidat à la traduction.

Les recherches en traduction automatique se poursuivent et ont ouvert la voie à d'autres recherche pour la traduction humaine assistée par ordinateur, notamment les *mémoires de traduction* : ce sont des bases de données d'exemples de traduction (au niveau phrase, expression ou mot) destinées à venir en aide au traducteur humain. L'alignement et l'extraction lexicale à partir de corpus multilingues permettent de proposer des candidats fiables (car déjà traduits) pour alimenter ces bases de traductions. Ces corpus peuvent aussi, en plus de la production de lexiques plus pertinents et précis, permettre de présenter les candidats à la traduction dans un contexte d'utilisation réel, pour renforcer la confiance du traducteur dans l'usage de la traduction proposée (Sharoff *et al.*, 2006).

1.2.4 Autres applications

Les corpus multilingues sont utilisés dans d'autres contextes, par exemple pour l'extraction d'information interlangue (*Cross-language information retrieval*, CLIR). Elle a pour but de permettre, à partir d'une requête dans une langue, d'extraire des documents pertinents dans d'autres langues. Elle s'adresse aux utilisateurs suffisamment expérimentés dans une langue pour être capable de la lire, sans être capable de la produire de façon suffisamment précise pour effectuer des recherches pertinentes.

Ils peuvent aussi être utilisés dans un but pédagogique dans l'enseignement des langues ou la formation des traducteurs (Pienemann, 1992 ; Jagtman, 1994 ; Bonhomme et Romary, 1995). Zanettin (1998) les utilise par exemple pour leur capacité à couvrir une terminologie particulière dans de nombreux domaines. Dans le cadre de la fouille textuelle, il relève également leur utilité pour mettre en évidence des phénomènes langagiers communs ou disjoints entre langues, et se sert également des mémoires de traduction toujours dans le but de former des traducteurs humains.

Par la suite, nous nous concentrons sur l'extraction lexicale bilingue à partir de corpus multilingues et retenons la définition 1.1.3 : nous utilisons les corpus comparables comme des ensembles de *sac de mots* dans lesquels nous alignons le vocabulaire qui nous intéresse (cf. chapitre 2).

1.3 Extraction bilingue à partir de corpus parallèles

Nous présentons ici le principe de l'extraction lexicale bilingue à partir de corpus parallèles. En effet, il nous paraît intéressant de pouvoir opposer les méthodes d'extraction dans les corpus parallèles aux méthodes mises au point pour les corpus comparables. Notons que nous employons indifféremment les termes *extraction* ou *alignement* pour désigner le processus de repérage des unités de traduction dans les corpus parallèles.

Les méthodes d'alignement s'appuient principalement sur la distribution, c'est-à-dire la position et la fréquence des éléments de traduction d'un document source dans son document cible en s'appuyant sur les deux premières propriétés des corpus parallèles énoncées en section 1.1.2. Les méthodes esquissées ici ne sont pas exclusives : elles sont généralement complémentaires et permettent, lorsque combinées, d'obtenir des résultats optimaux[3]. En particulier, un alignement au niveau des phrases facilitera l'alignement au niveau mots et expressions, et réciproquement (cf. section 1.3.2 – l'alignement au niveau phrastique permet de réduire l'espace de recherche des mots en relation de traduction, et les mots en relations de traduction permettent de repérer les phrases en correspondance).

[3] Toutes les approches résumées ici sont présentées plus en détails dans (Véronis, 2000)

1.3.1 Alignement des phrases

Une première approche pour extraire des couples de phrases part du constat qu'il est fort probable que la première phrase d'un document corresponde à la première phrase de sa traduction. De la même façon, la dernière phrase d'un document correspond sans doute à la dernière phrase de sa traduction et une phrase tirée au milieu d'un document source devrait apparaître dans une fenêtre définie dans le document cible (Kay et Röscheisen, 1988). La figure 1.4 représente l'espace de recherche d'une phrase cible (en ordonnée) pour une phrase source (en abscisse).

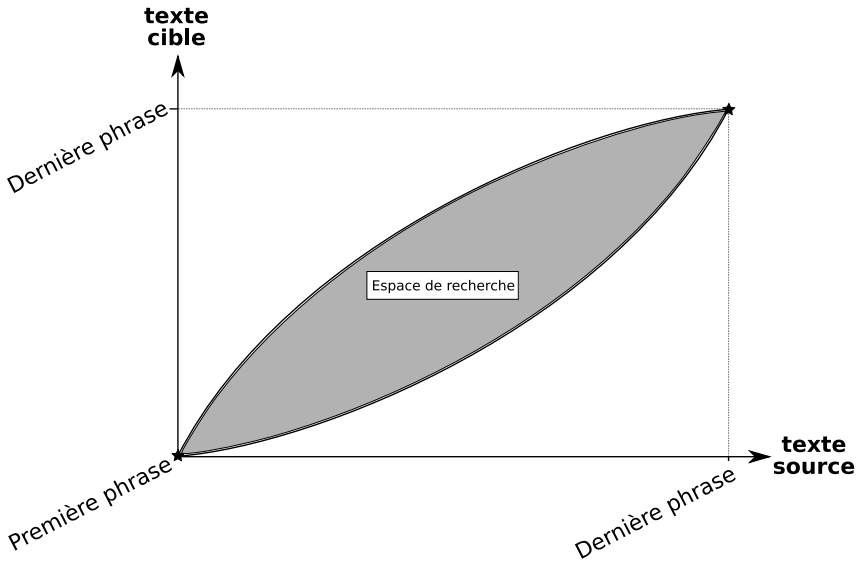


Figure 1.4 – Espace de recherche des phrases en relation de traduction.

La figure 1.4 met aussi en évidence le phénomène de *distorsion* : un mot, une expression ou une phrase ne sera pas traduit par le même nombre de caractères, de mots ou de phrases d'une langue à l'autre. Il apparaît par exemple que les textes français sont plus longs que leurs équivalents anglais (Véronis et Langlais, 2000). Cette différence semble relativement fixe pour un couple de langues donné (Gale et Church, 1991). D'autres phénomènes sont à prendre en compte dans l'alignement, par exemple le phénomène d'inversion (changement de l'ordre des phrases de la source vers la cible, ou au niveau lexical où l'ordre des mots est souvent contraint par la grammaire des langues concernées). Enfin, il est fréquent qu'entre un document et sa traduction, certaines phrases soient fusionnées en une seule ou éclatées en plusieurs, ne permettant plus une correspondance 1 : 1 stricte lors de l'alignement (Gale et Church, 1991).

1.3.2 Recherche autour de points d'ancrage

Il est possible d'affiner l'alignement au niveau phrase en s'appuyant sur des points d'ancrage, c'est-à-dire des informations permettant de relier deux parties des documents à aligner, de façon fiable. Les points d'ancrage peuvent être des informations structurelles associées au document : titre, sous-titre, légende, etc. (Romary et Bonhomme, 2000). Ils peuvent aussi être lexicaux (Kay et Röscheisen, 1988), et être extraits en s'appuyant sur un vocabulaire bilingue. Ce vocabulaire peut être proposé comme amorce, par exemple en utilisant un dictionnaire bilingue (Haruno et Yamazaki, 1996), ou inféré à partir des

propriétés des langues, par exemple en extrayant automatiquement les cognats[4] (Simard *et al.*, 1992) ou les translittérations[4]. Les points d'ancrage identifiés permettent de réduire l'espace de recherche des phrases à aligner. La figure 1.5 schématise le nouvel espace de recherche après identification de points d'ancrage.

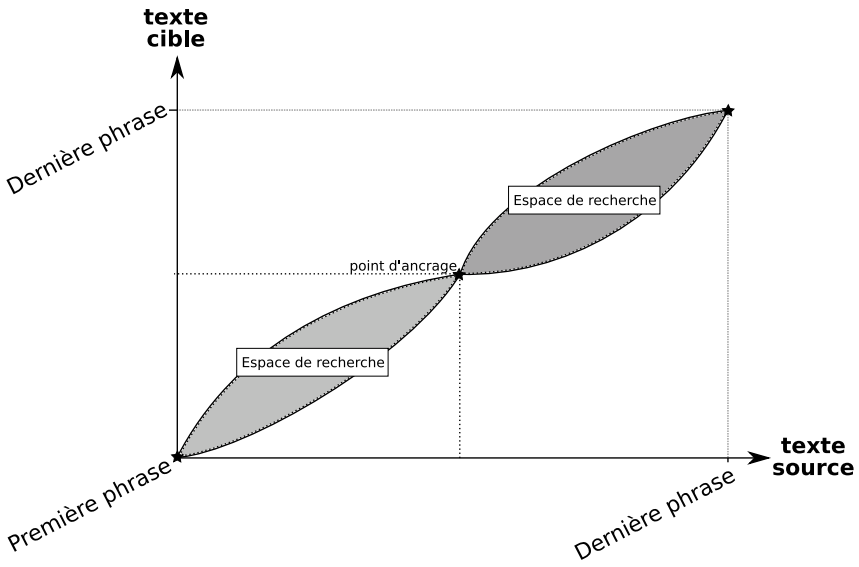


Figure 1.5 – Assistance des points d'ancrage pour l'alignement des phrases.

Les deux espaces de recherches sont disjoints aux phénomènes de distorsion et de fertilité près (qui n'ont un impact que local, autour du point d'ancrage). L'un correspond aux couples de phrases situés *avant* le point d'ancrage identifié, l'autre correspond aux couples de phrases situés *après* le point d'ancrage.

Une fois les couples de phrases alignés, il est possible d'affiner l'extraction en procédant à l'alignement au niveau mot (Brown *et al.*, 1990, 1993). C'est à ce niveau que les différences structurelles entre langues sont les plus visibles (revoir l'exemple de la figure 1.3). En effet, d'une langue à l'autre, l'organisation de la phrase et l'ordre des mots diffèrent grandement. Par exemple, l'épithète précède ce qu'il qualifie en anglais alors qu'il le succède le plus souvent en français. Citons également le cas des langues agglutinantes comme l'allemand qui, pour un terme complexe français ou anglais, proposera un terme simple en traduction. Ces différences sont encore plus marquées entre couples de langues plus éloignées. Ainsi le japonais place toujours le verbe à la fin d'une proposition alors qu'il est souvent placé entre le sujet et le complément d'objet en anglais ou en français. Ajoutons enfin que les mots fonctionnels sont rarement alignables d'une langue à l'autre (typiquement, le japonais n'emploie pas d'article).

Les méthodes d'alignement lexical s'appuient donc généralement sur des modèles de langues appliqués sur les phrases alignées, par exemple des patrons syntaxiques (Daille *et al.*, 1994 ; Blank, 2000) ou des n-grammes (Brown *et al.*, 1993), c'est-à-dire sur des connaissances préalables sur les deux langues alignées.

[4] Les notions de cognats et de translittérations sont précisées au chapitre 2.

1.3.3 Approche par comparaison de distribution

Une approche alternative consiste à construire, pour chaque mot, l'empreinte de sa distribution dans les documents sources et cibles. Cette approche s'appuie sur les hypothèses qu'il n'y a pas de traduction manquante d'un document à l'autre, qu'un mot n'a qu'un seul sens et une seule traduction entre deux documents (Fung, 1995a). Puisqu'aucune traduction ne manque, à chaque fois qu'un mot apparaît à un endroit dans un document source, il apparaît à une position comparable dans le document cible. De plus, puisqu'un mot n'a qu'une traduction, si celle-ci apparaît dans un document cible, elle correspond à une occurrence de ce mot dans le document source. En conséquence, la fréquence et la distribution d'un mot et de sa traduction sont similaires.

Le processus d'alignement enregistre cette distribution dans un vecteur de booléens, en notant la présence ou non des mots étudiés dans une unité textuelle donnée (partie, paragraphe, phrase…) sur l'ensemble du document, pour les deux langues. Reste alors à comparer les vecteurs sources avec les vecteurs cibles pour repérer les distributions les plus semblables, en utilisant une mesure de distance ou de similarité telle que la distance de Hamming (Hamming, 1950). La figure 1.6 schématise cette approche.

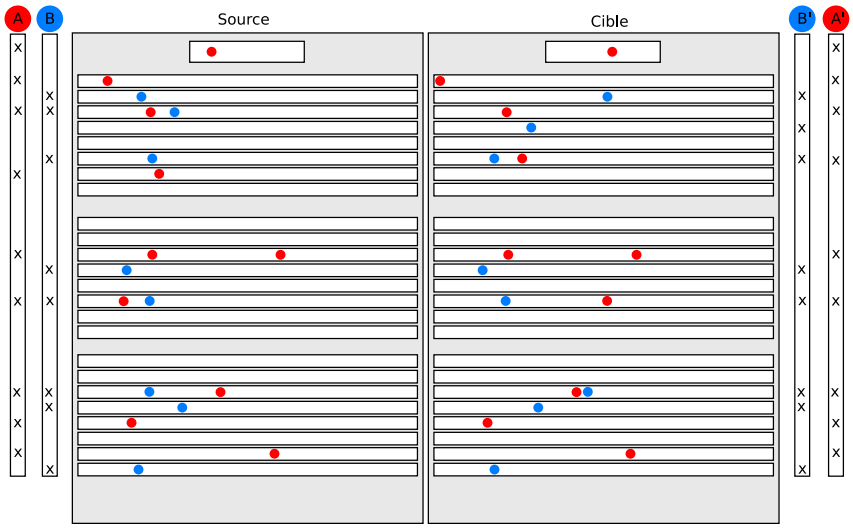


Figure 1.6 – Comparaison des distributions de deux mots et de leur traduction.

Elle met en évidence les distributions de deux mots A et B (et leur traduction, $A\prime$ et $B\prime$) et présente leurs vecteurs de distribution (sur les côtés). En comparant les paires de vecteurs source et cible, il est facile de voir que les distributions de A et $A\prime$ sont proches, tout comme les distributions de B et $B\prime$.

En pratique, les hypothèses énoncées ne sont pas parfaitement respectées comme nous l'avons vu avec le phénomène de distorsion. Toutefois, ces artefacts de traduction sont lissés avec l'augmentation de la taille des corpus.

1.3.4 Limites

L'extraction lexicale à partir de corpus parallèles présente toutefois des faiblesses. Ces faiblesses ne sont pas liées aux approches présentées ici, qui donnent de bien meilleurs résultats que celles utilisées pour les corpus comparables. Elles sont liées à l'intérêt que l'on peut porter aux ressources linguistiques ainsi acquises. En effet, Fung (1995a) remarque qu'extraire du vocabulaire à partir d'un texte parallèle

implique que ce texte a déjà été traduit par un traducteur, ce qui revient à faire de l'ingénierie à rebours pour retrouver le lexique utilisé par le traducteur[5]. C'est efficace dans de nombreux cas (en particulier pour alimenter les mémoires de traduction), mais cela exclut généralement l'exploration lexicographique. Abdul-Rauf et Schwenk (2009) constatent par ailleurs que les corpus parallèles exploités sont généralement caractérisés par un jargon peu utilisé par ailleurs, ce qui est typiquement le cas avec les corpus *Hansard* (parlement canadien) ou *EUROPARL* (parlement Européen), qui sont très largement utilisés (Tillmann et Ney, 2003).

Ce dernier exemple illustre un avantage proéminent des corpus comparables : les textes qui les composent sont plus disponibles sans la contrainte de traduction, permettant ainsi des travaux sur des groupes de langues et des domaines de spécialités peu représentés. L'alignement lexical à partir de corpus comparables est toutefois une opération délicate : il n'est plus possible de s'appuyer sur la distribution des mots dans le document ; de toute façon, rien ne garantit qu'un élément dont on cherche la traduction apparaisse effectivement dans le corpus *cible*[6].

1.4 Approches pour l'extraction à partir de corpus comparables

Nous présentons dans cette section l'évolution de l'exploitation des corpus comparables dans un but d'acquisition lexicographique en détaillant en particulier les travaux de Pascale Fung et Reinhard Rapp.

1.4.1 Premières approches

Les premières approches utilisant les corpus comparables pour construire des lexiques bilingues datent du milieu des années 1990. Rapp (1995) propose une première approche, par comparaison de motifs de cooccurrences ; il pose l'hypothèse que, si dans un texte de langue l_1, deux mots A et B cooccurrent plus souvent que *par hasard* alors, dans un texte en langue l_2, leurs traductions doivent également cooccurrer plus souvent. Ainsi un terme que l'on cherche à traduire et l'association avec ses voisins en langue source forment un motif spécifique. L'association entre deux termes indique dans quelle mesure un terme est relié à un autre (si les deux termes ne sont pas correlés, leur association sera faible), les façons de la caractériser sont plus largement discutées en section 1.5.2. Ce motif spécifique doit pouvoir se retrouver dans la langue cible, entre la traduction du terme et ses voisins. Ainsi, les comparaisons des différents motifs de la langue cible et de la langue source indiquent de potentielles relations de traduction.

Rapp (1995) propose un exemple simplifié, entre l'anglais et l'allemand, présentant des matrices d'association pour quelques mots. Ces matrices sont présentées en table 1.1, et l'alignement de leur motif en table 1.2.

Dans ces matrices, un point dans la case au croisement de deux mots indique qu'ils sont associés (c'est une simplification : en réalité cette association est un scalaire, calculé en utilisant l'information mutuelle, voir section 3). À chaque mot source et cible est attribué un identifiant numérique. L'algorithme d'alignement consiste à réorganiser les matrices de façon à trouver les motifs les plus similaires en langue source et cible, c'est-à-dire à réordonner les mots jusqu'à obtenir un motif similaire entre les

[5] « *the existence of a parallel corpus in a particular domain means* some *translator has translated it, therefore, the bilingual lexicon compiled from such a corpus is at best a reverse engineering of the lexicon this translator used.* » – page 173.

[6] Nous utiliserons toujours les notions de *source* et de *cible* dans les corpus comparables même si dans notre cas cette notion n'est plus aussi définie. Elle tient à la nature de l'alignement : à partir d'une liste de termes issue d'une partie du corpus (la *source*), nous cherchons les éléments de traduction dans les autres parties – les autres langues – du corpus (les *cibles*)

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

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

Table 1.1 – Cooccurrences d'une sélection de mots en anglais (gauche) et allemand (droite) – exemple issu de Rapp (1995).

		1	2	**5**	**6**	**3**	**4**	
blue	1		•	•				*blau*
green	2	•				•		*grün*
sky	5	•						*Himmel*
teacher	6						•	*Lehrer*
plant	3		•					*Pflanze*
school	4				•			*Schule*

Table 1.2 – Alignement des motifs.

deux matrices, comme dans la table 1.2. La matrice source a été réorganisée de telle manière à correspondre (ici, parfaitement) à la matrice cible. Cette approche est toutefois très couteuse en temps de calcul puisqu'elle nécessite un nombre exponentiel de réorganisations et de comparaisons de couples de matrice. Par ailleurs, pour être efficace, elle nécessite un corpus volumineux (dans ce cas, plusieurs millions de mots). Elle peut être améliorée en utilisant une amorce linguistique, c'est-à-dire une liste de mots déjà alignés qui vont contraindre les réorganisations possibles de chaque matrice en bloquant certaines combinaisons (Rapp, 1999).

Fung (1995a) s'intéresse au même problème mais propose une approche sensiblement différente. L'objectif est de mettre en évidence des caractéristiques discriminantes, communes entre un mot et sa traduction mais différentes entre des mots qui ne le sont pas. L'auteur propose de caractériser l'*hétérogénéité du contexte* d'un terme en utilisant deux mesures (une pour l'hétérogénéité à gauche, une pour la droite) indiquant pour un terme donné le nombre de mots différents le précédant immédiatement (à gauche) et le suivant immédiatement (à droite). Ainsi, l'article anglais *the* aura une hétérogénéité à droite très importante, car situé avant de nombreux mots différents, alors que le verbe *am* aura une hétérogénéité à gauche faible, car très fréquemment précédé de *I*. Ces mesures sont utilisées là encore comme une empreinte caractéristique de l'usage d'un terme dans une langue et sont comparées entre deux langues (dans le cadre de l'article de Fung, l'anglais et le chinois) pour extraire des candidats à la traduction.

Pour un mot W :

– hétérogénéité à droite, $x = \frac{a}{c}$

– hétérogénéité à gauche, $y = \frac{b}{c}$

– a : nombre de mots différents immédiatement avant W

– b : nombre de mots différents immédiatement après W

– c : nombre d'occurrence de W

L'auteur compare, à titre d'exemple, l'hétérogénéité du mot *air* (119/176 ; 47/176) = (0,676 ; 0,267) et de sa traduction 空气 (29/37 ; 17/37) = (0,784 ; 0,459) avec l'hétérogénéité du terme 休會/*ajournement* (37/175 ; 16/175) = (0,211 ; 0,091). Les termes *air* et 休會 ont une hétérogénéité très différente, indiquant

que *air* a des contextes bien plus productifs que 休會. En utilisant une mesure de distance adéquate (dans cet article, la distance euclidienne) l'auteur extrait des couples de traduction pertinents[7]. Elle obtient 50 % de candidats corrects parmi les 10 premiers candidats retournés.

1.4.1.1 Vers d'autres approches

Ces deux approches présentent des traits communs. On notera tout d'abord que les hypothèses proposées sont indépendantes du couple de langues concerné dans l'alignement et qu'elles ne s'appuient sur aucune connaissance linguistique préalable. De plus, elles réalisent toutes les deux le saut entre l'alignement à partir de corpus parallèles vers l'alignement à partir de corpus comparables (même si à ce moment là, ils parlent de textes ou de corpus *non-parallèles*). Ils se concentrent non plus sur la fréquence ou la distribution d'un mot dans les corpus, mais sur la comparaison de leur environnement. Fung (1995a) parlera la première de *contexte* en le limitant au voisin immédiat, Rapp (1995) s'intéressera à une fenêtre fixe de cinq mots à gauche et à droite du mot à traduire. Ils font tout deux écho à la citation de Firth (1957) : « *on reconnaît un mot à ses fréquentations*[8] ». C'est l'idée que des mots avec des sens similaires auront tendance à apparaître dans des contextes similaires ; nous y reviendrons dans le chapitre 3.

Le défaut de ces deux approches est qu'elles nécessitent des corpus comparables très volumineux. Rapp (1995) travaille sur un corpus anglais-allemand de 33/46 millions de mots, Fung (1995a) sur un corpus anglais-chinois de 73 618 phrases pour chaque langue. En effet, pour que les motifs de Rapp (1995) soient suffisamment discriminants, il faut qu'ils soient construits sur de grandes quantités de couples de termes pour affiner la pertinence des mesures d'association. De la même façon, la comparaison de l'hétérogénéité entre un mot et un candidat ne peut-être significative que si elle est opposée avec de nombreux autres candidats : la distance entre deux hétérogénéités importe peu, tant que cette distance est inférieure pour deux éléments en traduction qu'elle ne l'est pour les autres éléments. La littérature a rapidement proposé par la suite d'autres approches, plus efficaces mais s'appuyant sur des ressources linguistiques. Nous présentons dans les sections suivantes l'*approche directe* et ses améliorations.

1.5 Approche par traduction directe

L'approche par traduction directe (ou *approche directe*) a été introduite par Fung (1998), s'inspirant des modèles de contextes pour les corpus monolingues et de méthodes de *recherche d'information*[9] et remplaçant les concepts de *requête* et de *document* par ceux de *terme* et *contexte de terme*. Le principe de cette méthode est synthétisé dans le schéma de la figure 1.7. C'est l'approche que nous utiliserons tout au long de cette étude, nous en détaillons les choix d'implémentation au chapitre 4.

L'approche directe repose sur la caractérisation et la comparaison des contextes des termes à aligner en s'appuyant sur une nouvelle structure de données, les *vecteurs de contexte*. Ces vecteurs stockent, pour chaque mot en langues source[10] et cible, un ensemble d'unités lexicales représentatif de leur voisinage. Les candidats à la traduction sont ceux dont les vecteurs de contexte dans la langue cible sont les plus proches du vecteur de contexte (partiellement traduit) du terme à traduire. La traduction des vecteurs de

[7] Nous ne faisons ici qu'esquisser le propos de l'article. L'auteur y présente de nombreux détails permettant d'améliorer les résultats, notamment les méthodes de filtrage appliquées sur le corpus chinois, ainsi qu'une première analyse de cette nouvelle voie de recherche.

[8] *You shall know a word by the company it keeps.*

[9] En particulier le *modèle de Salton*, pionnier de la *recherche d'information* (Salton et Lesk, 1968).

[10] En pratique, les vecteurs de contexte pour la langue source ne seront construits que pour les mots dont la traduction est souhaitée.

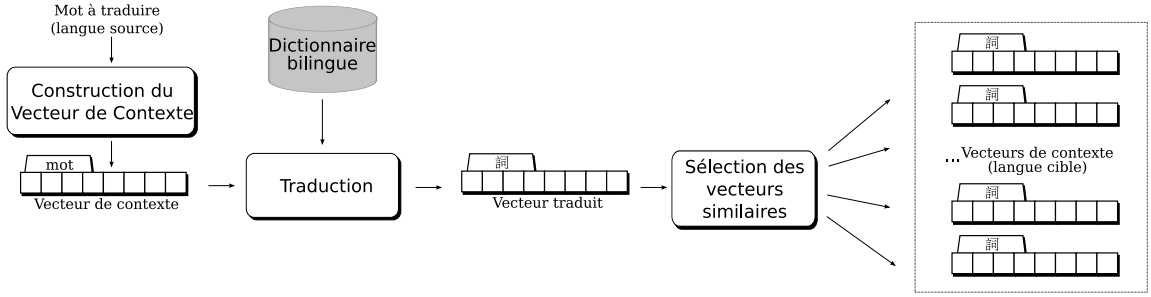


Figure 1.7 – Approche directe.

contexte source s'obtient à l'aide d'un dictionnaire bilingue. L'algorithme de cette méthode se présente ainsi :

- **Construction des vecteurs de contexte** pour chaque unité lexicale i, nous collectons toutes les unités lexicales cooccurrentes dans une fenêtre donnée. Nous obtenons, pour chaque unité lexicale i des corpus source et cible, un *vecteur de contexte* qui regroupe l'ensemble des unités j cooccurrant avec i, associées avec leur nombre de cooccurrences. Nous appelons i la *tête* du vecteur et j les *éléments* du vecteur. Les relations entre les éléments j et la tête i du vecteur sont alors évaluées avec une *mesure d'association*. Les vecteurs enregistrent alors le motif d'association du mot i avec ses voisins j.
- **Traduction des vecteurs de contexte** en utilisant un dictionnaire bilingue. Pour chaque mot dont nous voulons obtenir la traduction, nous traduisons les éléments de son vecteur de contexte.
- **Sélection des vecteurs de contexte proches** en utilisant des mesures de similarité. Plus deux vecteurs de contexte sont proches, plus il est probable qu'ils correspondent à des traductions.

Nous obtenons, pour chaque unité à traduire, une liste ordonnée (par ordre de similarité) des candidats à la traduction. La dernière étape du processus a pour but d'*aligner* les vecteurs traduits et les vecteurs cibles, c'est pourquoi nous parlons invariablement d'*alignement* ou d'*extraction* lexicale dans ce document. Nous en présentons les grandes lignes dans les paragraphes suivants.

1.5.1 Construction des vecteurs de contexte

Le rôle des vecteurs de contexte est de synthétiser, de représenter au mieux un terme donné ; il convient donc de choisir les paramètres de leur construction avec soin. Un premier paramètre est la taille de la fenêtre utilisée pour considérer qu'une unité est voisine d'une autre et devrait apparaître dans son vecteur de contexte. Cette fenêtre peut être fixe (n mot avant, n mot après l'unité considérée) ou variable (phrase, paragraphe…). Par exemple, Déjean et Gaussier (2002) considèrent tous les mots dans la phrase précédant et suivant l'unité étudiée ainsi que dans la phrase la contenant.

Un deuxième paramètre à calibrer soigneusement est la liste des mots que nous souhaitons voir apparaître dans les vecteurs. Il est par exemple maladroit de conserver les mots fonctionnels, peu informatifs et très fréquents, qui ne feront probablement qu'ajouter du bruit dans les vecteurs. Il est donc nécessaire de pré-traiter les corpus pour en extraire les unités lexicales, les lemmatiser pour regrouper les formes fléchies et les étiqueter pour ne pas mélanger les différentes lexies, ne conserver que les unités jugées pertinentes (substantifs, adverbes…) et filtrer certaines catégories de discours (articles, auxiliaires, conjonctions…).

D'autres paramètres peuvent-être pris en compte, par exemple, le nombre minimal de cooccurrences

entre deux mots pour qu'ils soient considérés comme voisins et enregistrés dans les vecteurs. Ce nombre sera faible pour les petits corpus mais pourra être élevé pour des corpus volumineux, pour ne garder que les éléments significatifs. La significativité d'une cooccurrence est aussi évaluée par les mesures d'association.

1.5.2 Mesures d'association

En statistique, une mesure d'association évalue l'indépendance statistique de deux variables aléatoires. Si deux variables sont correlées, leur mesure d'association sera importante, si elles sont indépendantes (si la réalisation de l'une n'influence pas la réalisation de l'autre), leur association sera nulle. En traitement des langues naturelles, ces mesures peuvent être utilisées notamment pour détecter les collocations (Manning et Schütze, 1999), c'est-à-dire, des expressions constituées de plusieurs mots et représentant une façon caractéristique de dire quelque chose[11], et dans un cadre général, pour extraire des relations sémantiques entre les mots. Ces notions sont plus largement développées dans le chapitre 3. Nous utilisons ces mesures d'association pour affiner la caractérisation d'une unité par son vecteur de contexte. Nous mesurons, pour chaque élément d'un vecteur, son association par rapport à sa tête. Si un mot w apparaît très fréquemment dans le corpus (c'est-à-dire qu'il est voisin de beaucoup d'autres unités et sera présent dans de nombreux vecteurs), son association avec la tête i d'un vecteur sera toutefois faible, car affaiblie par le nombre de fois où w apparaît sans i. En revanche, si deux unités i et j sont rares dans le corpus, mais cooccurrent fréquemment, leur mesure d'association sera élevée. L'utilisation de ces mesures à la place d'un simple dénombrement de cooccurrences permet de lisser les artéfacts dus à certains termes très fréquents, mais aussi de normaliser les poids des éléments des vecteurs de contexte entre le corpus cible et le corpus source, la fréquence des unités n'étant pas comparable d'un corpus à l'autre (seules les distributions et fréquences de leurs *contextes* le sont). La figure 1.8 schématise le motif d'un vecteur de contexte formé par ses scores d'associations entre la tête et les éléments du vecteur.

Dans la figure 1.8, la distance entre la tête du vecteur (le centre du motif) et ses éléments (les pointes du motif) représente leur association : plus ils sont éloignés, plus leur association est forte, ce qui se traduit par une surface plus importante dans le motif. Les éléments fortement associés « *déforment* » le motif. Cette représentation a l'avantage de présenter en deux dimensions des vecteurs d'un espace vectoriel dont le nombre de dimensions correspond au nombre de termes présents dans l'ensemble des vecteurs (les éléments dont l'association avec la tête est nulle n'y sont toutefois pas représentés, puisqu'il s'agit des mots qui ne cooccurrent jamais avec la tête du vecteur). C'est à partir des éléments communs de ces motifs que les vecteurs de contexte sources et cibles seront comparés, après l'étape de traduction.

1.5.3 Traduction des vecteurs

Contrairement aux méthodes introduites par Fung (1995a) et Rapp (1995), l'approche directe s'appuie sur des ressources linguistiques pour traduire les vecteurs de contexte. La couverture de ces ressources influence là encore la qualité de l'alignement. Si trop peu de mots sont traduits, la comparaison de vecteurs traduits et des vecteurs cibles ne sera pas significative puisque réalisée sur un échantillon trop faible du vocabulaire (Chiao, 2004). Le pouvoir de caractérisation des éléments non-traduits des vecteurs de contexte disparaîtra lorsque ce vecteur sera transféré en langue cible. D'un autre côté, il est impossible d'obtenir des ressources linguistiques exhaustives, ce qui donne tout son sens aux efforts réalisés pour

[11] « *A collocation is an expression consisting of two or more words that correspond to some conventional way of saying things* », (Manning et Schütze, 1999, page 151)

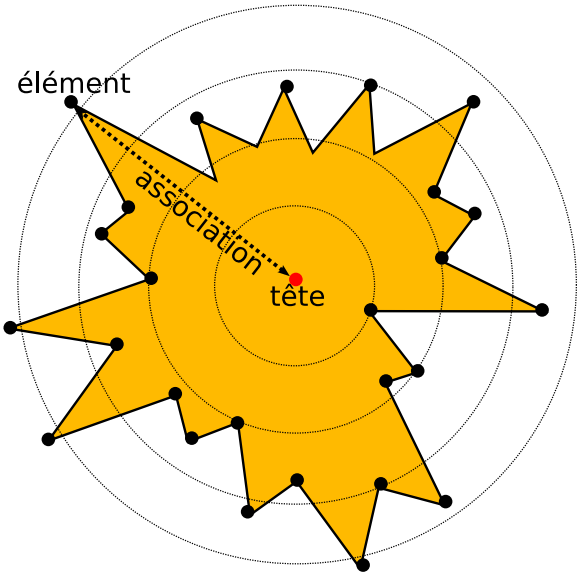


Figure 1.8 – Motif d'association d'un vecteur de contexte.

extraire et aligner du vocabulaire bilingue dans les corpus comparables, pour compléter les ressources linguistiques existantes.

L'utilisation de dictionnaires bilingues pose aussi des problèmes lorsqu'à un mot sont associées plusieurs traductions. Il peut y avoir plusieurs raisons à cela : les traductions sont homonymes ou le terme source est polysémique. Dans ce cas, et comme il est difficile d'évaluer, dans des ressources « plates » comme les lexiques bilingues généralement employés, quelles traductions sont les plus pertinentes (les différentes traductions sont toutes au même niveau), plusieurs approches ont été proposées. La première consiste à prendre en compte toutes les traductions disponibles et à les conserver avec la même priorité dans le vecteur traduit (Déjean et Gaussier, 2002). Fung (1998) propose de considérer les entrées des dictionnaires par ordre décroissant d'apparition. La première traduction proposée aura un poids plus important que la seconde. Cette démarche suppose que les entrées des dictionnaires soient classées par ordre d'importance, ce qui est rarement le cas. Les ressources bilingues et leur utilisation ont donc elles aussi un impact conséquent sur le déroulement du processus d'extraction. Pour contourner ce phénomène, Déjean et Gaussier (2002) proposent une nouvelle méthode dite *par similarité interlangue* que nous présentons en section 1.7.1. L'étape de traduction de l'approche directe permet de transférer les vecteurs de l'espace vectoriel source à l'espace vectoriel des termes de la langue cible. La figure 1.9 présente ce transfert en reprenant le schéma des motifs d'association.

La figure 1.9 présente un exemple de transfert de terme parfait (le terme transféré n'a qu'une traduction, qui bénéficiera du même score d'association), un exemple de traduction manquante (le terme transféré disparaîtra du vecteur traduit, et son association ne sera pas utilisée dans l'étape de comparaison des vecteurs) et un exemple de traduction multiple (dans ce cas, l'une des traductions sera omise).

1.5.4 Comparaison des vecteurs de contexte

Les vecteurs de contexte sont finalement comparés deux à deux (entre l'ensemble des vecteurs de contexte traduits et l'ensemble des vecteurs de contexte de la langue cible) en utilisant des mesures de

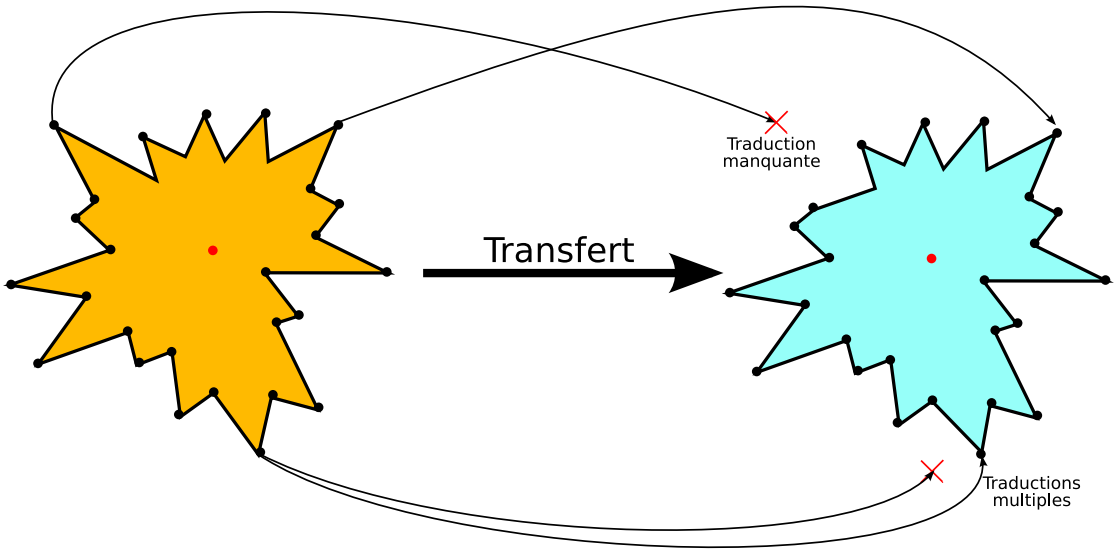


Figure 1.9 – Transfert des vecteurs de contexte.

similarité, telles que le cosinus, la distance de Jaccard ou la distance euclidienne. Les vecteurs les plus proches sont considérés comme des candidats à la traduction. Ces mesures sont développées en détail en section 4.1.6. Cette étape revient à comparer les motifs d'association introduits précédemment, voir figure 1.10.

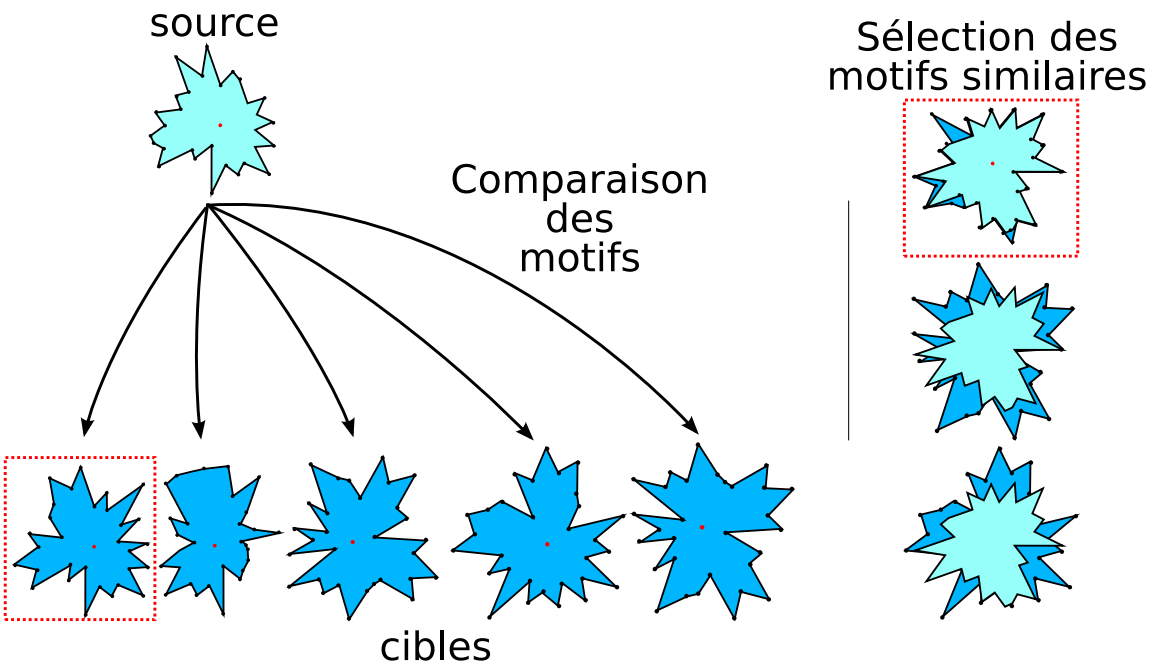


Figure 1.10 – Comparaison des motifs d'association entre vecteurs de contexte sources (traduit) et cibles.

La figure 1.10 montre que la similarité est calculée sur chaque dimension des vecteurs cibles et

vecteurs sources traduits. Les vecteurs sont alors classés par ordre de similarité (du plus similaire au moins similaire), les plus similaires étant les meilleurs candidats à la traduction.

Notons que l'approche directe, à première vue proche de celle proposée par Rapp (1995), est différente car elle ajoute au motif de cooccurrence une information lexicale supplémentaire : dans l'approche directe, les motifs sont basés sur l'association de leurs éléments communs (obtenus par traduction des vecteurs de la langue source – voir section 1.5.2 et 1.5.3) alors que dans le cas de Rapp (1995), ils sont construits et comparés uniquement sur l'ensemble des associations d'un mot à traduire. En d'autres termes, Rapp (1995) cherche à réorganiser les motifs pour qu'ils se superposent au mieux et déduit les relations de traductions à partir des branches superposées dans ces motifs. Les amorces lexicales utilisées dans Rapp (1999) permettent de *fixer* certaines superpositions et de limiter le nombre de réarrangements possibles pour les superpositions de motifs.

1.5.5 Résultats de l'approche directe

Le processus d'alignement renvoie une liste ordonnée de candidats à la traduction pour chaque terme à traduire, classée en fonction de la distance entre les vecteurs des candidats et le vecteur du terme à traduire. Les résultats sont évalués à partir d'une liste de traduction de référence, en comptant le nombre de candidats corrects trouvés dans les n premiers candidats renvoyés, le Top_n.

Il est difficile de comparer les résultats d'une expérience à l'autre, en raison des différences entre les corpus utilisés (en particulier leurs contraintes de construction et leur volume) mais aussi de la couverture et de la pertinence des ressources linguistiques utilisées pour la traduction. Ajoutons que les couples de langues sur lesquels portent les expériences influencent évidemment les résultats de l'alignement : les résultats sont généralement meilleurs entre langues voisines comme le français et l'anglais, qu'entre langues éloignées comme l'anglais et le chinois. À ce jour, il n'existe aucune expérience de référence, et aucun jeu de ressources de référence à notre connaissance[12].

Citons toutefois quelques résultats à titre indicatif. Rapp (1999) obtient 72 % de traductions correctes pour le Top_1 et 89 % pour le Top_{10} avec un corpus comparable composé d'articles de journaux (135 millions de mots pour la partie anglaise, 163 millions pour la partie allemande) et un dictionnaire bilingue contenant 16 380 entrées (termes simples). Chiao et Zweigenbaum (2002), en s'appuyant sur un corpus médical français-anglais de 600 000 mots environ pour chaque langue et un dictionnaire spécialisé de 18 437 entrées, obtiennent 20 % de précision pour le Top_1 et environ 60 % pour le Top_{20}. Ces résultats sont moins bons que ceux de Rapp (1999) mais s'expliquent aisément par la différence de taille des corpus utilisés : le corpus de Rapp (1999) est 200 fois plus grand que celui utilisé par Chiao et Zweigenbaum (2002).

Morin *et al.* (2007) se sont intéressés aux corpus de tailles modestes mais habilement constitués. Dans leur cas, le corpus était fortement contraint selon un type de discours et un thème définis ; il s'agit d'un corpus français-japonais spécialisé sur la thématique du *diabète et de l'alimentation* (que nous utilisons également, il est décrit dans le chapitre 2), scindé en deux corpus *scientifique* et *vulgarisé* de 426/234 et 267/572 milliers de mots chacun (fr-jp). Ils obtiennent 49 % de résultats corrects dans le Top_{10} avec la partie scientifique, en utilisant une ressource linguistique de 173 000 entrées.

[12] Nous avons tenté d'obtenir des ressources utilisées par d'autres chercheurs, mais ne sommes parvenus à en obtenir aucune, en particulier à cause des contraintes de droits d'auteurs. Ce problème est notamment traité par Sharoff (2006).

1.6 Améliorations de l'approche directe

L'approche par traduction directe a été améliorée en profondeur ou dans les détails par de nombreux travaux.

1.6.1 Ressources linguistiques

Les ressources linguistiques utilisées pour le transfert des vecteurs de contexte ont une influence directe sur la qualité de l'alignement. Quelques travaux se sont penchés sur ce problème : Koehn et Knight (2002) proposent notamment de compléter automatiquement ces ressources à partir des relations de cognats présentes à l'intérieur des corpus comparables. Les auteurs détectent automatiquement les mots avec une graphie proche entre l'anglais et l'allemand, en s'appuyant sur des règles de transformation simples, par exemple, en remplaçant `k` et `z` en allemand par `c` en anglais, comme dans *Elektrizität* → *Electricity*. Ils utilisent ce lexique généré pour la comparaison des contextes.

Déjean et Gaussier (2002) ; Déjean *et al.* (2002) s'appuient sur un thesaurus multilingue (le MeSH), ce qui leur permet d'exploiter la hiérarchie des classes de la ressource, hiérarchie indisponible dans le cas d'une ressource bilingue plate. Ils montrent un gain significatif en utilisant la hiérarchie du thesaurus. Avec cette ressource supplémentaire, ils font passer la précision de 57 % à 63 % pour le Top_{20}. Chiao (2004) utilise le métathesaurus UMLS (contenant notamment le thesaurus MeSH) en tant que ressource lexicale spécialisée, pour l'exploitation de corpus comparables spécialisés en médecine.

1.6.2 Hypothèse de symétrie distributionnelle

L'hypothèse de symétrie distributionnelle introduite par Chiao (2004) repose sur l'idée qu'un mot et sa traduction doivent avoir des contextes similaires dans un sens de traduction comme dans l'autre. Cette proposition n'est pas vraie dans tous les cas, typiquement pour les mots polysémiques dans une langue pour lesquels seule une direction de traduction favorisera leurs alignements.

Cette hypothèse permet de renforcer la qualité de l'alignement : le classement des candidats pour une direction de traduction pourra être révisé grâce au classement dans l'autre direction de traduction. Il s'agit de réaliser une mise en correspondance croisée dans les deux directions de traduction et de mettre en évidence les mots les plus proches des deux côtés simultanément.

Chiao (2004) introduit une mesure de *similarité croisée* qui, à partir des calculs de similarité classiques pour chaque direction, combine l'information apportée par le rang de chaque candidat à la traduction dans chaque direction. La mesure de similarité croisée est la moyenne harmonique de chaque rang, présentée en équation 1.1.

$$MH(r_{sc}, r_{cs}) = \frac{1}{\frac{1}{2}(\frac{1}{r_{sc}} + \frac{1}{r_{cs}})} = \frac{2 r_{sc} r_{cs}}{r_{sc} + r_{cs}} \tag{1.1}$$

Dans l'équation 1.1, r_{sc} est le rang d'une paire de candidats à la traduction de la langue source vers la langue cible, r_{cs} est le rang de la langue cible vers la langue source. Cette mesure favorise le fait d'être bien classé dans au moins une direction de traduction.

Ce travail montre une amélioration notable de la qualité des résultats sur un corpus spécialisé médical, notamment dans le classement du vocabulaire spécialisé (lexique de test construit à partir du thesaurus médical *Snomed*). L'auteur obtient un gain de 25 % dans le cas du Top_1 et de 20 % pour le Top_{30}.

1.6.3 Contraintes syntaxiques et lexicales

Sadat *et al.* (2003) proposent de contraindre les résultats retournés par l'approche directe en fonction de leur catégorie syntaxique. Ils supposent en effet que le rôle d'un mot et de sa traduction sont identiques (un nom sera traduit par un nom, un adjectif par un adjectif...). Ils définissent un ensemble d'équivalences autorisées entre l'anglais et le japonais, s'appuyant sur les résultats d'une étape d'étiquetage morphosyntaxique dans chaque langue, présentés en table 1.3.

anglais	japonais
Nom	Nom
Verbe	Verbe
Adverbe	Adverbe ou Adjectif
Adjectif	Adjectif ou Adverbe

Table 1.3 – Couple de catégories autorisées entre deux candidats à la traduction.

Grâce à ces contraintes, et en utilisant une méthode semblable à Chiao (2004), ils obtiennent un gain d'environ 12 % en précision pour la recherche d'information interlangue.

Shao et Ng (2004) utilisent l'information apportée par les translittérations, c'est-à-dire des mots adaptés d'une langue source vers une langue cible, sur la base de leur prononciation. Ils combinent l'information apportée par le contexte des traductions avec l'information apportée par les translittérations entre l'anglais et le chinois. L'intérêt de ce travail réside dans le fait qu'il permet l'alignement de mots très spécifiques (concernés par le phénomène de translittération) mais rares. Alors que l'approche directe fournit 10 résultats corrects, ce chiffre s'élève à 19 lorsqu'ils utilisent les relations de translittérations.

1.6.4 Traductions des termes peu fréquents

Dagan *et al.* (1999) se sont intéressés à l'estimation de la probabilité de cooccurrence entre deux mots m_1 et m_2 n'étant jamais observés ensemble, à partir des probabilités associées aux voisins m^* de m_1. Il s'agit du problème, connu dans le cadre de la linguistique de corpus, des *fréquences nulles* (*zero-frequency problem*, Witten et Bell, 1991). Ce problème apparaît dans le calcul et l'exploitation des n-grammes, où à l'analyse apparaît une configuration jamais rencontrée lors de l'apprentissage. Ce n'est pas un phénomène marginal, il apparaît même dans les corpus très larges. Essen et Steinbiss (1992) en font l'expérience sur le corpus LOB[13] (corpus d'anglais britannique d'un million de mots). Ils constatent que 12 % des bigrammes contenus dans un quart du corpus n'apparaissent jamais dans les trois quarts restants. Dagan *et al.* (1999) développent donc des techniques pour évaluer les probabilités d'évènements jamais rencontrés lors de l'observation, en inférant ces probabilités à partir des probabilités obtenues pour des éléments proches. Le nœud de ce problème est alors de caractériser les voisins m^* d'un mot m_1. Pour cela, plusieurs mesures sont proposées qui comparent les probabilités conditionnelles de cooccurrences des mots, en admettant l'hypothèse que deux mots voisins auront tendance à avoir des cooccurrences similaires.

S'inspirant de ces travaux, Pekar *et al.* (2006) se sont intéressés à l'extraction et à l'alignement de mots de faibles fréquences. En effet, ils constatent que la caractérisation par le contexte des mots à traduire est trop imprécise dans ce cas (en réalité, dans tous les cas sauf pour les mots les plus fréquents). Ce phénomène est agravé lors de l'étape de traduction, susceptible d'introduire du bruit en raison des

[13] *Lancaster-Oslo/Bergen corpus*, `http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/corpora/list/private/LOB/lob.html`.

traductions multiples et des cas de polysémie dans les ressources linguistiques : seuls les mots à fortes fréquences (c'est-à-dire caractérisés par de nombreuses cooccurrences avec de nombreux éléments) restent suffisamment stables et correctement caractérisés après le transfert en langue cible. Ils proposent une approche basée sur les *k-plus proches voisins*, pour caractériser la probabilité $p(v|n)$ du mot n, cooccurrant avec un mot v, dans le cas d'un nombre de cooccurrences observé nul. Cette probabilité est inférée par la probabilité estimée $p^*(v|n)$, moyenne pondérée des probabilités des voisins n^* de n. Pekar *et al.* (2006) proposent aussi de lisser les vecteurs de contexte dans le cas des cooccurrences observées rares, mais existantes. Dans ce cas, la probabilité estimée $p^*(v|n)$ est une combinaison linéaire de la probabilité observée $p(v|n)$ et de la probabilité estimée avec les plus proches voisins. Ils augmentent ainsi la qualité des résultats de l'alignement, obtenant une réduction significative du rang des candidats à la traduction (évalué par le *t-test*, avec une *p-value* inférieure à $0,001$).

1.7 Approches connexes

D'autres approches ont été proposées pour l'extraction lexicale bilingue à partir de corpus comparables qui divergent de l'approche directe.

1.7.1 Approche par similarité interlangue

L'approche par similarité interlangue repose sur l'hypothèse que *si deux mots ont des distributions similaires, alors ils sont reliés sémantiquement* (Déjean et Gaussier, 2002, page 4). Ainsi, un mot et ses synonymes auront tendance à avoir des contextes similaires même s'ils cooccurrent rarement. Le principe de cette approche consiste à identifier les vecteurs de contexte du dictionnaire bilingue qui sont proches (au sens d'une mesure de similarité) du vecteur de contexte du mot à traduire. Le dictionnaire va permettre de traduire les vecteurs de contexte dans leur globalité et non élément par élément. De cette manière, les vecteurs de contexte transférés perdent moins de leur potentiel de discrimination en langue cible. Le dictionnaire bilingue est alors mieux exploité puisque des traductions candidates peuvent être proposées pour un mot à traduire même si aucun élément de son vecteur de contexte ne peut être traduit. Le processus de cette approche peut alors se décomposer ainsi (voir aussi la figure 1.11) :

1. **Construction des vecteurs de contexte** pour les éléments des corpus source et cible.
2. **Construction des vecteurs de contexte** pour les éléments du dictionnaire bilingue (deux ensembles, un pour chaque langue).
3. **Sélection des vecteurs de contexte similaires** à celui du terme à traduire i dans la ressource bilingue (ensemble E de vecteurs).
4. **Sélection des vecteurs de contexte similaires** à E dans la langue cible.

Les vecteurs sélectionnés dans la langue cible sont les candidats à la traduction du terme i. Cette méthode exploite beaucoup plus en profondeur la ressource bilingue et a l'avantage de ne laisser aucun terme à traduire de côté, pour peu que les éléments de leur vecteur de contexte soient peu ou pas présents dans le dictionnaire, contournant ainsi les problèmes d'inadéquation des ressources bilingues rencontrés avec l'approche directe.

Le processus de construction des vecteurs de contexte est le même que pour l'approche directe. Les vecteurs de contexte des termes de la ressource linguistique sont construits à partir de leurs occurrences dans les corpus. Le vecteur de contexte v_i de l'élément i à traduire est comparé aux vecteurs de contexte de la ressource linguistique. Les vecteurs de la ressource linguistique sont naturellement reliés aux vecteurs de contexte de leur traduction. Ils sont, comme le disent Déjean et Gaussier (2002) à « *double*

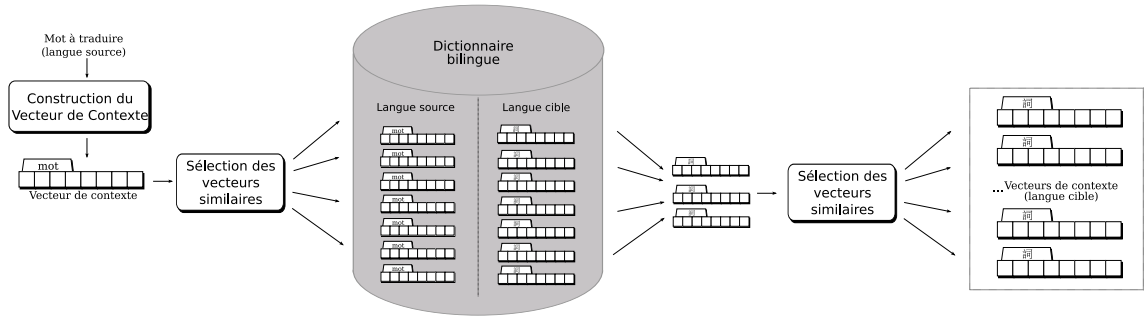


Figure 1.11 – Approche par similarité interlangue.

face », une face en langue source et une face en langue cible. Les vecteurs de la ressource proche du vecteur à traduire sont sélectionnés et leur face cible est comparée avec les vecteurs de contexte issus de la partie cible du corpus. Les vecteurs cibles les plus similaires à ceux de la ressource sont à leur tour extraits pour être candidats à la traduction du terme i.

Avec cette méthode, Déjean et Gaussier (2002) obtiennent pour des termes simples français-allemand une précision, pour les Top_{10} et Top_{20}, de 43 % et 51 % pour un corpus médical de 100 000 mots (respectivement 44 % et 57 % avec l'approche directe) et de 79 % et 84 % pour un corpus de sciences sociales de 8 millions de mots (respectivement 35 % et 42 % avec l'approche directe).

1.7.2 Approches géométriques

Gaussier *et al.* (2004) proposent de regarder le problème de l'extraction lexicale à partir de corpus comparables d'un point de vue géométrique. Cette approche est assez naturelle, puisque nous travaillons sur des *vecteurs* de contexte, c'est-à-dire des entités mathématiques appartenant à des *espaces vectoriels* et sur lesquels nous pouvons effectuer des opérations algébriques classiques. De ce point de vue, le processus d'alignement repose sur deux espaces vectoriels : l'espace des mots *source* et l'espace des mots *cible*, chaque mot distinct représentant une dimension. Les vecteurs s'inscrivent dans ces espaces : chaque composante des vecteurs correspond à la mesure d'association des éléments des vecteurs, chaque élément étant une dimension de l'espace vectoriel, les éléments non associés avec la tête du vecteur ont donc une composante nulle pour cette dimension. L'étape de traduction consiste donc à transférer les vecteurs en langue source vers l'espace vectoriel de la langue cible.

Ce point de vue met en évidence certains défauts de l'approche directe. Typiquement, elle fait l'hypothèse d'orthogonalité des dimensions des espaces vectoriels (Chiao, 2004). En d'autres termes, elle part du principe que tous les mots sont indépendants entre eux, ce qui est simplificateur en raison des relations qu'entretiennent les mots entre eux (synonymie, hyperonymie, antonymie, collocations…). Gaussier *et al.* (2004) proposent deux approches pour pallier ce problème. La première consiste à représenter les vecteurs de contexte non plus dans l'espace vectoriel des mots du corpus, mais dans l'espace vectoriel des *vecteurs de contexte construits à partir des dictionnaires bilingues*. Il est intéressant de noter que cette approche généralise l'approche par similarité interlangue proposée dans Déjean et Gaussier (2002) avec des motivations différentes.

La seconde approche s'inspire de l'*Analyse Sémantique Latente* (*Latent Semantic Analysis* – LSA, Deerwester *et al.*, 1990). Cette méthode a pour but de réduire les dimensions des espaces vectoriels en proposant une combinaison linéaire des dimensions initiales. Cette méthode a été notamment exploitée

en *recherche d'information* (Berry *et al.*, 1995) : elle est censée permettre de caractériser au mieux des documents, dérivant automatiquement des indices statistiques au lieu de reposer sur le lexique des documents. Hofmann (1999) en a proposé une amélioration, l'*Analyse Sémantique Latente Probabiliste* (PLSA), utilisée par Gaussier *et al.* (2004). Le principe reste sensiblement le même : il s'agit de réduire les dimensions d'espaces vectoriels à des combinaisons linéaires *pertinentes* de dimensions (de plusieurs milliers à quelques centaines de dimensions).

Les résultats de ces deux approches restent moins bons que ceux obtenus avec l'approche directe simple. Elles permettent toutefois un gain en précision lorsque combinées avec d'autres approches (voir section 1.7.4). Cette étude apporte un regard différent sur l'approche directe et ses faiblesses.

1.7.3 Traduction compositionnelle

Toutes les approches présentées jusque là se concentraient sur l'alignement de termes simples (formé d'un seul mot – nous revenons sur la notion de terme dans le chapitre 2 – par exemple *diabète*). Nous avons laissé de côté pour le moment l'alignement de termes complexes (composés de plusieurs mots – par exemple *diabète de type 2*). En premier lieu se pose la question de leur identification dans un cadre monolingue pour la construction des vecteurs de contexte. Il faut en effet être capable de reconnaître et de regrouper les variantes des termes. Morin et Daille (2004) regroupent ainsi les variantes morphologiques dérivationnelles synonymiques telles que *produit de la forêt* et *produit forestier*. Les vecteurs de contexte ainsi constitués sont plus spécifiques car les termes complexes tendent à porter un sens plus précis. Ils permettent donc une meilleure caractérisation des termes dans un contexte monolingue, mais pas nécessairement un meilleur alignement bilingue. En effet, les termes complexes ont des fréquences d'apparition plus faibles que les termes simples (le terme *produit de la forêt* aura une fréquence d'apparition inférieure ou égale à la fréquence de *produit* et de *forêt*). L'association d'un terme complexe avec la tête d'un vecteur est donc moins discriminante dans l'alignement que celle des termes simples.

Au delà du paradoxe de la représentativité des termes complexes se pose aussi celui de leur traduction dans le cadre de la méthode directe (section 1.5.3). En effet, les ressources linguistiques sont généralement très pauvres en traduction directe de termes complexes (à quelques exceptions notables telles que les expressions idiomatiques et les collocations fréquentes). Une approche classique consiste à réaliser une traduction compositionnelle, c'est-à-dire à traduire chaque mot du terme complexe indépendamment et à observer l'ensemble des combinaisons possibles (Grefenstette, 1999). La traduction compositionnelle est un processus complexe, en raison de plusieurs phénomènes linguistiques (Robitaille *et al.*, 2006 ; Morin et Daille, 2004, 2008) :

- **Fertilité** : un terme complexe source et sa traduction en langue cible peuvent être de taille différente. Exemple : *table de vérité* – . La traduction peut même être non-compositionnelle, par exemple 赤点 (litt. *rouge point*), traduction japonaise de *échec*.
- **Variabilité de traduction** : un terme complexe source peut avoir plusieurs traductions concurrentes, tel que *champ électromagnétique* traduisible en 電磁場 (littéralement *électromagnétique champ*) et 電磁界 (litt. *électromagnétique région*).
- **Modifications de motifs syntaxiques** : le terme source et sa traduction ne partagent pas nécessairement la même syntaxe dans leur composition. Par exemple le terme *cellule graisseuse* (Nom-Adjectif) sera traduit en japonais par 脂肪細胞 (Nom-Nom – litt. *cellule gras*).
- **Instabilité graphique** en particulier dans le cas des noms propres, parfois translittérés, parfois repris tels quels. Par exemple : *syndrome de Cushing* se traduit en カッシング症候群 ou en *Cushing* 症候群 – 症候群 = *syndrome*.

Morin et Daille (2008) proposent d'intégrer un processus de traduction des termes complexes au pro-

cessus global de l'approche directe, en ajoutant des règles morphologiques pour améliorer la traduction compositionnelle. Cette approche suppose une connaissance préalable des langues impliquées dans la traduction.

Les résultats de l'alignement de termes complexes sont généralement moins bons qu'avec les termes simples. Ainsi, Morin *et al.* (2007), qui obtenaient 49 % de résultats corrects avec des termes simples (voir section 1.5.5), n'obtiennent plus que 18 % de résultats corrects dans le Top_{10}, dans les mêmes conditions mais en cherchant à aligner des termes complexes. L'effort pour aligner des termes complexes n'est toutefois pas inutile, en particulier concernant le vocabulaire spécialisé, car ils sont généralement plus précis et spécifiques et donc d'autant plus précieux pour le lexicographe ou le traducteur.

1.7.4 Combinaison d'approche

Il semble naturel de chercher à combiner les différentes approches proposées pour améliorer la qualité des résultats, en accord avec la remarque de Zweigenbaum et Habert (2006, page 38) :

> « *Lorsque l'on dispose de plusieurs méthodes pour résoudre un problème, il est souvent plus productif de chercher à les combiner*. »

Ainsi en réalisant une combinaison linéaire des probabilités de traduction associées aux approches directes et par similarité interlangue, Déjean et Gaussier (2002) indiquent un gain absolu de 20 % au niveau du Top_{10} par rapport aux meilleurs résultats obtenus avec les méthodes utilisées individuellement. Dans l'article où Gaussier *et al.* (2004) introduisent la PLSA, la même stratégie est appliquée. Ainsi pour les résultats exprimés avec la f-mesure pour le Top_{100}, les méthodes individuelles obtiennent les scores suivants : 0,24 (approche directe), 0,27 (approche par similarité interlangue) et 0,20 (approche PLSA) ; et la combinaison des modèles : 0,32 (approches directe et par similarité interlangue) et 0,28 (approches par similarité interlangue et PLSA). Morin (2009) combine les résultats obtenus à partir d'un corpus échantillonné français-anglais[14], en considérant d'une part la moyenne arithmétique des scores de similarité des traductions candidates, mais aussi la moyenne harmonique des rangs des traductions candidates. Il montre un gain immédiat dès l'utilisation d'une combinaison de deux résultats d'alignement. Ce gain atteint quatorze points avec la combinaison de huit candidats ou plus, soit une précision de 69,7 % (contre 54,9 % pour le meilleur alignement isolé).

1.8 Conclusion

Cette section présente l'objet d'étude de cette thèse : les corpus multilingues et en particulier les corpus comparables. Nous avons proposé un tour d'horizon de l'exploitation des corpus multilingues, passant des corpus parallèles, aux corpus parallèles bruités puis aux corpus comparables. Nous avons présenté différents cas d'utilisation en nous concentrant sur l'extraction et l'alignement d'un lexique commun dans des documents n'étant pas en relation de traduction, dans le but de compiler automatiquement des ressources bilingues, utilisables notamment dans le cas de l'assistance à la traduction automatique. Nous avons en particulier étudié l'approche directe, sur laquelle nous revenons plus en détail dans le chapitre 4. Ces quelques bases définissent le cadre de cette thèse, cadre que nous précisons encore dans les chapitres suivants.

[14] Ce travail est plus largement présenté au chapitre 6. Le corpus utilisé est un corpus déséquilibré français-anglais. La partie anglaise est découpée en 14 échantillons de taille comparable à la partie française. Pour chaque mot à traduire, il y donc 14 alignements disponibles, ce sont ces alignements qui seront combinés.

Le chapitre suivant est consacré à la description du cadre linguistique de nos travaux. Nous y présentons en particulier les corpus comparables utilisés pour nos observations et nos expériences ainsi que les choix motivant leur usage. Nous y étudierons également en détail le phénomène des translittérations, observé sur l'un des corpus que nous exploiterons par la suite dans le chapitre 5.

CHAPITRE 2

Contexte, matériel

Ce chapitre est destiné à présenter l'ensemble des ressources linguistiques utilisées dans nos expériences, en particulier les corpus. Nous nous sommes intéressés à des corpus relevant du discours scientifique. Nous disposons de deux corpus dans le domaine médical : l'un traite de l'*alimentation et du diabète*, initialement en français-japonais et complété par une partie anglaise ; l'autre corpus traite du *cancer du sein* en français et en anglais. Nous introduisons les notions de terme et de langue de spécialité, puis nous décrirons les ressources linguistiques utilisées pour l'alignement, c'est-à-dire les dictionnaires bilingues et les listes de références pour l'évaluation. Nous présentons enfin le phénomène linguistique des translittérations en japonais et leurs occurrences dans les textes de spécialité. Cette étude nous permettra par ailleurs d'entrer plus en détail dans l'un des corpus comparable utilisé.

2.1 Langues de spécialité

Nous nous intéressons dans cette étude à des corpus spécialisés, qui reflètent un vocabulaire particulier propre aux domaines des textes utilisés pour les constituer. L'extraction lexicale prend tout son sens dans ce contexte : elle doit permettre de relever une terminologie particulière et de constituer automatiquement des lexiques précis, notamment dans le but de faciliter le travail des lexicographes et des terminologues. Nous revenons d'abord à la notion de terme, puis introduisons celle de sous-langage.

2.1.1 Le terme

Les termes constituent la manifestation linguistique d'objet réels ou immatériels qui partagent des propriétés communes (Sager, 1990). Un terme est, dans une vision classique, « *l'étiquette linguistique d'un concept* » (Jacquemin et Bourigault, 2003, page 599)[1]. Cette première définition est valable à l'intérieur d'un domaine particulier où un terme est censé n'avoir qu'un seul sens. Jacquemin et Bourigault précisent cette définition dans le cas général : « *un terme est le résultat d'une analyse terminologique* ». En d'autres mots, un terme est défini comme tel lorsqu'un terminologue l'a décidé.

Sans entrer plus en détails dans le définition de *terme*, nous retiendrons qu'il s'agit généralement d'un nom ou d'un groupe nominal (Daille, 2002), utilisé dans un contexte précis pour un sens précis. Le terme *clavier* a un sens en informatique, sensiblement différent de celui utilisé en musique, bien que le concept soit finalement proche (L'Homme, 2004). Par ailleurs, la littérature distingue les termes simples, composés d'un seul mot, des termes complexes, c'est-à-dire des syntagmes nominaux constitués d'au moins deux unités lexicales pleines (Sager, 1990). Par la suite, nous nous efforcerons d'employer *terme* lorsque nous nous référerons à un vocabulaire spécialisé et *mot* dans le cas général.

[1] « *a term is considered as the linguistic label of a concept* ».

2.1.2 Sous-langages

Les langues de spécialité se définissent en opposition à la langue générale : la langue générale correspond au langage parlé tous les jours, dans des situations courantes pour exprimer des concepts communs, alors que les langues de spécialité servent à définir des concepts spécialisés dans des domaines de connaissances spécialisés. En d'autres termes, la langue générale est parlée par les locuteurs de naissance d'une langue, alors que les langues de spécialité sont utilisées par les spécialistes d'un domaine scientifique et technique. Pour une même langue, il existe une langue générale et plusieurs langues de spécialité (Bowker et Pearson, 2002).

Les langues de spécialité se caractérisent par l'emploi d'un vocabulaire particulier, mais aussi par des structures grammaticales différentes de la langue générale (Harris, 1988)[2]. Sager (1986) définit un sous-langage comme « *la langue parlée par une communauté spécifique, par exemple, celle concernée par un sujet spécifique ou impliquée dans une activité spécifique*[3] ». Les phrases utilisent un nombre plus limité de structures grammaticales, mais elles peuvent être plus complexes, en particulier à cause de termes composés plus longs (Pearson, 1999). Le choix du terme *sous-langage* est lui même discuté dans la littérature (Habert *et al.*, 1997) car les applications qui les exploitent n'ont pas les mêmes objectifs. Nous utiliserons ce terme comme un équivalent de *langue de spécialité* et nous les exploiterons pour leurs propriétés à être révélateurs d'un usage spécifique d'un vocabulaire spécifique.

Nous distinguons deux critères principaux utilisés pour constituer des corpus représentatifs d'un sous-langage : le domaine et le type de discours.

Le domaine indique les thématiques des documents utilisés. Il peut généralement être affiné en sous-domaine voir en micro-domaine (Chiao, 2004). Par exemple, le domaine des documents d'un corpus peut être l'architecture, la physique nucléaire ou la médecine. Dans le cas de l'architecture, les sous-domaines pourront être l'architecture médiévale ou contemporaine.

Le type de discours indique la nature des documents : qui les a écrits et dans quel but. Nous nous intéresserons exclusivement aux documents *scientifiques*, c'est-à-dire écrits par des experts à destination d'autres experts (Pearson, 1999). Dans notre cas, il s'agit majoritairement d'articles de recherche, comme nous le présentons dans la section suivante.

2.2 Ressources linguistiques

Nous avons à notre disposition deux corpus comparables. Le premier est constitué de trois sous-corpus en anglais, français et japonais et traite du thème « *diabète et alimentation* », le second est un corpus anglais-français sur le thème du « *cancer du sein* ».

2.2.1 Corpus « diabète et alimentation »

Nous avons à notre disposition un corpus comparable en français et japonais, que nous avons complété avec une partie en anglais. Les documents qui le composent ont été extraits à partir du Web et concernent le domaine médical. Ils traitent tous du thème « *diabète et alimentation* ».

[2] Pour Harris (1988), les sous-langages se caractérisent par un non respect des règles transformationnelles proposées dans le cadre de la langue générale, et par l'ajout d'un certain nombre de règles spécifiques. Il le montra en utilisant un corpus sur le domaine de l'immunologie.

[3] « *the language used by a particular community of speakers, say, those concerned by a particular subject matter or those engaged in a specialized occupation* ».

Pour les parties en français et en japonais, trois étapes ont été nécessaires à la constitution de ces corpus. La première a consisté à moissonner le web pour extraire les documents. Deux approches sont possibles, une approche en largeur, qui consiste à traiter l'ensemble des réponses retournées par une seule requête, et une approche en profondeur qui consiste à n'exploiter que les premières réponses, mais en exploitant les liens internes disponibles sur ces pages. Pour le français, les documents ont été sélectionnés en utilisant une stratégie de recherche en profondeur, partant des mots clés *diabète*, *alimentation* et *obésité* puis en poursuivant la recherche à partir des premières réponses (il existe quelques portails médicaux qui regroupent de nombreux documents). Pour le japonais, nous avons effectué une recherche en largeur à partir des 180 000 réponses à la requête 糖尿病 (diabète) et 食事.療法 (régime alimentaire).

La deuxième étape a consisté à classer les documents selon leur type de discours, pour séparer les documents de vulgarisation des articles scientifiques. Cette étape a consisté à choisir manuellement les documents qui allaient être intégrés aux corpus : la qualité des résultats des moteurs de recherche n'est jamais garantie et doit être évaluée manuellement. Les documents ont été triés par des locuteurs de naissance pour chaque langue. Lorsque la classification d'un document n'était pas claire (désaccord entre les personnes réalisant le tri, ou simplement difficulté à catégoriser précisément un document) il était écarté. Enfin, l'ensemble des documents a été normalisé : ils ont tous été convertis en texte brut au format UTF-8 (Goeuriot *et al.*, 2008).

Pour la partie anglaise, nous avons utilisé le site médical *PubMed*[4], de la *U.S. National Libray of Medicine* et du *National Institute of Health*, qui recense des publications scientifiques, classées manuellement par de nombreux relecteurs suivant les catégories du thésaurus *MeSH*. Les documents ont donc pu être facilement extraits à partir d'une seule requête, en sélectionnant les articles en anglais et en ne retournant que les articles disponibles gratuitement :

```
("Diabetes Mellitus/diet therapy"[MeSH] OR
"Diabetes Mellitus/etiology"[MeSH] OR
"Diabetes Mellitus/prevention and control"[MeSH]) AND
("nutrition" OR "feeding")
```

Les documents ont été classés en fonction de leur année de publication et de la nationalité du premier auteur, pour séparer les documents britanniques, australiens, américains, canadiens…

Le tableau 2.1 indique la taille de chaque sous-corpus, après conversion en texte brut et au format UTF-8. Notons que la partie anglaise effectivement récoltée est plus importante. De manière à obtenir des parties équilibrées en nombre de mots[5], seule une sous-partie a été conservée et traitée. Les documents sélectionnés ont tous été publiés après l'année 2000 et les documents en anglais britannique conservés en priorité.

	Français	Japonais	Anglais
Nombre de mots	257 000	235 000	250 000
Nombre de documents	65	119	61

Table 2.1 – Taille de chaque partie pour le corpus *diabète et alimentation*.

[4] `http://www.ncbi.nlm.nih.gov/pubmed/`.

[5] Le problème de l'équilibre des corpus comparables est présenté dans le chapitre 6, notamment pour souligner que c'était en réalité une précaution inutile dans notre cas.

2.2.2 Corpus « cancer du sein »

Nous avons construit un corpus comparable spécialisé français-anglais à partir du portail d'articles scientifiques *Elsevier*[6]. Les documents collectés relèvent du domaine médical et concernent la thématique *cancer du sein*. Les documents ont été collectés en utilisant l'interface de recherche du portail pour sélectionner les publications scientifiques comportant dans le titre les mots clés `cancer du sein` en français et `breast cancer` en anglais, pour la période 2001-2008.

Nous avons obtenu 130 documents pour la partie française et 1 640 pour la partie anglaise. La partie anglaise étant plus volumineuse que la partie française, elle a été découpée aléatoirement en 14 partitions de tailles comparables à celle du sous-corpus français. Nous n'utilisons toutefois qu'une seule partie du corpus anglais-français dans le cadre de l'alignement. L'utilisation de l'ensemble du corpus, dans le but d'étudier l'extraction lexicale bilingue à partir de corpus déséquilibrés est présentée dans Morin (2009) et exploitée dans le chapitre 6.

Les documents ont été convertis en texte brut, codés en UTF8 et nettoyés. Le tableau 2.2 résume différentes informations sur ce corpus.

	Français	Anglais
Nombre de mots	530 000	$14 \times 530\,000$
Nombre de documents	130	1 640

Table 2.2 – Taille de chaque partie pour le corpus *cancer du sein*.

2.2.3 Dictionnaires bilingues

Nous avons également besoin de dictionnaires bilingues[7] pour l'étape de traduction de la méthode directe. Le dictionnaire français-japonais est composé de quatre dictionnaires disponibles gratuitement sur Internet[8] ainsi que du *Dictionnaire scientifique français-japonais* (Hakusuisha, 1989). Il contient 173 156 entrées, dont 114 461 sont des termes simples (composé d'un seul mot), avec une moyenne de 2,1 traductions par entrée.

Pour l'anglais-japonais, nous avons utilisé le dictionnaire *JMDict*[9] qui est disponible librement sous une licence *Creative-Commons* (Attribution-ShareAlike). Nous l'avons complété de traductions de termes techniques issus de différents domaines, composés d'une liste compilée par le Ministère de l'Éducation japonais et le *National Institute of Informatics*[10] ainsi que du *Dictionary of Technical Term* (Kotani et Kori, 1990). Il contient 589 956 entrées avec une moyenne de 2,3 traductions par entrée dont 49 208 termes simples.

Enfin, nous disposons d'un dictionnaire français-anglais, extrait à partir de diverses ressources disponibles sur Internet. Il contient 541 600 entrées avec en moyenne 1,6 traduction par entrée.

[6] `http://www.elsevier.com`.

[7] Il s'agit en réalité d'un lexique bilingue, qui à un mot associe une traduction. Certains mots ont plusieurs entrées dans ces lexiques, correspondant à plusieurs traductions.

[8] `http://kanji.free.fr`; `http://quebec-japon.com/lexique/index.php?a=index&d=25`; `http://dico.fj.free.fr/index.php`; `http://quebec-japon.com/lexique/index.php?a=index&d=3`

[9] `http://www.csse.monash.edu.au/~jwb/j_jmdict.html`.

[10] `http://sciterm.nii.ac.jp/cgi-bin/reference.cgi`.

2.2.4 Listes de références

Pour évaluer la qualité des résultats de l'alignement, nous avons constitué des listes de références, contenant des ensembles de paires de traductions.

2.2.4.1 Corpus « cancer du sein »

Dans le cas du corpus *cancer du sein*, nous utilisons une liste de référence de 122 termes, utilisée notamment dans Morin (2009) et que nous appelons [En-Fr-122]. Cette liste a été construite à partir du méta-thesaurus UMLS[11] et du *Grand dictionnaire terminologique*[12] en ne conservant que les termes simples apparaissant au moins cinq fois dans les parties anglaise et française du corpus. Nous utilisons également une liste conçue de façon semi-automatique, appelée [En-Fr-648] : nous avons extrait l'ensemble des mots du corpus anglais dont la fréquence est supérieure à 15. Nous avons croisé la liste des traductions (obtenues grâce à *Google Translate*[13]) avec l'ensemble des mots de fréquence supérieure à 15 dans le corpus français pour obtenir une liste de 648 mots anglais, avec une ou plusieurs traductions à chaque fois. Il est à noter que certains mots de cette liste, bien que correctement traduits, ne sont pas forcément représentatifs de la terminologie du corpus. Ainsi, cette liste contient des mots tels que *avancer*, *alternative* ou *exprimer*. Elle a l'avantage d'être beaucoup plus volumineuse que la liste [En-Fr-122], ce qui nous permettra de tirer des conclusions plus générales sur la qualité de l'alignement anglais-français sur ce corpus. Cette liste offre aussi une plus grande variété dans la fréquence des mots : certains sont relativement rares (entre 15 et 20 occurrences dans le corpus – par exemple *chose*) alors que d'autres sont très fréquents (plus de 500 occurrences, par exemple *cancer*, qui apparaît 3 235 fois dans la partie française du corpus). Environ 18 % des éléments de la liste [En-Fr-122] sont présents dans la liste [En-Fr-648].

2.2.4.2 Corpus « diabète et alimentation »

Pour le corpus *diabète et alimentation*, nous avons constitué deux listes, une pour l'alignement français-japonais et une pour l'anglais-japonais. La liste pour le français-japonais est constituée de 98 termes sélectionnés (liste [Fr-Jp-98]), la liste pour l'anglais-japonais contient 99 termes ([En-Jp-99]). Elles ont été construites à partir des mêmes ressources que la liste [En-Fr-122] présentée précédemment.

Nous n'avons pas constitué de liste plus générique pour l'alignement vers le japonais. Nous le montrons au chapitre 5 : les résultats de ces alignements sont très différents de ceux obtenus pour l'alignement anglais-français. Des listes construites automatiquement, telle que la liste [En-Fr-648], donnent dans le cas de l'alignement vers le japonais des résultats très dégradés et peu exploitables. Cela s'explique notamment par la difficulté à trouver des ressources linguistiques fiables et couvrants un vocabulaire spécialisé. Cela explique également pourquoi les listes [Fr-Jp-98] et [En-Jp-99] sont plus réduites que la liste [En-Fr-122].

La section suivante traite du phénomène des translittérations, que nous retrouvons dans le corpus *diabète et alimentation*.

[11] `www.nlm.nih.gov/research/umls`
[12] `www.granddictionnaire.com`
[13] `http://translate.google.com`

2.3 Étude des translittérations

Cette section concerne les translittérations, c'est-à-dire des mots empruntés par une langue cible à une langue source, ayant subi des transformations graphiques et phonétiques. Nous nous concentrerons sur les caractéristiques des translittérations, en particulier en langue japonaise pour présenter une étude de cas sur un corpus comparable trilingue anglais-français-japonais. Ces préliminaires nous permettent d'introduire les concepts nécessaires à la compréhension des sections suivantes où nous montrons comment nous exploitons les translittérations dans le cas de l'extraction lexicale à partir de corpus comparables.

Nous appelons *translittération* le phénomène d'emprunt d'un mot d'une langue source à destination d'une langue cible, ne partageant pas nécessairement les mêmes phonèmes, ni les mêmes symboles d'écriture. Sherif et Kondrak (2007) en donnent une définition dont nous proposons une traduction[14] :

> « *Les* translittérations *sont des mots qui sont convertis d'un script à un autre sur la base de leur prononciation plutôt que des transcriptions basées sur leur sens.* »

De ce point de vue, les translittérations sont donc des cas particuliers d'emprunts lexicaux. Ces définitions ne couvrent toutefois pas le cas où des mots sont empruntés et adaptés sans nécessairement conserver leur sens initial. Les exemples sont nombreux dans toutes les langues, citons le cas de *taliban*, signifiant *étudiants en religion* en pachtoun (langue officielle afghane), mais ayant un sens bien différent en français comme en anglais.

Ainsi, la relation de translittération est fréquente entre deux langues possédant des systèmes d'écriture ne partageant aucun caractère en commun, comme c'est le cas entre des langues utilisant l'alphabet latin et d'autres langues utilisant un système d'écriture distinct, comme le chinois ou l'arabe. Le mot emprunté est souvent radicalement modifié pour respecter les contraintes de la langue cible. C'est notamment le cas avec la langue arabe littéraire, qui, à l'écrit, ne précise généralement pas les voyelles. Les contraintes sont différentes avec le chinois et le japonais qui, bien qu'étant des langues très différentes, d'un point de vue grammatical et phonétique, n'autorisent pas de consonne autonome (à quelques exceptions près ; les autres consonnes seront toujours suivies d'une voyelle). La translittération d'un mot vers le chinois ou le japonais introduira donc fréquemment des voyelles entre les consonnes saillantes, et est susceptible de supprimer les consonnes finales. À titre d'exemple, Yoon *et al.* (2007) présentent la translittération de *Bagdad*, adapté en 巴格达 (`ba-ge-da`) en chinois (une voyelle a été insérée et le `d` final supprimé).

Le phénomène de translittération peut donc être vu comme *la projection d'un mot d'une langue source vers une langue cible*. Le processus de projection est un processus avec pertes, certaines informations pouvant être omises ou transformées pour se conformer au cadre graphique et phonétique de la langue cible (Knight et Graehl, 1997).

Les translittérations sont fréquentes dans de nombreuses langues, principalement pour exprimer des concepts récents qui n'ont pas encore de vocabulaire consacré ou qui sont difficiles à transposer en langue cible. Elles sont nombreuses dans le vocabulaire technique, qui est dans un premier temps employé par des experts avant d'être, le cas échéant, assimilé dans le langage courant. Dans certains cas, les translittérations sont utilisées alors même qu'un équivalent peut facilement être trouvé dans la langue cible, comme par exemple en japonais, ライス (`ra-i-su` – *rice*) pour une présentation de riz, alors que plusieurs kanjis sont disponibles (米, *kome*, le riz ou 飯, *meshi*, riz préparé, synonyme de repas).

Les translittérations sont employées pour les mots difficiles à traduire, comme c'est souvent le cas des noms propres. On écrira *Tokyo* en français (ou *Tokio* en allemand) car la graphie originale, 東京

[14] « *Transliterations are words that are converted from one writing script to another on the basis of their pronunciation, rather than being translated on the basis of their meaning.* »

(phonétiquement `to-o-kyo-o`), ne peut pas être employée naturellement dans ces langues. Notons que la traduction des noms propres est une question épineuse pour le traducteur humain qui dispose de plusieurs choix (Humbley, 2006). Il peut se contenter de l'emprunter, comme *Washington* qui sera repris à l'identique en français, mais aussi de le traduire lorsque le nom propre le permet, par exemple : *Organisation des Nations Unis* / *United Nations* / 国際連合 (lit. 国際 - international ; 連合 - union, association) ou *Nouvelle Orléans* (*New Orleans*). Les noms propres peuvent également être modifiés, comme dans l'exemple de *Nova Zeelandia* (néerlandais, nom d'origine), adapté en *New Zealand* en anglais et *Nouvelle-Zélande* en français. Enfin, le traducteur peut être amené à translittérer le nom propre lorsque la traduction ou l'emprunt direct ne sont pas pertinents (Agafonov *et al.*, 2006). Par exemple フランス / `fu-ra-n-su`, *France*. Par la suite nous nous intéresserons aux translittérations dans la langue japonaise et aux relations avec le français et l'anglais.

2.4 Translittérations en japonais

Les translittérations sont un phénomène très fréquent en japonais, nous le montrons par la suite en étudiant le corpus *diabète et alimentation* ; dans un premier temps nous présentons quelques caractéristiques propres à cette langue.

2.4.1 Caractéristiques de la langue japonaise

La langue japonaise moderne dispose de trois ensembles de symboles :

- **Les kanjis**, à l'origine des symboles chinois, portent un ou plusieurs sens. Ils sont utilisés seuls ou en combinaison avec d'autres kanjis pour former les mots courants de la langue japonaise. Ils possèdent généralement plusieurs prononciations différentes en fonction du contexte. Ainsi 寺, signifiant *temple* se prononcera `(o)-tera` lorsqu'employé de façon isolée (la particule `o` étant optionnelle et honorifique), sera sonorisé en `-dera` dans 山寺 (`yama-dera`, *temple de montagne*) et prononcé `ji`, à la chinoise, dans 仏寺/`bu-tsu-ji`, *temple bouddhique*.
- **Les hiraganas** sont un syllabaire phonétique (voir table 2.3) et ont un rôle grammatical. Ils indiquent les déclinaisons des verbes et les particules grammaticales. Ils servent aussi pour certains mots lorsqu'aucun kanji n'est disponible (typiquement, pour les mots fonctionnels) ou lorsque les kanjis sont inconnus du scripteur ou potentiellement inconnus du lecteur (cas des livres pour enfants).
- **Les katakanas** sont un autre syllabaire phonétique, équivalent au syllabaire hiragana. Ils servent principalement pour les mots d'emprunts (ce qui nous donne un précieux indice pour les repérer) mais aussi pour indiquer l'emphase, notamment dans les communications publicitaires et les onomatopées (les onomatopées émises par des voix humaines étant écrites en hiragana). La figure 2.1 présente quelques exemples d'utilisations des hiraganas et des katakanas dans un *manga*.

La table 2.3 recense l'ensemble des symboles hiraganas et katana utilisés dans le japonais moderne. Le japonais n'autorise pas de séquence de consonnes (à l'exception du *n*, représentant une more consonantique autonome). Ainsi le `dri` de *Sandrine* est retranscrit par `do-ri`. Par extension, il est impossible d'avoir une consonne isolée en fin de mot, comme le `d` final de `David` (toujours à l'exception du `n`). Dans ce cas, soit une voyelle sera ajoutée, soit la consonne sera élidée, comme dans `guitar/guitare` (en-fr), transcrit en キター (`gi-ta-a`). Enfin, les accents toniques seront généralement conservés et la voyelle concernée sera doublée dans la retranscription en japonais, comme dans `accordéon`, retranscrit en アコーデオン / (`a-ko-o- dé-o-n`). Une dernière more, n'apparaissant pas dans la table permet d'introduire un silence dans la prononciation d'un mot, le petit « tsu » (っ / ッ) retranscrit en alphabet latin

Figure 2.1 – Un exemple du manga *GTO*. Ici le personnage enrage はあ～～～, `ha-a...` (*hiraganas*) ; les essuie-glaces produisent le son ユッコ ユッコ, `yu-kko yu-kko` (katakana) (il y a d'autres onomatopés tel que le tonnerre ピカツ (pi-ka-tsu), le son des gouttes de pluies sur le pare-brise ビチ (bi-chi)...)

par le doublement de la consonne suivante, par exemple dans スタッフ, `su-ta-ffu`, *staff*, permettant là aussi de marquer l'accent tonique dans le cas des translittérations.

	1	2	3	4	5	6	7	8
1	あ/ア/a	い/イ/i	う/ウ/u	え/エ/e	お/オ/o			
2	か/カ/ka/	き/キ/ki	く/ク/ku	け/ケ/ke	こ/コ/ko	きゃ/キャ/kya	きゅ/キュ/kyu	きょ/キョ/kyo
3	さ/サ/sa	し/シ/shi	す/ス/su	せ/セ/se	そ/ソ/so	しゃ/シャ/sha	しゅ/シュ/shu	しょ/ショ/sho
4	た/タ/ta	ち/チ/chi	つ/ツ/tsu	て/テ/te	と/ト/to	ちゃ/チャ/cha	ちゅ/チュ/chu	ちょ/チョ/cho
5	な/ナ/na	に/ニ/ni	ぬ/ヌ/nu	ね/ネ/ne	の/ノ/no	にゃ/ニャ/nya	にゅ/ニュ/nyu	にょ/ニョ/nyo
6	は/ハ/ha	ひ/ヒ/hi	ふ/フ/fu	へ/ヘ/he	ほ/ホ/ho	ひゃ/ヒャ/hya	ひゅ/ヒュ/hyu	ひょ/ヒョ/hyo
7	ま/マ/ma	み/ミ/mi	む/ム/mu	め/メ/me	も/モ/mo	みゃ/ミャ/mya	みゅ/ミュ/myu	みょ/ミョ/myo
8	ら/ラ/ra	り/リ/ri	る/ル/ru	れ/レ/re	ろ/ロ/ro	りゃ/リャ/rya	りゅ/リュ/ryu	りょ/リョ/ryo
9	や/ヤ/ya		ゆ/ユ/yu		よ/ヨ/yo			
10	わ/ワ/wa				を/ヲ/wo			
11	が/ガ/ga	ぎ/ギ/gi	ぐ/グ/gu	げ/ゲ/ge	ご/ゴ/go	ぎゃ/ギャ/gya	ぎゅ/ギュ/gyu	ぎょ/ギョ/gyo
12	ざ/ザ/za	じ/ジ/ji	ず/ズ/zu	ぜ/ゼ/ze	ぞ/ゾ/zo	じゃ/ジャ/ja	じゅ/ジュ/ju	じょ/ジョ/jo
13	だ/ダ/da			で/デ/de	ど/ド/do			
14	ば/バ/ba	び/ビ/bi	ぶ/ブ/bu	べ/ベ/be	ぼ/ボ/bo	びゃ/ビャ/bya	びゅ/ビュ/byu	びょ/ビョ/byo
15	ぱ/パ/pa	び/ピ/pi	ぷ/プ/pu	ぺ/ペ/pe	ぼ/ポ/po	びゃ/ピャ/pya	びゅ/ピュ/pyu	びょ/ピョ/pyo
16	ん/ン/n							

Table 2.3 – Les hiraganas et katakanas de base et leur prononciation. À ces symboles peuvent s'ajouter ceux associées au `v` et au `f`, introduits tardivement dans le but de faciliter les translittérations.

2.4.2 Place des translittérations en japonais

Les études existantes sur les mots empruntés et translittérés révèlent quelques traits quantitatifs et qualitatifs. Quantitativement, le volume des translittérations tend à augmenter lorsque le volume de la

terminologie augmente (Kageura, 2003). L'augmentation de la présence de mots d'emprunt est aussi clairement observée dans les textes ; Ito (2007), comparant le nombre de mots par type d'origine dans 90 magazines publiés en 1956 et dans 70 magazines publiés en 1994, montre que la proportion d'éléments translittérés passe de 2,7 % à 8 %. Au niveau caractère, cette proportion passe de 9,2 % à 24,7 %. Ceci montre que les translittérations constituent une part importante du vocabulaire japonais. Nous confirmons cette observation sur notre corpus comparable trilingue en section 2.5.

Par ailleurs, les translittérations tendent à être utilisées dans des contextes spécifiques. C'est en particulier vrai dans les domaines spécialisés. Ainsi, *small world*[15] sera translitteré en スモールワールド / su-mo-o-ru-wa-a-ru-do alors même que 小さな世界 (lit. 小さな, petit – notons l'utilisation des hiraganas さい pour décliner l'adjectif –, 世界, monde) est une traduction correcte. De même, il arrive fréquemment que les noms propres soient utilisés pour désigner des concepts, tel que le *pivot de Gauss*, le *théorème de Pythagore*, les *automates de Markov* (Bodenreider et Zweigenbaum, 2000). Dans ces cas, l'emploi de translittérations est fréquent pour l'entité nommée.

2.4.3 Typologie des translittérations japonaises

La langue japonaise emprunte beaucoup de vocabulaire aux langues occidentales (anglais, français, allemand...) mais aussi aux langues orientales, en particulier le chinois et le coréen. Nous nous concentrerons ici uniquement sur les emprunts aux langues occidentales. Nous distinguons différents types de translittérations dans la langue japonaise en fonction du terme source et de son adaptation :

- de terme simple vers terme simple, translittération complète.
  Exemple : *aluminium* → アルミニウム / a-ru-mi-ni-u-mu,
  computer → コンピュータ/co-n-pyu-u-ta ;
- de terme complexe vers terme complexe, translittération complète.
  Exemple : *front glass* →フロント・ガラス /fu-ro-n-to ga-la-su ;
- terme complexe vers terme simple, translittération complète.
  Exemple : *christmas tree* → クリスマスツリー / ku-ri-su-ma-su-tsu-ri-i ;
- terme simple ou complexe vers terme simple ou complexe, translittération partielle.
  Exemple : *costume play* → コスプレ / ko-su-pu-re (cosplay),
  personal computer → パソコン/pa-so-ko-n ;

La plupart des emprunts aux langues occidentales proviennent de l'anglais (même certains noms propres sont empruntés à leur adaptation en anglais, comme スペイン/su-pe-i-n (*Spain*) – España). Quelques translittérations sont toutefois empruntées au français, citons クロワサン/ku-ro-wa-sa-n – croissant ou エスカルゴ/e-su-ka-ru-go – escargot ; ceux cuisinés, que l'on mange, le nom de l'animal étant 蝸牛/カタツムリ / ka-ta-tsu-mu-ri. Ce dernier exemple montre que les noms d'animaux sont souvent écrits en katakana même lorsque des kanjis sont disponibles. Certains termes sont empruntés à l'allemand, comme レントゲン/re-n-to-ge-n, pour rayon-X, du nom de leur découvreur, Wilhelm Röntgen.

2.4.4 Relation avec la langue française

Bien que les emprunts à la langue française soient très spécifiques, nous constatons qu'une grande quantité de translittérations du japonais peut en réalité être reliée à un terme français, car l'anglais et le français partagent beaucoup de vocabulaire commun. Cela est dû aux nombreux emprunts lexicaux,

[15] Terme utilisé pour parler de réseaux où chaque nœud peut-être atteint en peu de sauts.

mais aussi aux relations de cognats[16] entre ces langues. Les cognats sont des mots issus d'une langue commune mais ayant évolué parallèlement dans des langues différentes. Ainsi, l'origine des translittérations en japonais importe peu et est difficile à déterminer par l'observation, car le mot emprunté est souvent lui même un mot dérivé d'une langue plus ancienne telle que le grec ou le latin. Nous ne nous intéresserons donc pas à l'étymologie des translittérations, mais chercherons uniquement à savoir si une translittération peut être associée à un terme français ou anglais.

La table 2.4 donne quelques exemples de translittérations correspondant aussi bien à un terme français qu'à un terme anglais.

japonais / romanisation	anglais	français
インスリン / `i-n-su-ri-n`	insulin	insuline
ホルモン / `ho-ru-mo-n`	hormone	hormone
ミネラル / `mi-ne-ra-ru`	mineral	minéral
ヘモグロビン / `he-mo-gu-ro-bi-n`	hemoglobin	hémoglobine
ビタミン / `bi-ta-mi-n`	vitamin	vitamine
カルシウム / `ka-ru-shi-u-mu`	calcium	calcium

Table 2.4 – Exemple de relations de translittérations indirectes entre le français et le japonais.

Nous arrêterons donc ici une définition de la relation de translittération que nous utiliserons par la suite, en accord avec les discussions de la section 2.3 et les observations sur la langue japonaise.

Définition 2.4.4.1. *Deux mots w_1 et w_2 de langue l_1 et l_2 sont en relation de translittération s'ils sont traductions l'un de l'autre, s'ils sont phonétiquement proches et dans des systèmes d'écriture différents.*

Cette définition permet également de mettre de côté les *faux-amis*, qui ne sont que phonétiquement proches. Citons フィルム/`fi-ru-mu`, translittérations de *film*. Alors qu'en français, *film* signifie principalement une œuvre audiovisuelle, le sens commun en anglais (c'est aussi le sens de フィルム) est celui de *pellicule*. Ainsi, même si ces sens sont liés et que l'on peut aussi appeler une pellicule un *film* en français, nous n'admettrons pas la relation de translittérations entre フィルム et *film* en français.

Dans l'optique de l'alignement multilingue à partir de corpus comparables, nous pouvons déjà proposer quelques hypothèses suite aux observations présentées dans cette partie :

1. Quantitativement, les translittérations constituent une part non négligeable du vocabulaire japonais ;
2. Qualitativement, les translittérations portent des concepts spécifiques, puisqu'ils recouvrent une partie du vocabulaire technique en japonais.

D'un point de vue applicatif, nous devons aussi noter que les dictionnaires existants ne peuvent pas couvrir efficacement le vocabulaire translittéré car il évolue très rapidement, ce qui est le cas des noms propres et des termes techniques.

Dans les sections suivantes, nous testons les hypothèses proposées ici sur un corpus trilingue anglais-français-japonais, puis faisons le point sur les méthodes utilisées dans la littérature pour détecter automatiquement les translittérations.

[16] Du latin *cognatus*, « parent par les liens du sang », signifiant *personne liée à une autre par un lien de parenté naturelle sans distinction de lignes*, d'après le *Trésor de la Langue Française*.

2.5 Observation des translittérations dans le corpus « diabète et alimentation »

Cette partie présente une étude de cas à partir du corpus comparable trilingue « *diabète et alimentation* » présentée en section 2.2.1. Nous y observons la place des translittérations et les relations qu'elles offrent entre les différentes langues. Notre première observation porte sur la place des translittérations dans le corpus japonais puis sur les liens de translittérations avec les corpus français et anglais pour mettre en évidence l'importance de ce phénomène dans les corpus spécialisés.

2.5.1 À partir du corpus japonais

Dans un premier temps, nous avons extrait toutes les séquences maximales de katakanas à partir du corpus japonais (quelle que soit la taille). Ces séquences sont pour la plupart des translittérations, en accord avec ce que nous avons présenté précédemment. Nous obtenons alors 493 séquences valables, c'est-à-dire des termes qui existent effectivement dans la langue japonaise (les termes extraits ont été validés par des Japonais) en katakanas dans le corpus. Une partie des séquences de katakanas a été écartée car résultant d'erreurs de segmentation des mots introduites par la conversion des documents en texte. Les candidats conservés représentent environ 8 % du vocabulaire total unique du corpus. Nous avons ensuite traduit manuellement ces candidats en anglais, mais aussi en français lorsque cela était possible.

Nous avons classé les candidats extraits en plusieurs catégories. Certains ne pouvaient être alignés qu'avec l'anglais, ou qu'avec le français, alors que d'autres pouvaient être alignés avec un terme appartenant aux deux langues. D'autres étaient des translittérations, mais ne s'alignant ni avec un mot anglais, ni avec un mot français. Certains candidats ne correspondaient pas à des translittérations (typiquement les noms d'animaux, mais aussi l'exception notable de タンパク/`ta-n-pa-ku` signifiant *protéine*). Enfin, certaines translittérations sont des adaptations et ne correspondent pas phonétiquement au terme source, par exemple コンビニ/`ko-n-bi-ni`, adaptation de *convenient store*. Le nombre d'éléments de chacune de ces catégories est présenté en table 2.5.

	#occ	ratio	Exemples
fran. seul.	4	0.8 %	リール/`ri-i-ru`/*Lille*
ang. seul.	228	46 %	ヘルス/`he-ru-su`/*health*, ダイエット/`da-i-e-tto`/*diet*
fran./ang.	221	45 %	ヒスタミン/`hi-su-ta-mi-n`/*histamine*
adapté	12	2 %	ビル/`bi-ru`/*building*, テレビ/`te-re-bi`/*television*
non fran./ang.	5	1 %	カリウム/`ka-ri-wa-mu`/*potassium*, タイ/`ta-i`/*Thailand*
non trans.	23	5 %	ムカデ/`mu-ka-de`/*mille-pattes*, カキ/`ka-ki`/*huître*
TOTAL	**493**	**100** %	

Table 2.5 – Statistiques concernant les séquences de katakanas dans le corpus japonais.

2.5.2 Relations avec les corpus anglais et français

Il est intéressant de vérifier s'il est possible de trouver les correspondances des translittérations japonaises mises en évidence dans la section précédente dans les parties française et anglaise. Dans le corpus japonais, nous avons trouvé 225 translittérations pouvant s'aligner avec un terme français (voir table 2.5

– 221 translittérations pour français-anglais et 4 pour le français seulement). Cela signifie que nous pourrons, au mieux, trouver 225 relations de translittération entre les corpus japonais et français. De même pour l'anglais, nous avons trouvé 449 translittérations japonaises susceptibles de s'aligner avec un terme anglais. Nous trouvons finalement 140 relations de translittérations entre les corpus japonais et français, et 314 entre les corpus japonais et anglais. La table 2.6 résume ces résultats, nous y précisons également le nombre de relations impliquant un hapax.

	Max.	Trouvé	Ratio	Dont hapax
Fr-Jp	225	**140**	62 %	16
En-Jp	449	**314**	70 %	26

Table 2.6 – Relations de translittérations entre les corpus.

Ces résultats sont intéressants car ils montrent que, non seulement les translittérations dans la langue japonaise ne sont pas un phénomène marginal, mais aussi qu'elles offrent des liens entre nos différents corpus puisque nous retrouvons la majorité de leurs correspondances dans les corpus anglais et français. Ces résultats indiquent également que les translittérations couvrent une partie du vocabulaire partagé entre corpus, vocabulaire que nous chercherons à extraire par la suite.

2.6 Détection automatique des translittérations

De nombreux efforts ont été réalisés pour aligner automatiquement les translittérations et il nous paraît intéressant de décrire les difficultés rencontrées et les méthodes déployées. Citons notamment Yaser et Knight (2002) et Sherif et Kondrak (2007), qui ont travaillé sur l'alignement arabe-anglais, Tao *et al.* (2006) qui ont travaillé sur l'arabe, le chinois et l'anglais, Sproat *et al.* (2006), Virga et Khudanpur (2003) et Gao *et al.* (2004) sur les entités nommées en chinois et Knight et Graehl (1997) ; Nakamura-Delloye (2007) et Tsuji *et al.* (2002) sur les translittérations en langue japonaise.

2.6.1 Enjeux de la détection de translittérations

Il peut paraître aisé, au premier abord, de relier un mot translitteré au mot source, à partir duquel il a été dérivé. Toutefois nous l'avons vu, le processus de translittération est un processus avec pertes et le rapprochement entre une translittération et sa source n'est pas toujours évident. Sproat *et al.* (2006) donnent l'exemple, en chinois, de 威廉姆斯 (`wei-lian-mu-si`), translittérations de *Williams*, délicate à traduire, même pour un humain. La détection et l'alignement automatique de translittérations est alors une assistance précieuse pour le traducteur et le lexicographe.

C'est une tâche d'autant plus difficile dans le cadre d'un processus automatique en raison des ambiguïtés phonétiques que l'on rencontre dans les langues sources. Par exemple en français, la séquence `ch` possède deux prononciations distinctes, dans *chorale* ou dans *château*, c'est aussi le cas en anglais ou *live* aura une prononciation différente s'il s'agit du verbe, prononcé [lɪv] ou de l'adjectif [laɪv]. Nous présentons par la suite une approche proposée par Knight et Graehl (1997), s'appuyant sur une série de modèles probabilistes.

2.6.2 Exemple d'approche

Les translittérations étant *des transcriptions basées sur la prononciation plutôt que sur leur sens*, la plupart des approches s'appuie sur des connaissances phonétiques. Nous présentons en détail l'exemple

de Knight et Graehl (1997) car ils introduisent un premier processus statistique complet, repris et amélioré par la suite. Ils proposent dans un premier temps un processus d'apprentissage pour produire des translittérations japonaises à partir d'un mot anglais :

- le mot est phonétisé à l'aide d'un automate ;
- la prononciation est modifiée pour s'adapter aux phonèmes du japonais ;
- la prononciation modifiée est retranscrite en katakanas ;
- les katakanas sont écrits.

Ces cinq problèmes sont modélisés à l'aide du théorème de Bayes, cinq distributions de probabilités sont définies. Le but est d'être capable de reprendre ce processus en sens inverse (c'est-à-dire entre autre d'appliquer un traitement OCR[17] pour remonter des katakanas écrits vers la séquence originale en anglais). Les modèles sont définis ainsi :

1. $P(w)$ – génère les séquences de mots en anglais ;
2. $P(e|w)$ – prononce la séquence anglaise ;
3. $P(j|e)$ – convertit le son anglais en son japonais ;
4. $P(k|j)$ – convertit le son japonais en katakanas ;
5. $P(o|k)$ – introduit les erreurs causées par l'OCR.

Ainsi, partant d'une chaîne o en katakanas, il s'agit de maximiser la somme sur e, j et k :

$$\sum_{e,j,k} P(w).P(e|w).P(j|e).P(k|j).P(o|k) \tag{2.1}$$

2.6.2.1 Modèles probabilistes

Pour reprendre à rebours le processus d'apprentissage dans le but d'obtenir le mot source à partir d'une translittération en katakanas, plusieurs modèles probabilistes sont définis.

- **5. Pour les séquences de mots anglais**, l'objectif est qu'une séquence de caractères valable obtienne un meilleur score qu'une séquence erronée. Par exemple, *ice cream* doit avoir une meilleure probabilité que *ice creme*, qui doit avoir une meilleure probabilité que *aice kreem*. Dans leur étude, ce modèle s'appuie sur les unigrammes : 262 000 fréquences obtenues du *Wall Street Journal* ainsi que sur un corpus de noms et de lieux en anglais.
- **4. Pour la traduction anglais écrit vers anglais phonétisé**, les auteurs construisent un automate à états finis en utilisant le CMU[18], qui décrit la prononciation de 125 000 mots (Anglais nord-américain – 110 000 mots à la rédaction de l'article) à partir de 39 phonèmes auxquels ils adjoignent un symbole spécial pour les pauses entre les mots. Ce transducteur prend donc en entrée le mot écrit et propose différentes prononciations en sortie.
- **3. Pour la traduction phonétique anglaise vers phonétique japonaise**, ils tentent d'aligner les phonèmes anglais avec des phonèmes japonais, processus qui perd nécessairement une partie de l'information (par ex., il n'y pas de de différence entre le `R` et le `L` en japonais). Se pose donc la difficulté de savoir quel ensemble de symboles cibles pourra correspondre à l'ensemble de symboles représentant le terme phonétisé source, exemple : `P R OW PAUSE S AA K ER` pourra s'aligner avec `p u r o pause s a kk a a` en japonais. Notons que la séquence en japonais est volontairement plus longue – la séquence `a a` étant délibérément éclatée – correspondant

[17] *Optical Character Recognition*, Reconnaissance Optique de l'Écriture, voir notamment Crettez et Lorette (1998).
[18] Le *Carnegie Mellon University Pronouncing Dictionary*.

à la réalité, puisque généralement les translittérations japonaises sont plus longues. L'alignement se fait en 6 étapes :

1. pour chaque paire anglais-japonais, calculer tous les alignements possibles entre leurs éléments (`L OW` et `r o o` pourront s'aligner de deux façons différentes : `L-r / OW-oo` ou `L-r o / OW-o`) ;
2. répartir équitablement les probabilités de chaque alignement (somme des probabilités égale à 1) ;
3. pour chacun des phonèmes anglais, compter les instances de leurs alignements et leur attribuer un score en fonction de leur fréquence ;
4. recalculer les scores d'alignement : le produit des scores d'alignement des phonèmes ;
5. normaliser les scores d'alignement (somme pour chaque paire égale à 1) ;
6. répéter 3–6 jusqu'à ce que les probabilités d'alignement des symboles convergent.

– **2. Pour la traduction japonais phonétisée vers katakanas**, deux automates sont construits manuellement avec l'aide de recommandations linguistiques sur le japonais. Le premier automate regroupe les séquences de phonèmes construites précédemment, le second traduit en katakana en essayant d'absorber un maximum de symboles avant de produire un katakana. Ce deuxième automate intègre des règles probabilistes pour, par exemple, représenter une suite de voyelles : une voyelle appuyée (doublée) en katakanas sera généralement suivie du symbole ー, mais sera parfois simplement répétée (`oo` pourra donc être retranscrit en オー ou en オオ).
– **1. Traduction katakana-OCR**. Les auteurs impriment des pages de katakanas sur lesquelles ils appliquent le processus de reconnaissance de l'écriture, pour analyser les erreurs produites par le système. Ils obtiennent donc, pour chaque caractère d'origine k et chaque caractère reconnu o la probabilité $P(o|k)$. Certains caractères sont parfaitement reconnus alors que d'autres sont très ambigus (ピ/`bi` est reconnu comme ピ/`pi` dans près de 40 % des cas).

2.6.2.2 Exemple

Munis de ces différents modèles probabilistes, conçus pour chaque étape de l'apprentissage, les auteurs remontent chaque étape à partir d'une chaîne de caractères imprimée. La séquence reconnue par l'OCR est マスクーズトーチメント (`ma-su-ku-u-zu-to chi-me-n-to`). Il y a deux erreurs de reconnaissance ici : le ク/`ku` est en fait un タ/`ta`, et le チ/`chi` est en fait un ナ/`na`. En utilisant les différents modèles à rebours, jusqu'à l'étape 2. (japonais phonétisé) la chaîne la plus probable obtenue est la chaîne `m a s u t a a zu t o o ch i m e n t o`. En remontant encore d'une étape, ils obtiennent la phonétisation anglaise `M AE S T AE AE DH UH T AO AO CH IH M EH N T AO` puis la chaîne en anglais probable (étape 4.) *masters tone am ent awe* et enfin, en dernière étape, la chaîne correcte *masters tournament*.

Cet exemple est d'autant plus intéressant qu'un traducteur humain a beaucoup de difficultés pour retrouver le sens anglais d'origine à partir de la translittération マスターズトーナメント (`ma-su-ta-a-zu-to-o-na-me-n-to`) alors que le système proposé donne un résultat correct.

Les auteurs confirment cette impression en exécutant une série d'expériences opposant traducteurs humains et résultats de leur chaîne de traitement. En utilisant un modèle de langue adapté pour les noms (pour l'étape $P(w)$), ils proposent d'opposer les résultats de leur traitement à une traduction réalisée par un humain. L'expérience consiste à demander à quatre humains (anglophones de naissance) de traduire des noms propres translittérés (dans cette expérience, des personnalités politiques des États-Unis) et de les translittérer automatiquement avec leur algorithme. Les résultats sont présentés table 2.7.

	humain	machine
correct	27 %	64 %
équivalent phonétiquement	7 %	12 %
incorrect	66 %	24 %

Table 2.7 – Comparaison des résultats avec des traducteurs humains.

Ces résultats montrent que 27 % des alignements réalisés par les humains sont corrects alors que la machine atteint 64 % de traductions correctes. 7 % des alignements humains sont *phonétiquement équivalents*, c'est-à-dire presque corrects, mais mal orthographiés (exemple *brian/bryan*) contre 12 % pour les ordinateurs. Enfin, 66 % des traductions réalisées par des humains sont incorrectes alors que seules 24 % sont incorrectes pour l'ordinateur. Ces résultats montrent une nette supériorité de l'algorithme de Knight et Graehl (1997) pour la réalisation de cette tâche mais doivent être modérés : les quatres humains connaissaient bien l'anglais mais peu le japonais, il est probable qu'un Japonais de naissance soit meilleur pour cette tâche, même s'il n'est pas un expert en anglais.

2.6.3 Autres approches

Par la suite, cette première approche a été perfectionnée, notamment en utilisant des outils de génération de la parole avancés tel que *Festival*[19] (Taylor *et al.*, 1998) ou en améliorant les modèles de langage pour la désambiguïsation des résultats (dans l'exemple, l'étape consistant à retrouver la séquence en anglais la plus probable). Enfin, des travaux plus récents s'appuient sur les corpus comparables : comme nous le montrons en partie 2.5 sur une étude de cas, les translittérations représentent une part importante du vocabulaire commun des corpus comparables. Ainsi, il est possible d'opposer les différents candidats obtenus pour une translittération appartenant à un corpus source au vocabulaire du corpus cible (Sproat *et al.*, 2006 ; Tao *et al.*, 2006).

D'autres approches cherchent, à partir de couples de chaînes de caractères translittérées déjà alignés, à entraîner une mesure de distance entre paire de candidats. Il est possible par exemple, d'apprendre les coûts d'édition (le nombre et le coût des transformations à appliquer à une chaîne pour obtenir l'autre, voir exemple 2.8) en généralisant la distance de Levenshtein (Levenshtein, 1966). Le coût des transformations fréquentes sera faible, par exemple, entre la syllabe française *ca* et la more japonaise カ–ka. Le coût des transformations rares ou inexistantes sera élevé (par exemple entre la syllabe *ca* et la more ス–su).

Transformation	P	O	U	L	E	T	**Coût**
suppression	P	O	U	L	E		1
substitution	P	L	U	L	E		1
substitution	P	L	U	M	E		1
Total	P	L	U	M	E		**3**

Table 2.8 – Coût d'édition entre les mots *poulet* et *plume*.

Il est ensuite possible d'utiliser ces mesures pour évaluer la similarité entre deux candidats à la translittération. Sherif et Kondrak (2007) proposent un algorithme d'apprentissage basé sur la distance de Levenshtein généralisée pour aligner des paires de translittérations entre l'arabe et l'anglais. Ces

[19] Réalisé par le *Centre for Speech Technology Research* de l'Université d'Edinburgh `http://www.cstr.ed.ac.uk/projects/festival/`.

approches sont toutefois très sensibles aux ambiguïtés phonétiques décrites précédemment, criantes dans la langue française. Par exemple, les règles utilisées par Tsuji *et al.* (2002) pour l'alignement français-japonais, appliquées sur notre corpus comparable, donnent des résultats peu fiables, des exemples sont présentés en table 2.9, pour la plupart erronés. Cela est dû, d'une part, à des problèmes de segmentation du corpus japonais, réalisée pour le japonais avec *Chasen*[20], retournant des séquences de katakanas courtes et ne correspondant à aucun mot japonais (permettant de nombreuses correspondances avec des mots courts – souvent erronés également – en français), d'autre part, pour les mots plus longs, à des ambiguïtés phonétiques. Ainsi, la syllabe `cha` en français est le plus souvent prononcée [ʃa], comme dans *chapiteau*, mais dans certains rare cas, elle est prononcée [ka], comme dans *charisme*, ce pourquoi le mot *charente* se retrouve associé à la prononciation `ka-re-n-to`.

カレント	`ka-re-n-to`	courant, charente, garantie
エル	`e-ru`	au, are, er, al, ill, ir, hill, ar
ラン	`ra-n`	lent, ligne, lin, rang, lm
エン	`e-n`	an, en, im, am, int, ang, in

Table 2.9 – Alignements proposés pour la détection automatique de translittérations sur le corpus comparable français-japonais.

2.7 Conclusion

Ce chapitre apporte de nouvelles informations sur les motivations de notre travail : nous nous concentrons sur l'exploitation de deux corpus comparables, conçus pour refléter des *langues de spécialité*. Ces langues de spécialité tendent notamment à employer une terminologie particulière, rarement disponible dans des dictionnaires de langue générale. En exploitant ces corpus, nous espérons améliorer l'acquisition automatique de paires de traductions pour cette terminologie, en nous appuyant par exemple sur les translittérations.

Le chapitre suivant se consacre aux méthodes utilisées notamment pour l'acquisition terminologique, dédiées à la *caractérisation sémantique* des termes, dans le but de distinguer leurs différents sens en s'aidant de leurs cooccurrences. Nous montrons comment l'approche directe est reliée aux techniques d'acquisition sémantique d'un point de vue linguistique et technique.

[20] `http://chasen.aist-nara.ac.jp/hiki/ChaSen/`.

CHAPITRE 3

Caractérisation sémantique

Ce chapitre revient sur les hypothèses motivant les approches d'alignement de lexique bilingue à partir de corpus comparables. Nous montrons que ces hypothèses ont été dans un premier temps exploitées dans un cadre monolingue, par exemple pour l'aide à la conception de thesaurus, avant d'être étendues au cadre multilingue.

3.1 Acquisition sémantique

Un même concept[1] peut être exprimé de plusieurs façons, ce qui rend la tâche de *faire comprendre* une information textuelle à un ordinateur délicate. Ce problème de la *variabilité sémantique* (un des « saint Graal » de la linguistique computationnelle, d'après David Evans – préface de Grefenstette, 1994b) se retrouve typiquement dans la recherche d'information où un document peut être omis comme résultat d'une requête car le vocabulaire de la requête ne s'y trouve pas. Une requête contenant *voiture* devrait pouvoir retourner les documents parlant également de *monospace* ou de *berline* : bien que ces mots aient chacun une graphie différente, ils sont tous reliés aux mêmes concepts de *véhicule terrestre motorisé* auquel nous pouvons ajouter par exemple la notion de *véhicule personnel* (pour les distinguer des bus ou des trains – bien que *voiture* renvoie aussi au concept de *wagon de train*, comme dans *voiture bar*) ou de *véhicule muni de roues* (pour les distinguer des chars d'assaut). L'acquisition sémantique, ou *analyse sémantique*, a donc pour but de franchir le pas entre la manipulation de simples chaînes de caractères et l'étude du sens des mots (Grefenstette, 1994b).

Nous présentons rapidement cet axe de recherche, car même si les objectifs d'origine ne sont pas les nôtres, nous nous inspirons largement des hypothèses formulées et de la méthodologie utilisée que nous présentons en section 3.1.2, dans le cadre de l'alignement de lexique bilingue à partir de corpus comparables.

3.1.1 Tâches

Grefenstette (1994b) situe l'apparition du besoin d'analyse sémantique au milieu des années 50, lorsque les premiers travaux en traduction automatique ont constaté que les simples substitutions de

[1] Un concept s'entend par rapport aux mots qui l'expriment. *Fauteuil* décrit un objet, généralement un siège, comportant quatre pieds, un dossier et des accoudoirs. Il se définit en opposition à *chaise* ou *tabouret*, qui servent également à s'assoir mais ont des propriétés différentes (Habert *et al.*, 1997, page 75). Par extension, *fauteuil* s'emploie également dans la métonymie *fauteuil de président – Présider* est lui même emprunté au latin *praesidere*, signifiant *être assis devant*, d'après le *Trésor de la Langue Française*.

mots piochés dans des dictionnaires bilingues étaient inadéquates et que le contexte des mots dans une phrase était de première importance. L'acquisition sémantique *a pour but de constituer ou d'accroître des dictionnaires sémantiques* (Zweigenbaum et Habert, 2006, page 22), c'est-à-dire des ressources linguistiques associant à un mot des mots sémantiquement proches ou reliés (synonymes, antonymes, couples hyponyme-hyperonyme). En d'autres termes, il s'agit de permettre la contextualisation des mots étudiés, généralement des termes, et d'être capable de les classer selon leurs différents sens. Plus largement, il s'agit de répondre au problème de *variabilité sémantique des termes* propre aux langues naturelles. Cette contextualisation est particulièrement prisée par les terminologues pour différencier les termes selon leurs usages et les domaines dans lesquels ils sont employés (L'Homme, 2004). Le mot *arbre* aura un sens bien différent pour un botaniste (c'est une plante), un informaticien (c'est un graphe acyclique) ou un électricien (c'est un réseau électrique). Dans le cadre des systèmes mécaniques, un *arbre* ne représente plus un objet *arborescent*, mais une pièce, relié au moteur qui entraine le reste du mécanisme (par exemple, l'hélice d'un bateau). Dans ce cas, c'est un synonyme d'*axe*, éloigné du concept de la structure de données informatique.

Habert *et al.* (1997) distinguent trois applications principales de l'acquisition sémantique : l'analyse de contenu, l'acquisition de connaissances et la recherche documentaire. L'analyse sémantique permet en effet de rendre compte du *contenu* d'un corpus en analysant son lexique et les relations entretenues par les mots au sein de ce corpus. Elle permet d'extraire les thèmes principaux des ouvrages, ce qui a une utilité directe pour l'organisation et la classification de documents dans le cadre de la recherche documentaire. L'analyse sémantique permet enfin d'acquérir de nouvelles connaissances à partir de corpus dans le but d'aider à la construction de bases de connaissances lexicographiques (dictionnaires, terminologies, thesaurus…). C'est cette application en particulier qui nous intéresse et sur laquelle nous nous concentrons par la suite.

3.1.2 Procédés

Deux axes principaux ont été proposés pour caractériser les relations sémantiques entre les mots. Une première approche s'inspire des travaux de Chomsky et Schutzenberg sur les langages formels et les grammaires génératives. Elle cherche à *marquer* un texte à l'aide d'un ensemble défini d'étiquettes sémantiques, pour désambiguïser les différents sens des mots (Katz et Fodor, 1963). Cette approche, bien que prometteuse, a été vivement critiquée, notamment parce qu'elle impose une nouvelle interprétation des étiquettes ajoutées au document pour le comprendre. Les étiquettes ne sont qu'un *meta-langage* (Lewis, 1972) qui n'est pas plus facilement interprétable que le texte qu'il cherche à caractériser sémantiquement. Cette approche a été suivie de plusieurs autres (Grefenstette les classe dans les approches d'*Intelligence Artificielle* et de *science cognitive*) qui ont toutes le défaut de devoir caractériser *a priori* exhaustivement la réalité pour faire face à toutes les situations ; on parle d'approche à base de connaissances (*knowledge-rich*). En d'autres termes, n'importe quel mot doit être caractérisable par une liste d'étiquettes définies par un humain. C'est en pratique possible pour des domaines restreints mais difficile à étendre à des domaines inconnus. Cette approche, dépendante de connaissances spécifiques *a priori*, a été étendue en utilisant des ressources linguistiques (dictionnaire, thésaurus) comme amorce pour construire automatiquement la base de connaissances nécessaire.

Grefenstette (1994b) propose une approche différente qui ne nécessite aucune connaissance *a priori*. Elle est basée sur l'étude statistique de la distribution des mots dans les corpus. Elle s'appuie sur une hypothèse proche de la *sémantique distributionnelle* de Firth (cf. section 1.4.1.1) :

> « *[...] si un concept est discuté au sein de deux textes différents, il y aura un large chevauchement des mots utilisés pour le décrire* »

Là encore, deux approches sont comparées : les approches syntaxiques et celles utilisant une fenêtre contextuelle ; Evert (2008) parle de *cooccurrence relationnelle* pour la première et *distributionnelle* pour la seconde. L'approche syntaxique s'intéresse, dans la tradition des travaux de Harris (voir notamment Harris, 1988), aux dépendances entre opérateurs et opérandes, entre gouverneurs et gouvernés. Ces travaux sont synthétisés notamment dans Habert et Zweigenbaum (2002), en particulier l'idée que les propriétés structurelles des langues ne se concrétisent pas à travers les cooccurrences (prenant le contre-pied de Firth), mais sur les dépendances entre un mot et l'ensemble de ses opérateurs et opérandes[2]. Elle nécessite un traitement préalable du corpus pour faire ressortir ces dépendances (en ne conservant que les plus simples). Grefenstette (1994b) présente le logiciel SEXTANT, qui se concentre sur les groupes nominaux et verbaux. Pour chaque mot étudié, les contextes sont les mots avec lesquels il entretient une relation de dépendance. Ces contextes sont alors comparés à l'aide d'une mesure de similarité (dans le cas de Grefenstette (1994b), la *Mesure de Jaccard Pondérée*, voir équation 4.16). Les contextes les plus proches sont alors regroupés et triés pour offrir des listes de mots similaires sémantiquement. L'approche utilisant une fenêtre contextuelle n'exploite plus les dépendances syntaxiques mais se contente d'extraire tous les mots cooccurrants dans une fenêtre donnée (d'une façon semblable à celle présentée pour l'approche directe en section 1.5.1) pour faire ressortir les cooccurrences significatives à l'aide d'une mesure d'association. Là encore, la mesure de Jaccard pondérée est utilisée pour ordonner les mots similaires. Les deux approches donnent des résultats différents en fonction des corpus utilisés et des mots analysés. Grefenstette (1996) compare précisément les deux, à l'aide des mêmes ressources et des mêmes résultats de référence et constate que l'approche syntaxique offre les meilleurs résultats pour les mots très fréquents (dans son cas, les 600 mots les plus fréquents sur un corpus de 400 000 mots). À l'inverse, alors que l'utilisation de fenêtres contextuelles est jugée comme une faiblesse (Zweigenbaum et Habert, 2006, page 33), elles donnent de meilleurs résultats pour les mots rares. Ce résultat n'est pas surprenant : les dépendances sont plus riches sémantiquement mais diminuent le nombre d'éléments de rapprochement entre les mots. Les mots moins fréquents ne peuvent supporter cette « disette » de caractérisation et ont besoin de plus de points de rapprochement pour obtenir des partionnements satisfaisants.

3.1.3 Affinités du premier, deuxième et troisième ordre

L'acquisition sémantique permet de détecter automatiquement trois types de relation particulière entre les mots. Grefenstette (1994a) parle d'affinité du premier, deuxième et troisième ordre.

Les affinités du premier ordre décrivent quels mots sont susceptibles d'apparaître dans le voisinage immédiat d'un mot donné [3]. Elles regroupent les *mots fortement associés*, c'est-à-dire les cooccurrences significatives et les *collocations*.

Les affinités du second ordre indiquent quels mots partagent le même environnement[4]. Ces mots n'apparaissent en principe jamais ensemble. Il peut s'agir, par exemple, de mot à la graphie très proche, par exemple *color* en anglais américain et *colour* en anglais britannique. Les deux sont corrects, mais apparaîtront rarement dans un même document. Hindle (1990) les extrait en calculant l'association, à l'aide de l'information mutuelle, entre des relations Verbe-Nom dans un corpus de 6 millions de mots (AP 1987). Les noms les plus associés au verbe *drink* sont *bunch-beer*, *tea*, *Pepsi*, *champagne*. Ces

[2] « *the structural property is not merely co-occurrence, or even frequent co-occurence, but rather dependence of a word on a set : an operator dos not appear in a sentence unless a word – one or another – of its argument set is there (or has been zeroed there). When that relation is satisfied in a word-sequence, the words constitute a sentence* ». (page 332 Harris, 1988).

[3] « *First-order term affinities describe what other words are likely to be found in the immediate vicinity of a given word* », (Grefenstette, 1994a, page 1).

[4] « *Second-order affinities show which words share the same environnement* »

affinités reflètent donc les mots *sémantiquement proches* et permettra par exemple de relever des relations de synonymie.

Les affinités du troisième ordre sont un reclassement d'affinités du second ordre selon des axes sémantiques. Chaque axe représente une caractéristique sémantique particulière (par exemple *animal*, *concret*, *humain*...). Chaque mot peut alors être représenté dans l'espace ainsi formé. Pour construire l'espace vectoriel – définir les axes sémantiques – il est possible de partir d'affinité du second ordre. Si A et B sont reliés (si A est relié à B, et B est réciproquement proche de A), alors A-B sera un axe de l'espace. Un mot C sera inscrit en fonction de son affinité avec A et avec B. Grefenstette (1994a) construit par exemple un axe *tumor – carcinoma* et y place *cancer disease*, qui sera également associé à l'axe *tumor – lesion*. Les affinités du troisième ordre permettent donc de dégager et de regrouper des nuances sémantiques. Elles permettent en particulier de séparer les différents sens d'un mot polysémique.

3.1.4 Vers l'acquisition sémantique bilingue

Dans notre cas nous ne cherchons plus les mots les plus similaires au sein d'un même corpus mais entre deux corpus dans des langues différentes. Un mot et sa traduction sont *sémantiquement proches*, voire sémantiquement équivalent dans le meilleur des cas. Il est donc naturel de s'appuyer sur les travaux en acquisition sémantique monolingue pour motiver les travaux dans un cadre multilingue. Ainsi, nous reprenons le concept de *contextualisation sémantique* des termes (à travers les vecteurs de contexte) et de *similarité sémantique* (à travers les mesures de distance).

Ce regard en arrière apporte un éclairage sur les travaux réalisés dans un cadre multilingue et nous permet également de relativiser la qualité des résultats de l'alignement de lexique bilingue par rapport au cadre monolingue. Le problème multilingue est plus délicat : techniquement, en raison de la nécessité d'un lexique pivot pour le transfert des vecteurs de contexte, mais aussi méthodologiquement, puisque nous ne cherchons plus idéalement des mots sémantiquement proches, mais des mots sémantiquement équivalents lorsqu'ils sont disponibles. Ainsi, *foie* sera relié à *organe* par une relation d'hyperonymie, mais *organe* ne sera pas une traduction satisfaisante de *liver*.

Zweigenbaum et Habert (2006) relèvent par ailleurs que le cadre multilingue contribue au cadre monolingue, bien que traditionnellement présenté comme une *version dégradée de ce qui est possible en corpus monolingue* (page 35). En effet, il a fallu pallier les difficultés amenées par les spécificités du cadre multilingue, par exemple, en introduisant des amorces lexicales, inexistantes à l'origine, notamment dans les premiers travaux de Rapp (1995) et Fung (1995a) (cf. section 1.4.1.

Nous développons dans les sections suivantes la mesure de l'*association* entre deux mots, pour évaluer dans quelle mesure ces mots sont sémantiquement reliés.

3.2 Statistique de la cooccurrence de termes

Cette section est largement inspirée des travaux de Stefan Evert concernant la *statistique de la cooccurrence de termes*. Elle permet d'introduire de nombreuses notions utiles pour la suite de notre étude. Ces travaux synthétisent de nombreuses mesures d'association et développent leur sens, d'un point de vue statistique et du point de vue de la *théorie de l'information* (Evert, 2004, 2008). L'auteur s'intéresse en particulier aux phénomènes des cooccurrences et des collocations. Nous présentons d'abord la problématique des collocations car elle est proche de celle de la caractérisation d'un terme par son environnement dans les vecteurs de contexte. Nous revenons ensuite sur les mesures d'association pour comprendre ce qu'elles évaluent et de quelle manière.

3.2.1 Collocations et cooccurrences

Revenons tout d'abord sur l'hypothèse *firthienne* (Firth, 1957) : *on reconnait un mot à ses fréquentations*[5]. Cette proposition appuie le fait que les mots n'apparaissent pas ensemble par hasard et ne se combinent pas par hasard pour former des phrases sur la seule base de la syntaxe.

Le concept de collocation est discuté (Tutin et Grossmann, 2002), mais il est admis qu'il reflète l'intuition que certains mots ont tendance à apparaître ensemble. Tutin et Grossmann (2002) reprennent cette notion tout en admettant qu'elle est floue. Les collocations peuvent être paradigmatiques (*médecin...patient*) ou syntagmatiques (*argument de poids*). Elles peuvent aussi être des combinaisons de mots privilégiés dont les différents composants sont difficilement interchangeables. Cette dernière définition caractérise les collocations par le fait que le sens est transparent en réception (il se comprend facilement) mais qu'il est difficile à produire pour un locuteur non natif. C'est le cas par exemple de *pluie torrentielle*, qui apparaît plus naturel que *précipitations torrentielles*. Une autre propriété de ces collocations est leur non-compositionnalité : elles ne peuvent être traduites mot à mot.

Dans notre cas, nous chercherons les cooccurrences syntagmatiques et paradigmatiques, qui absorberont de toutes façons les cooccurrences figées (ce sont des affinités du premier ordre). En pratique, nous nous intéresserons aux cooccurrences *significatives*, c'est-à-dire des paires de mots *significativement associés*, nous permettant de caractériser un mot par ses voisins.

3.2.2 Association entre deux termes

Un simple décompte des cooccurrences permet rarement de tirer des conclusions sur la force de la relation entre deux mots. Typiquement les articles français *le/la/l'* ont une fréquence très élevée dans les documents francophones sans pour autant être reliés à un mot précis puisqu'ils cooccurrent avec énormément de mots différents. Il est donc indispensable d'introduire des mesures statistiques plus subtiles pour évaluer la force de la relation entre deux mots. Ces mesures doivent combiner la fréquence de cooccurrence de deux mots avec les fréquences d'apparitions de l'un sans l'autre, mais aussi avec le nombre de mots total dans les documents ou tout autre indice pertinent pour caractériser cette association. Ajoutons que cette information est contextuelle et propre au corpus duquel sont extraits les termes évalués.

Les *mesures d'association* permettent d'évaluer quantitativement la force d'une telle relation entre deux termes. Elles doivent idéalement combiner dans un seul nombre réel à la fois l'association statistique entre deux termes (l'*effet* de leurs cooccurrences), mais aussi la confiance que l'on peut avoir dans une telle mesure (la *significativité* de leurs cooccurrences). Nous en présentons quelques unes parmi les plus classiques en les décrivant brièvement.

3.2.3 Mesures d'association simples

Une première mesure d'association, naturelle, va comparer les cooccurrences de deux mots observées avec les cooccurrences attendues dans le cas d'une hypothèse nulle. L'hypothèse nulle ici est que les mots sont distribués aléatoirement dans le corpus. L'association s'évalue en utilisant deux valeurs. La première, O, correspond au nombre de cooccurrences mesuré dans le corpus. La seconde correspond à la valeur E espérée sous l'hypothèse nulle. La valeur E est calculée à partir des fréquences des deux

[5] Cette citation est elle-même inspirée d'une fable d'Ésope dont la morale est que l'« *on reconnait un homme à ses fréquentations* ». Dans cette fable, un homme achète un âne mais le rend très rapidement à son propriétaire initial, après avoir remarqué que l'âne s'était immédiatement acoquiné avec les ânes paresseux. Fort de ce constat, il a jugé l'âne nouvellement acheté comme similaire à ses fréquentions. Le sens de la citation d'Ésope est donc sensiblement différent de celle de Firth. Pour Ésope, on ressemble à ses fréquentations, pour Firth, on est caractérisé par ses fréquentations.

mots. Evert (2008) donne l'exemple suivant, à partir de l'observation de *is to*. Ce bigramme est obtenu 260 fois dans le corpus étudié[6]. Ces deux mots sont fréquents l'un sans l'autre : *is* apparaît environ 10 000 fois et *to* environ 26 000 fois parmi le million de mots du corpus. Si les mots du corpus sont distribués aléatoirement, alors chacune des 10 000 occurrences de *is* a $26\,000/1\,000\,000 = 26/1\,000$ chances d'être suivie par *to*, soit $10\,000 \times (26/1\,000) = 260$, ce qui correspond à la fréquence observée du bigramme *is to*. Dans ce cas, il n'est pas possible de rejeter l'hypothèse nulle : ce bigramme apparaît autant de fois qu'attendu, leur association est nulle. La réalisation de l'évènement « le terme *is* apparaît » est statistiquement indépendante de l'évènement « le terme *to* apparaît ».

L'information mutuelle ponctuelle (equation 3.1 – par la suite, nous l'appelerons simplement information mutuelle) compare le nombre de cooccurrences observées au nombre de cooccurrences attendues sous l'hypothèse nulle.

$$IM_{ponctuelle} = log_2 \frac{O}{E} \tag{3.1}$$

Dans cette équation, l'usage du logarithme permet d'atténuer l'importance du quotient. De plus, si $O = E$, la mesure d'association sera nulle ($log_2(1) = 0$). Par ailleurs, si $O < E$ (si l'observation est en deçà de l'attente), la mesure d'association sera négative, indiquant que les deux termes apparaissent moins souvent ensemble qu'attendu, ce qui peut se révéler être un indice important (on parlera d'*anti-association*). En pratique, cette mesure ne réflète pas la confiance que l'on peut lui donner. Ainsi, l'information mutuelle sera identique pour $(O = 2; E = 1)$ et $(O = 100; E = 50)$. Pour pallier ce problème, il a été proposé de renforcer l'influence de l'observation en introduisant la famille de mesure $IM^k = log_2(O^k/E)$, par exemple dans Daille (1994) avec $k = 3$. Une autre façon de réduire ce problème consiste à pondérer l'information mutuelle par l'observation, c'est l'information mutuelle locale (équation 3.2) :

$$\textit{local-IM} = O \cdot log_2(O/E) \tag{3.2}$$

Toutefois, cette nouvelle mesure présente le défaut de ne plus révéler de façon homogène l'anti-association (au delà d'une valeur donnée, l'association va augmenter lentement alors que O diminue, pour une valeur fixe de E). À l'inverse, la famille des IM^k n'a pas ce défaut, mais n'évalue pas la confiance accordée à l'observation. Dunning (1993), constatant de nombreuses utilisations erronées de certains tests statistiques (en particulier, le test du χ_2) introduit le taux de vraisemblance (en anglais, *Log likelihood* – version simple en équation 3.3).

$$ll_{simple} = 2 \cdot (O \cdot log(O/E) - (O - E)) \tag{3.3}$$

Ce test de vraisemblance (qui évalue dans quelle mesure il est *vraisemblable* d'obtenir une telle observation par rapport à l'hypothèse nulle) est moins sensible aux valeurs de E faibles, ce qui est le cas dans des corpus de taille modeste. Toutefois, ce test présente le défaut de retourner des scores d'association positifs pour les associations comme pour les anti-associations, ce qui peut toutefois être corrigé assez simplement (lorsque $O < E$, il suffit de considérer l'opposé de l'association pour obtenir l'anti-association correspondante).

[6] Dans ce cas, le *Brown corpus*, `http://khnt.aksis.uib.no/icame/manuals/brown/`.

3.2.4 Table de contingence

Une table de contingence (table 3.1) permet de regrouper, pour deux termes i et j, le nombre de leurs cooccurrences, mais aussi le nombre de cooccurrence de l'un sans l'autre ; $occ(i, j)$ est le nombre de cooccurrences des éléments i et j, $\neg i$ représente toutes les unités considérées sauf i.

	j	$\neg j$	
i	$O_{11} = occ(i, j)$	$O_{12} = occ(i, \neg j)$	L_1
$\neg i$	$O_{21} = occ(\neg i, j)$	$O_{22} = occ(\neg i, \neg j)$	L_2
	C_1	C_2	N

	j	$\neg j$
i	$E_{11} = \frac{L_1 C_1}{N}$	$E_{12} = \frac{L_1 C_2}{N}$
$\neg i$	$E_{21} = \frac{R_2 C_1}{N}$	$E_{22} = \frac{R_2 C_2}{N}$

Table 3.1 – Table de contingence observée (à gauche) et espérée (à droite) pour un couple i et j.

Les valeurs L_i, C_i et N sont les sommes des lignes et des colonnes. Ce sont les valeurs *marginales*. Les tables de contingence permettent des calculs équivalents à ceux proposés en section 3.2.3 (le nombre d'occurrences observées correspond à la valeur O_{11}, le nombre de cooccurrences attendues est $E_{11} = L_1 C_1 / N$). Toutefois, elles permettent aussi des statistiques plus subtiles puisqu'elles décrivent plus finement l'observation. Il est possible de construire la table de contingence des valeurs E_{ij} attendue, et de comparer l'ensemble de ces valeurs avec les valeurs effectivement observées. Une table de contingence résume donc plus d'informations qu'un simple dénombrement de cooccurrences.

Le *Test Exact de Fisher* (equation 3.4) calcule sans approximation la significativité d'une table de contingence observée.

$$Fisher = \frac{(O_{11} + O_{12})! \cdot (O_{21} + O_{22})! \cdot (O_{11} + O_{21})! \cdot (O_{12} + O_{22})!}{N! \prod_{ij} (O_{ij}!)} \tag{3.4}$$

Ce test calcule la probabilité d'obtenir une telle table de contingence parmi l'ensemble des tables de contingence possibles pour des valeurs marginales fixes. En pratique ce test est difficile à utiliser en raison de son coût calculatoire. Toutefois, le taux de vraisemblance est une bonne approximation du test de Fisher, et peut donc être utilisé pour un coût calculatoire moindre (voir équation 3.5).

$$ll = 2 \sum_{ij} O_{ij} log \frac{O_{ij}}{E_{ij}} \tag{3.5}$$

Ce test est sensiblement une somme de taux de vraisemblance simple (eq. 3.3) appliqué à chaque case de la table de contingence. De même, la mesure d'information mutuelle locale peut-être généralisée sur une table de contingence, voir équation 3.6.

$$IM_{locale} = \sum_{ij} O_{ij} log_2 \frac{O_{ij}}{E_{ij}} \tag{3.6}$$

Cette mesure est alors homogène au taux de vraisemblance de l'équation 3.5 (à une constante près). Toutefois, elle diffère de l'information mutuelle ponctuelle (équation 3.1). L'information mutuelle ponctuelle donne une indication de la force de la relation entre deux termes i et j (en quelque sorte, dans quelle mesure l'apparition de i influence l'apparition de j) alors que l'information mutuelle généralisée indique dans quelle mesure la distribution de i influence la distribution de j dans le corpus. L'information mutuelle ponctuelle mesure un *effet* (la force d'une relation, très sensible à l'évolution de la taille de l'échantillon) alors que l'information mutuelle généralisée mesure la *significativité* de la relation entre i et j.

Nous introduisons une dernière mesure d'association, les *Odds Ratio*[7] équation 3.7.

$$odds = log\frac{O_{11}}{O_{21}}/\frac{O_{12}}{O_{22}} = log\frac{(O_{11})(O_{22})}{(O_{12})(O_{21})} \tag{3.7}$$

Cette mesure est le rapport des colonnes de la table de contingence. En effet, sous l'hypothèse nulle, ces rapports doivent être égaux (c'est évident au regard de la table de contingence espérée – table 3.1).

Ces mesures d'association sont des outils statistiques puissants pour caractériser la relation entre les mot, reflétant des relations sémantiques. Toutefois, elles sont indispensables mais pas suffisantes. Il reste à savoir sur quels mots les calculer et comment enregistrer globalement ces associations dans le but de les comparer d'une langue à l'autre. Il s'agit d'enregistrer le *contexte* d'un mot.

3.3 Contextes des mots

C'est le rôle des vecteurs de contexte de stocker, dans une seule structure, l'environnement d'un mot que l'on cherche à caractériser – le *mot vedette*, terme emprunté à Habert *et al.* (1997). Il en existe d'autres, par exemple, l'empreinte d'hétérogénéité de Fung (1995a) (voir chapitre 1). Toutefois, les vecteurs de contexte se sont imposés dans les approches statistiques, sous l'influence de Salton et Lesk (1968), que ce soit dans un cadre monolingue (cf. section 3.1) ou multilingue (cf. sections 1.5 et 1.7).

Nous l'avons évoqué dans les sections précédentes : deux approches dominent pour la constitution des vecteurs de contexte, c'est-à-dire pour le choix des mots à intégrer dans la caractérisation du mot vedette. La première s'appuie sur les relations syntaxiques et va enregistrer par exemple des associations Nom-Verbe. La seconde s'appuie sur une fenêtre d'une taille définie et enregistre tous les mots apparaissant dans la fenêtre entourant les différentes occurrences d'un mot vedette. Il en existe une troisième qui s'appuie sur des unités linguistiques, par exemple la phrase, le paragraphe ou le document entier.

En réalité, les approches syntaxiques et à base de fenêtre contextuelle ne sont pas en concurrence. En effet, les fenêtres de petite taille auront tendance à refléter des dépendances syntagmatiques, tout comme les approches syntaxiques. À l'inverse, les fenêtres de grande taille feront apparaître des dépendances paradigmatiques. Zweigenbaum et Habert (2006) soulignent que « *réduire la fenêtre à quelques mots à gauche et à droite du mot à caractériser revient à fournir une version simpliste des contextes syntaxiques* ». C'est une propriété intéressante que nous tâchons d'exploiter par la suite, car la tâche du découpage syntaxique automatique n'est pas évidente, en particulier dans un cadre multilingue. Précisons toutefois que la qualité des contextes syntaxiques ainsi construits est sensiblement dégradée : les vecteurs ne contiendront pas *uniquement* des dépendances syntagmatiques.

Dans ces approches, la catégorie morpho-syntaxique des mots enregistrés dans les vecteurs est d'importance aussi. C'est naturellement fait dans le cas de l'approche syntaxique et un filtrage peut être réalisé dans le cas de l'utilisation de fenêtres contextuelles. Il s'agit par exemple de ne pas enregistrer les mots outils (fréquents et peu informatifs) mais de ne garder que les mots les plus pertinents d'un point de vue sémantique.

3.4 Conclusion

Les mesures d'association nous permettent d'enregistrer des informations pertinentes supplémentaires dans les vecteurs de contexte : au lieu d'un simple dénombrement de cooccurrences, ces mesures

[7] Terme laissé tel quel car délicat à traduire. Il fait référence au fait que cette mesure utilise les valeurs en colonne, indiquant des « chances » (*odds* réfèrent par exemple aux chances de gagner un pari).

permettent de refléter certaines informations *sémantiques* sur les liens que les mots entretiennent entre eux. Dans le cadre de l'extraction lexicale bilingue à partir de corpus comparables, ce sont ces informations qui vont être comparées d'une langue à l'autre, là où un simple dénombrement surfacique aurait été trop instable et peu informatif en pratique.

Le chapitre suivant est consacré à l'implémentation de l'approche directe, que nous utilisons pour l'ensemble des expériences de nos travaux. Nous y discutons différentes notions techniques, en particulier concernant la construction des vecteurs de contexte et les mesures de similarité utilisées pour leurs comparaisons. Il présente l'ensemble des paramètres disponibles et leur influence à travers une série d'expériences de référence qui sont le point de départ des propositions d'améliorations que nous faisons par la suite.

CHAPITRE 4

Approche par traduction directe

Nous présentons dans ce chapitre l'implémentation que nous proposons de l'approche directe. La première partie, plus technique que les précédents chapitres, détaille chaque étape de l'alignement et l'ensemble des choix effectués pour les différents paramètres ou algorithmes. La seconde partie présente une série d'expériences sur notre implémentation pour d'une part, présenter des résultats de référence utilisés comme étalon pour la suite, d'autre part pour mettre en évidence quelques points d'observations pertinents.

4.1 Implémentation de l'approche directe

La chaîne de traitement implémentée permet de travailler sur des termes simples ou complexes, toutefois nous nous concentrerons uniquement sur les termes simples, seuls objets de cette étude. Nous avons implémenté l'approche directe grâce à une chaîne de traitement, appelé *Ziggurat*, initialement conçue par Samuel DUFOUR-KOWALSKY et Emmanuel MORIN.

4.1.1 Pré-traitements

Plusieurs pré-traitements doivent être appliqués sur les documents du corpus avant de pouvoir les exploiter :

- **Nettoyage du corpus** : conversion des documents en texte brut, à partir de document en PDF ou en HTML. Cette première étape entraîne généralement de nombreuses erreurs. En effet, les documents PDF (*Portable Document File*) contiennent au pire le document sous forme d'image, qu'il faut convertir en texte en utilisant un processus d'OCR[1], au mieux un ensemble de caractères et leurs positions absolues dans le document (sans caractère d'espace ou de retour à la ligne). Il faut alors recomposer les mots, les phrases et les paragraphes à partir de ces informations, ce qui entraîne généralement des erreurs de segmentation. Dans le cas des documents HTML, le traitement est plus simple et plus efficace mais ils contiennent généralement beaucoup d'informations peu reliées aux contenus qui nous intéressent dans le document (informations de navigation, bannières publicitaires…).
- **Segmentation** pour séparer les mots simples (espacement de la ponctuation lorsque pertinent, découpage en phrase…).
- **Étiquetage morpho-syntaxique** pour identifier, à partir de la morphologie de chaque mot son rôle syntaxique.

[1] *Optical Character Recognition* – reconnaissance optique de l'écriture hors-ligne.

– **Filtrage** pour éliminer les mots de faibles fréquences (paramétrable), ou que nous jugeons peu informatifs, en fonction de leur étiquette morpho-syntaxique (paramétrable également). Certains mots sont filtrés à partir d'une liste donnée (*stop words*).

Une étape d'extraction terminologique peut également être ajoutée, notamment pour identifier les termes complexes.

4.1.2 Collecte des contextes

À partir des résultats du pré-traitement, nous collectons tous les mots apparaissant dans une fenêtre donnée autour des mots à caractériser en s'appuyant sur quelques critères. Les critères utilisés sont une fréquence minimale d'apparition d'un terme dans l'ensemble du corpus, pour qu'il soit jugé suffisamment significatif. Cela permet par exemple de filtrer les erreurs de segmentation, qui donnent des résultats singuliers, mais aussi une liste de mots ou d'étiquettes syntaxiques à écarter. L'étape de collecte est présentée en pseudocode dans l'algorithme 4.1

ALG. 4.1 – Collecte des contextes d'un mot.

$t\_document \leftarrow document$ {Enregistrement du document sous la forme d'un tableau, un mot par entrée}
$h\_contextes \leftarrow Hachage$ {Déclaration d'une table de hachage pour stocker les résultats}
$fenetre \leftarrow n$ {Taille de la fenêtre contextuelle, centrée}
for all $mot, index\_mot \in t\_document$ **do**
  $debut \leftarrow (index\_mot - \frac{fenetre}{2})$
  $fin \leftarrow (index\_mot - \frac{fenetre}{2})$
  $t\_contexte = t\_document[debut, fin]$ {Sélection des mots dans la fenêtre}
  for all $cooc \in t\_contexte$ **do**
    if $est\_acceptable(cooc)$ **then**
      $h\_contextes[mot][cooc] = h\_contextes[mot][cooc] + 1$
    end if
  end for
end for

À la fin de cet algorithme, la table de hachage $h\_contextes$ contient, pour chaque mot i à caractériser, une table de hachage contenant son vecteur de contexte, associant à chaque mot i le nombre de ses cooccurrences avec j.

Les tables 4.1 et 4.2 présentent le résultat des étapes de nettoyage et de filtrage, à partir des séquences suivantes, extraites de la partie anglaise du corpus *cancer du sein* (voir section 2.2.2) :

> *What is the risk of recurrence or death for this patient occurring within five years after the diagnosis date ? Full-size image (52K)*
> *It did not have a direct prediction value for the real outcome of the patient or the state of the patient five years after diagnosis ; but it did predict the risk group with moderate Kappa values.*

Appliquons l'algorithme 4.1 sur ces exemples, pour produire le vecteur de contexte des mots *year* et *patient* (tableau 4.3). Nous considérons une taille de fenêtre de 3 (trois mots avant, trois mots après le mot à caractériser). Pour cet exemple, nous ne considérons pas de limite de fréquence minimale pour l'acceptation d'un mot.

0	risk
1	recurrence
2	death
3	patient
4	occur
5	year
6	diagnosis
7	date
8	full-size
9	images

Table 4.1 – Séquence extraite du corpus anglais *cancer du sein*, nettoyée et filtrée (1).

0	prediction
1	value
2	outcome
3	patient
4	state
5	patient
6	year
7	diagnosis
8	predict
9	risk
10	kappa
11	value

Table 4.2 – Séquence extraite du corpus anglais *cancer du sein*, nettoyée et filtrée (2).

mot	fréquence	mot	fréquence
year	**2**	*patient*	**3**
patient	3	year	3
diagnosis	2	outcome	2
death	1	state	2
occur	1	patient	2
date	1	risk	1
full-size	1	recurrence	1
state	1	death	1
predict	1	occur	1
risk	1	prediction	1
		value	1
		predict	1

Table 4.3 – Exemples de vecteurs de contexte.

Il est intéressant de noter dans les exemples du tableau 4.3 que le vecteur de contexte de *patient* contient *patient* (deux cooccurrences), ce qui n'est probablement pas une information discriminante. Par ailleurs, le second extrait contient deux occurrences proches du mot *patient* : leurs contextes se chevauchent et une certaine redondance apparaît dans le vecteur. La première occurrence apparaît encadrée par *prediction, value, outcome, state, patient* et *year*, la seconde occurrence est encadrée par *outcome, patient, state, year, diagnosis* et *predict*. Les occurrences de *outcome*, *state* et *year* sont donc comptées deux fois dans le vecteur. De plus, le premier extrait présente un exemple de bruit restant dans le corpus. La séquence *Full-size image (52K)* correspond à la légende d'une image affichée à cet endroit dans le document, qui est délicate à filtrer dans des documents peu structurés. Elle sera donc traitée normalement.

4.1.3 Calcul des associations

Les mesures d'associations ont été présentées en section 3.2. Dans le cadre de l'approche directe, les mesures les plus fréquemment utilisées sont l'information mutuelle (eq. 3.1) et le taux de vraisemblance (eq. 3.5).

La première étape du calcul des associations entre tête et élément des vecteurs de contexte consiste à calculer les tables de contingences pour toutes les paires de mots. Elles se calculent à partir de l'ensemble des vecteurs de contextes collectés. La table 4.4 et les équations 4.1 à 4.10 détaillent le calcul de ces tables, en reprenant la notation utilisée en section 3.2.1.

	i	$\neg i$	
j	O_{11}	O_{12}	L_1
$\neg j$	O_{21}	O_{22}	L_2
	C_1	C_2	N

Table 4.4 – Table de contingence pour deux mots i et j.

La valeur O_{11} (le nombre de cooccurrences de i et j) correspond à la valeur collectée dans le vecteur de contexte du mot i pour l'élément j (et réciproquement), elle peut donc être complétée directement. Nous disposons par ailleurs du nombre d'occurrences des éléments i et j, enregistrées lors de la construction du corpus et accessible à travers une fonction $occ(i)$. Ainsi les valeurs L_1 et C_1 peuvent être extraites des vecteurs de i (pour C_1) et de j (pour L_1), comme indiqué dans les équations suivantes (la fonction $cooc(a, b)$ correspond à la valeur O_{11} pour la table de contingence des mots a et b, c'est-à-dire au nombre de leurs cooccurrences).

$$L_1 = \sum_k cooc(j, k) \tag{4.1}$$

$$C_1 = \sum_k cooc(i, k) \tag{4.2}$$

Les valeurs O_{12} et O_{21} peuvent être inférées à partir des valeurs calculées précédemment :

$$O_{12} = L_1 - O_{11} \tag{4.3}$$
$$O_{21} = C_1 - O_{11} \tag{4.4}$$

La valeur N correspond à la somme de toutes les occurrences du corpus, elle est calculée à l'aide la fonction $occ(k)$ et est constante pour toutes les tables de contingences d'un même corpus :

$$N = \sum_k occ(k) \tag{4.5}$$

À partir de cette dernière valeur, nous pouvons calculer L_2 et C_2 :

$$L_2 = N - L_1 \tag{4.6}$$
$$C_2 = N - C_1 \tag{4.7}$$

Nous pouvons finalement calculer O_{22} (de plusieurs façons équivalentes) et compléter la table de contingence :

$$O_{22} = N - O_{11} - O_{21} - O_{12} \tag{4.8}$$

ou encore :

$$O_{22} = L_2 - O_{21} \tag{4.9}$$
$$O_{22} = C_2 - O_{12} \tag{4.10}$$

Le table 4.5 est la table de contingence des éléments *patient* et *year* à partir de deux séquences d'exemples précédentes (tableau 4.2 et 4.2).

	`patient`	`¬patient`
`year`	$O_{11} = 3$	$O_{12} =?$
`¬year`	$O_{21} =?$	$O_{22} =?$

Table 4.5 – Table de contingence (incomplète) des éléments *patient* et *year*.

La seule valeur proprement définie, en utilisant l'algorithme 4.1 est le nombre de cooccurrences de ces deux mots. Le nombre de cooccurrences de *patient* sans *year* (O_{21}) peut être calculé de deux façons. Intuitivement, d'après l'observation, il n'y a aucune occurrence de *patient* qui n'a pas *year* dans son contexte. Toutefois, d'après l'interprétation des tables de contingence, cette valeur doit pouvoir s'obtenir en soustrayant le nombre de cooccurrences des deux (O_{11}) à la fréquence de *year*. Or, dans ce cas, nous obtenons une valeur négative de -1, aberrante.

En réalité, pour obtenir une table de contingence valable dans ce cadre, il ne faut pas chercher à compter *dans combien de contextes i et j cooccurrent, dans combien de contextes l'un cooccure sans l'autre et dans combien de contextes ils ne cooccurent pas*, mais plutôt *dans combien de relation de cooccurrences i et j sont impliqués, dans combien l'un est impliqué sans l'autre, et dans combien aucun n'est impliqué*. La somme N de toutes les valeurs de la table n'est donc pas le nombre de fenêtres calculées, mais le nombre de paires de mots analysées, c'est-à-dire la somme du nombre de cooccurrences. La table de contingence complétée est présentée en table 4.6.

4.1.4 Filtrage des vecteurs de contexte

Les vecteurs de contexte ainsi construits sont généralement volumineux : ils rassemblent souvent plusieurs milliers d'éléments dont les associations sont parfois très proches de zéro. C'est typiquement le cas pour les éléments très fréquents qui ont beaucoup de voisins différents mais avec lesquels ils ne

	patient	¬patient	
year	$O_{11} = 3$	$O_{12} = 9$	$L1 = 12$
¬year	$O_{21} = 15$	$O_{22} = 69$	$L2 = 84$
	$C1 = 18$	$C2 = 78$	$N = 96$

Table 4.6 – Table de contingence complète des éléments *year* et *patient*.

sont pas fortement associés. Notre implémentation de l'approche directe propose donc de les élaguer. L'approche la plus simple et celle que nous avons implémentée consiste simplement à ne conserver que les premiers éléments (le seuil par défaut de notre approche est fixé à 5000 – les éléments sont triés par ordre décroissant d'association dans les vecteurs). Nous avons également implémenté d'autres techniques d'élagage, par exemple en fixant un seuil minimum pour le score d'association, de manière à retirer les éléments les moins significatifs. Ces techniques n'ont pas donné de résultats concluants. Il semble que ce filtrage ne retire pas suffisamment d'informations inutiles dans le cas des vecteurs très peuplés (contenant un grand nombre d'éléments) : restent toujours de nombreux éléments peu significatifs qui pénalisent l'étape de comparaison. Pour les vecteurs moins bien peuplés, cet élagage entraîne une plus grande pauvreté d'informations qui pénalise à son tour l'étape de comparaison. D'autre approches ont alors été tentées, notamment pour retirer les éléments les moins significatifs par rapport aux éléments les plus fortement associés (par exemple, en supprimant une fraction des derniers éléments du vecteur, ou en élagant par rapport à un indice de position calculé sur l'ensemble des valeurs). Nous n'avons observé dans ces cas qu'une altération marginale des résultats. Il semble que la méthode la plus simple soit aussi la plus efficace.

4.1.5 Traduction des vecteurs de contexte source

L'étape suivante consiste à transférer les vecteurs de contexte source dans l'espace vectoriel du corpus cible. Nous montrons à la section suivante que la comparaison entre vecteurs s'effectue sur les éléments communs (lexicalement parlant) entre les vecteurs sources et cibles. Il est donc nécessaire de les traduire, d'un point de vue linguistique, mais aussi d'un point de vue algébrique, pour qu'ils se conforment à l'espace vectoriel dans lequel s'inscrivent les vecteurs de la langue cible.

En fonction des traductions disponibles dans les lexiques bilingues utilisés, trois cas de figure peuvent apparaître :

1. aucune traduction n'est disponible dans le lexique pour un élément i donné ;
2. une seule traduction est disponible pour i ;
3. plusieurs traductions sont disponibles pour i.

1. Dans le cas où aucune traduction n'est disponible, l'élément n'est pas transféré et n'apparaîtra pas dans le vecteur traduit. C'est un problème car son pouvoir de caractérisation disparaitra.

2. Dans le cas où une seule traduction est disponible, elle se substitue à l'élément et prendra son score d'association dans le vecteur traduit. C'est un cas idéal, car il peut signifier que cet élément n'est pas ambigu. Il peut aussi être dû à des lacunes dans les ressources linguistiques et l'élément traduit peut ne pas correspondre à la traduction que l'on souhaiterait dans ce contexte. À titre d'exemple dans le cadre médical, la traduction de *drug* sera probablement *médicament*, alors qu'un lexique générique pourra proposer uniquement la traduction *drogue*.

3. Dans le cas où plusieurs traductions sont disponibles, elles sont prises en compte en fonction de leur fréquence dans le corpus cible. C'est-à-dire que le score d'association de l'élément i sera réparti

entre les différentes traductions j_x, cette répartition se faisant en fonction de la fréquence des éléments j_x dans le corpus cible. Le nouveau score d'association se calcule ainsi :

$$\forall j_x; j_x \in Traductions(i)$$
$$assoc(j_x) = assoc(i) \times \frac{freq(j_x)}{\sum_k freq(j_k)}$$

Cette méthode permet de faire ressortir les traductions les plus pertinentes pour un contexte donné. En reprenant l'exemple précédent, si le lexique bilingue fournit *drogue* et *médicament* comme traduction de *drug*, ces deux traductions seront pondérées différemment en fonction de leurs occurrences dans le corpus cible. Dans un cadre médical, il est probable que la traduction *médicament* sera favorisée (*drogue* pourra même obtenir une association nulle s'il n'apparaît jamais dans le corpus). La figure 4.1 schématise ces trois cas en reprenant l'image des motifs d'association du chapitre 1.

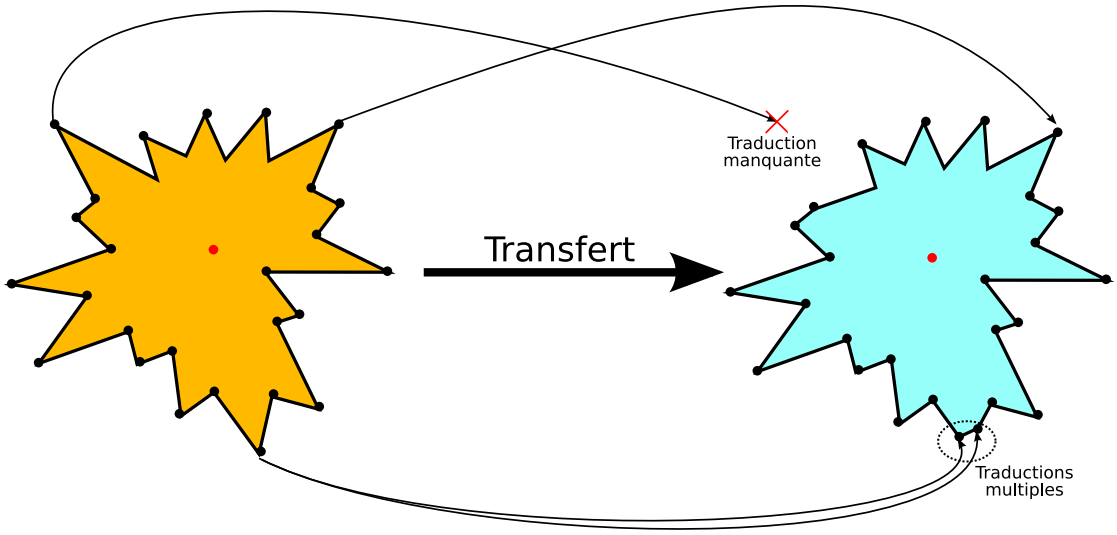


Figure 4.1 – Transfert des vecteurs de contexte.

Après l'étape de traduction des vecteurs sources, une nouvelle étape de filtrage peut être appliquée puisque l'étape de traduction augmente généralement considérablement la taille des vecteurs de contexte traduits, en raison des traductions multiples. L'étape suivante consiste à comparer les vecteurs de contexte source-traduits avec les vecteurs de contexte cible. Nous ouvrons une parenthèse pour aborder la notion de mesure de similarité dans la section suivante.

4.1.6 Mesures de similarité et de distance

4.1.6.1 Mesures de similarité

Dans le cadre général, une mesure de similarité va évaluer la ressemblance entre deux objets. Ces mesures retournent un intervalle de valeurs borné (généralement entre 0 et 1, 1 indiquant une similarité parfaite, 0 indiquant que les deux objets n'ont rien en commun). La comparaison des deux éléments se fait sur un ou plusieurs critères, quantitatifs ou qualitatifs. Par exemple si l'on veut comparer deux espèces animales, plusieurs critères peuvent entrer en compte. Qualitativement, on peut s'intéresser à la

présence ou l'absence d'un organe donné (par exemple, des ailes) ou la famille (mammifère, ovipare). Quantitativement, on peut observer la taille moyenne de l'animal, la durée de gestation ou mesurer la différence entre le génome de deux espèces. Cette différence est une mesure de distance, qui, à l'inverse des mesures de similarité va retourner 0 pour deux objets identiques et des valeurs élevées non bornées pour des éléments très différents.

Une première mesure de similarité, introduite par le botaniste Paul Jaccard a pour but de comparer le nombre de traits communs entre deux espèces afin de les classer. Elle est présentée en équation 4.11.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{4.11}$$

Elle correspond au rapport entre le nombre d'éléments communs dans les ensembles A et B par rapport au nombre total d'éléments de A et B. Deux ensembles identiques auront une similarité de 1 ($A \cap B = A \cup B \rightarrow A \cap B = A$), deux ensembles disjoints auront une similarité nulle ($A \cap B = \emptyset$). Cette mesure évalue donc le nombre de caractères *qualitatifs* communs entre les deux objets comparés.

Le quotien de similarité de Sorenson (équivalent au coefficient *Dice*) s'applique aussi à des critères qualitatifs. Il est présenté en équation 4.12 :

$$Sorenson(A,B) = \frac{2|A \cap B|}{|A| + |B|} \tag{4.12}$$

Ce quotien est le rapport du nombre d'éléments communs entre A et B divisé par le nombre d'éléments de A et de B. Il dérive directement de l'indice de Jaccard : $|A| + |B| = |A \cup B| + |A \cap B|$. Si $A = B$ alors $|A \cap B| = |A \cup B|$, l'indice devient :

$$Sorenson(A,A) = \frac{2|A|}{|A| + |A|} = \frac{2|A|}{2|A|} = 1$$

L'indice de Jaccard a été généralisé aux vecteurs de données quantitatives, c'est-à-dire des critères quantitatifs associés deux à deux. Il s'agit du *coefficient de Tanimoto* (Tanimoto, 1958), équation 4.13.

$$Tanimoto(A,B) = \frac{A \cdot B}{\|A\|^2 + \|B\|^2 - A \cdot B} \tag{4.13}$$

Dans cette équation, $\|V\|$ est la norme du vecteur V, c'est-à-dire la racine carrée du produit scalaire de V avec V ; $A \cdot B$ correspond au produit scalaire des vecteurs A et B.

Cette formule fait le lien entre l'indice de Jaccard et la mesure du cosinus, présenté en équation 4.14.

$$Cosinus(A,B) = \frac{A \cdot B}{\|A\|\|B\|} \tag{4.14}$$

Cette mesure retourne des valeurs entre -1 et 1. Une valeur de 0 indique que les deux vecteurs sont indépendants (c'est-à-dire orthogonaux – ce qui peut signifier qu'ils n'ont aucun critère en commun), une valeur de 1 indique deux vecteurs identiques et une valeur de -1 deux vecteurs opposés.

4.1.6.2 Mesure de distance

Une *mesure de distance* $D(A,B)$ est une fonction qui évalue la distance entre les éléments d'un ensemble. Ce sont des indices comparables aux mesures de similarités, il est d'ailleurs très simple de convertir une similarité en distance, et réciproquement.

La distance la plus intuitive est la distance euclidienne qui, dans un espace euclidien en deux dimensions correspond à la longueur du segment reliant deux points. Cette distance se généralise facilement dans un espace à *n* dimensions, voir l'équation 4.15.

$$Euclide(A, B) = \sqrt{\sum_i |a_i - b_i|^2} \tag{4.15}$$

Dans cette équation, $|x|$ est la valeur absolue de x. Cette mesure présente toutefois le défaut d'augmenter avec le nombre de critères pris en compte. Ainsi, au lieu de rendre plus pertinente une comparaison sur un ensemble de critères plus significatifs, cette distance favorisera les objets qui n'ont que peu de critères en commun (deux objets n'ayant aucune dimension commune se verront attribuer une distance de 0, car ils ne présentent aucune dissimilarité mesurable). Cette mesure de distance a rapidement été écartée dans le cadre de l'approche directe.

Une autre mesure de distance, très classique en théorie de l'information, est la distance de Levenshtein (Levenshtein, 1966), dont un exemple est donné en section 2.6.3. Elle mesure le nombre de modifications atomiques à effectuer pour transformer un objet A en B. Elle n'est pas utilisée ici puisque peu pertinente pour notre problématique.

Pour conclure cette parenthèse, l'intérêt des mesures de similarité ou de distance n'est pas dans l'interprétation brute du résultat qu'elles retournent, mais dans le fait qu'elles retournent des valeurs numériques comparables. Dans l'application de l'approche directe, nous ne nous intéressons pas au fait que deux vecteurs de contexte soient proches, mais au fait que deux vecteurs soient *plus similaires* (ou *moins distants*) que les autres paires de vecteurs candidats. Ces mesures nous permettent donc d'effectuer un classement des résultats.

4.1.7 Recherche des candidats à la traduction

La dernière étape majeure consiste donc à comparer les vecteurs de contexte « *sources-traduits* » avec l'ensemble des vecteurs de contexte de la langue cible, en utilisant les mesures de similarité décrites précédemment. C'est un problème délicat en raison des propriétés intrinsèques de ces mesures, en particulier le problème des valeurs nulles : le fait qu'un vecteur soit nul pour une composante n'est généralement pas un indice très significatif en soi. C'est en tout cas beaucoup moins significatif que la réciproque. Dans notre cas, que la composante d'un vecteur de contexte soit nulle pour une dimension j signifie que l'association entre la tête i du vecteur et l'élément j est nulle, ce qui signifie généralement que les deux termes i et j ne sont pas apparus ensemble dans une fenêtre donnée dans le corpus. Ce n'est pas une information très importante dans la mesure où les échantillons sur lesquels nous travaillons ne sont pas universellement représentatifs, même dans le cas de gros corpus.

La question se pose de savoir si des valeurs nulles communes à deux vecteurs sont un indice de leur similarité. Mais la question se pose aussi de savoir s'il faut comparer une composante nulle ou indéfinie avec une composante non nulle. Dans le cas de l'approche directe, cela revient à savoir s'il faut prendre en compte le fait qu'un élément soit défini pour une dimension dans un vecteur mais pas dans l'autre. Une approche classique consiste à ignorer les éléments qui ne sont pas communs aux deux vecteurs (c'est typiquement ce qui se passe avec la mesure du cosinus). Là encore, ce choix se discute car l'absence de nombreux éléments entre deux vecteurs est un indice de leur dissimilarité. Cette absence peut toutefois être due aux lacunes des ressources bilingues utilisées pour le transfert des vecteurs. Ces arguments font une fois encore apparaître la difficulté de trouver une mesure de similarité optimale dans l'approche directe en raison des nombreux paramètres qui entrent en jeu.

Grefenstette (1994a) propose une mesure basée sur la mesure de Jaccard (eq. 4.11) qui ne prend pas en compte les valeurs nulles, mais utilise le poids des valeurs communes aux deux vecteurs de contexte. Il s'agit de la *mesure de Jaccard pondérée*, présentée en équation 4.16. Dans cette équation, $V_s[i]$ (respectivement $V_t[i]$) correspond à l'association de l'élément i dans le vecteur source V_s (respectivement, dans le vecteur cible V_t).

$$JP(V_s, V_t) = \frac{\sum_i min(V_s[i], V_t[i])}{\sum_i max(V_s[i], V_t[i])} \tag{4.16}$$

Si les deux vecteurs sont identiques ou très proches, ils auront des associations très proches pour les éléments communs : $min(V_s[i], V_t[i])$ sera sensiblement égal à $max(V_s[i], V_t[i])$ quel que soit l'élément i, le quotient des deux sera donc proche de 1.

Nous avons implémenté cette mesure ainsi que la mesure du cosinus, que nous présentons à nouveau en équation 4.17 pour l'adapter aux vecteurs de contexte.

$$Cosinus_{CV}(V_s, V_t) = \frac{\sum_i V_s[i] \times V_t[i]}{\sqrt{\sum_i V_s[i]^2}\sqrt{\sum_j V_t[j]^2}} \tag{4.17}$$

La chaîne de traitement renvoie donc, pour chaque élément dont nous souhaitons obtenir une traduction, une liste de candidats ordonnées par similarité. Nous utilisons ces listes (ainsi que la liste des traductions de référence) pour produire un ensemble de statistiques sur les résultats, par exemple :

- Pour un mot à traduire donné nous relevons la position de sa traduction dans la liste et par extension, pour l'ensemble des mots à traduire, le nombre de traductions obtenues en première position (Top_1), avant la cinquième position (Top_5)...
- La liste des mots qui ont effectivement obtenus une traduction correcte et leurs positions (lorsque plusieurs traductions sont acceptées dans la liste d'évaluation).
- Quelques graphiques pour visualiser la qualité de l'alignement et pour comparer rapidement les résultats d'une expérience à l'autre.

4.1.8 Évaluation de la complexité

L'ensemble des traitements se fait en temps polynomial, ce qui est raisonnable mais parfois long en raison de la taille des jeux de données. Nous analysons la complexité de chaque étape de la chaîne de traitement pour évaluer le coût global. Nous commençons par la collecte des vecteurs de contexte, puisque nous n'avons pas la main sur les outils de pré-traitements[2].

La collecte des vecteurs de contexte dépend de la taille de la fenêtre contextuelle utilisée (w) et du nombre de mots dans les corpus à analyser (n_s et n_t, pour les corpus sources et cibles). Elle a un coût en $O(w(n_s + n_t))$. Par ailleurs, cette étape n'est généralement effectuée qu'une fois. En effet, sauf modification du corpus, l'ensemble des vecteurs de contexte peut être généré pour différentes combinaisons de paramètres, pour obtenir un *corpus de vecteurs de contexte* sur lesquels nous travaillons par la suite, en faisant varier les autres paramètres.

La construction des tables de contingence dépend du nombre de vecteurs de contexte (lui-même dépendant du nombre de mots dans le corpus), majoré par n_s et n_t, ainsi que du nombre d'éléments par vecteur de contexte, également majoré par n_s et n_t. Le coût final est donc en $O(n_s^2 + n_t^2)$ dans le pire cas.

[2] Les pré-traitements sont de toutes façons réalisés en amont, donc de complexité négligeable.

Le calcul des scores d'association est fonction de la mesure d'association utilisée. Pour l'information mutuelle comme pour le *Taux de vraisemblance*, le coût de chaque opération est fixe mais doit être répété pour chaque vecteur de contexte. Le coût est donc en $O(n_s + n_t)$.

La traduction des vecteurs sources dépend de la taille du lexique bilingue utilisé (accès au dictionnaire, $O(ln(n_l))$), du nombre de vecteur à traduire et du nombre de mots par vecteur à traduire (majorés par n_s^2). En théorie, son coût est en $O(ln(n_l) \cdot n_s^2)$.

La comparaison des vecteurs de contexte est l'opération la plus gourmande en mémoire. Elle nécessite des comparaisons de tous les éléments de tous les vecteurs sources (n_s) avec tous les éléments de tous les vecteurs cibles (n_t) – c'est une des raisons pour lesquelles seuls les vecteurs sources pour lesquels nous cherchons une traduction sont conservés. Pour un vecteur source, le calcul est en $O(n_s n_t^2)$, pour l'ensemble des vecteurs sources, il est donc en $O(n_s^2 n_t^2)$.

L'ensemble du processus est la somme de chaque étape, soit $O(w(n_s + n_t) + n_s^2 + n_t^2 + ln(n_l) \cdot n_s^2) + n_s^2 n_t^2)$, soit $O(n_s^2 n_t^2)$.

Dans la section suivante, nous éprouvons notre implémentation de l'approche directe sur l'un des corpus présentés au chapitre 2, ainsi que les ressources linguistiques associées en combinant différents paramètres. Cette section nous donne un étalon des résultats que nous obtenons et nous permet de tirer quelques conclusions quant aux combinaisons optimales de paramètres.

4.2 Évaluation de l'approche directe pour l'alignement bilingue

Dans cette section, nous présentons un ensemble de résultats de référence, pour évaluer l'influence des différents paramètres sur la qualité de l'alignement. Cette section nous permet de présenter en pratique les résultats de l'approche directe et de mettre en application les différentes notions abordées dans ce chapitre et dans les précédents.

4.2.1 Étalonnage

Nous avons étalonné la chaîne de traitement sur le corpus *cancer du sein* en utilisant les listes de références présentées en section 2.2.4. Les résultats de l'alignement sont résumés dans la table 4.7 et dans la figure 4.2. Ce sont les résultats optimaux obtenus sur ce corpus, en utilisant le *Taux de vraisemblance* et la mesure de *Jaccard pondérée*, pour un alignement de l'anglais vers le français en utilisant une taille de fenêtre contextuelle de 3 (3 mots avant, trois mots après le mot vedette).

	[En-Fr-122]	[En-Fr-648]
Top_1	25 (20,5 %)	83 (12,8 %)
Top_{10}	57 (46,7 %)	223 (34,4 %)
Top_{20}	69 (56,6 %)	263 (40,6 %)

Table 4.7 – Résultats de l'alignement anglais français, corpus *cancer du sein*.

La figure 4.2 présente en abscisse le rang des traductions obtenues et en ordonnée leurs scores de similarité, exprimés à l'aide de croix sur le graphique. Elle indique sous forme de courbe le nombre de traductions trouvées pour chaque Top (échelle en ordonnée à droite). Ce type de représentation, que nous réutilisons par la suite, permet d'évaluer visuellement la qualité et les défauts de chaque combinaison de paramètres.

Ces deux expériences montrent une tendance que nous retrouverons par la suite : la plupart des traductions sont obtenues pour des Top faibles, comme le montre la forte croissance des courbes pour les

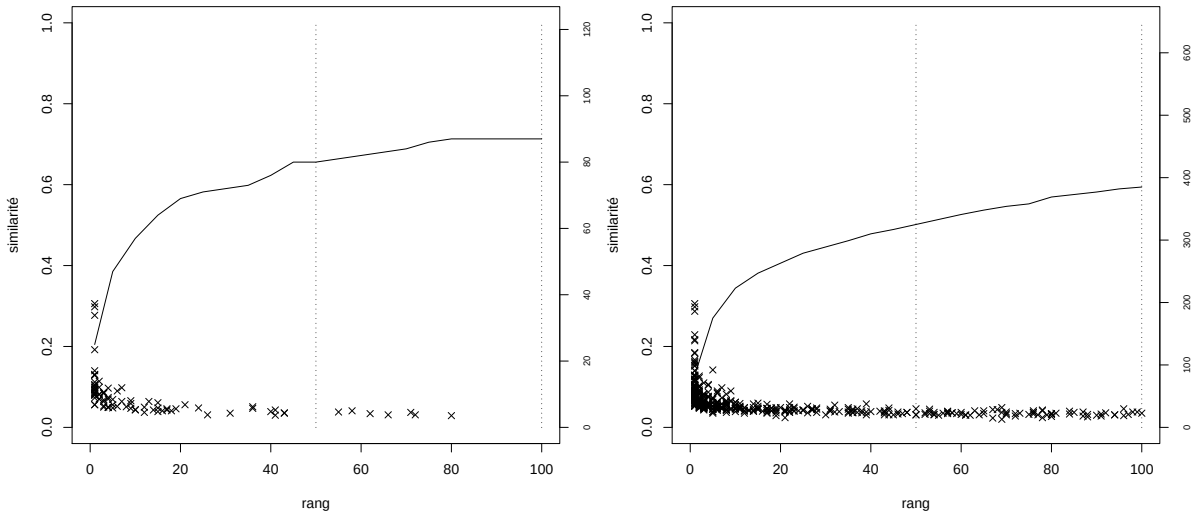


Figure 4.2 – Résultats de l'alignement anglais français, corpus *cancer du sein*. Liste [En-Fr-122] à gauche, [En-Fr-648] à droite.

rangs faibles (environ jusqu'au rang 20). D'autres traductions apparaissent par la suite mais de manière moins abondantes, comme l'indique la faible croissance des courbes au delà du rang 20.

La figure 4.3 compare les résultats pour différentes classes de fréquences[3]. Elle nous permet d'évaluer la qualité de l'alignement en fonction de la fréquence des termes à traduire.

L'influence de la fréquence est nette : les mots les plus fréquents obtiennent plus de traductions dans les premiers rangs des Top. Bien que l'effectif soit faible pour les mots occurrant plus de 800 fois, ils sont tous traduits au dela du Top_{10} pour la liste [En-Fr-122] et la précision atteint 95 % au dela du Top_{30} pour la liste [En-Fr-648]. À l'inverse, les termes occurrant peu fréquemment (en particulier la classe la plus basse) sont beaucoup plus mal traduits.

Nous réalisons par la suite trois expériences ; elles partagent les mêmes paramètres, à l'exception du paramètre dont nous voulons mesurer l'influence :

1. la taille de la fenêtre contextuelle ;
2. les mesures d'association ;
3. les mesures de distance.

4.2.2 Expériences

La première expérience consiste à faire varier la taille de la fenêtre contextuelle, de 2 à 25 pour observer la qualité des résultats. Les résultats sont présentés dans la figure 4.4.

Ces résultats indiquent que la taille de la fenêtre a une influence peu significative sur l'alignement : les courbes se chevauchent, il est délicat de déterminer une taille de fenêtre optimale dans ce cas. Cette question sera traitée plus en détail dans le chapitre 5.

[3] Il s'agit ici en réalité d'un dénombrement d'occurrences dans le corpus, c'est-à-dire d'un effectif. Toutefois, les expériences étant menées sur le même corpus, l'effectif est directement proportionnel à la fréquence dans ce cas.

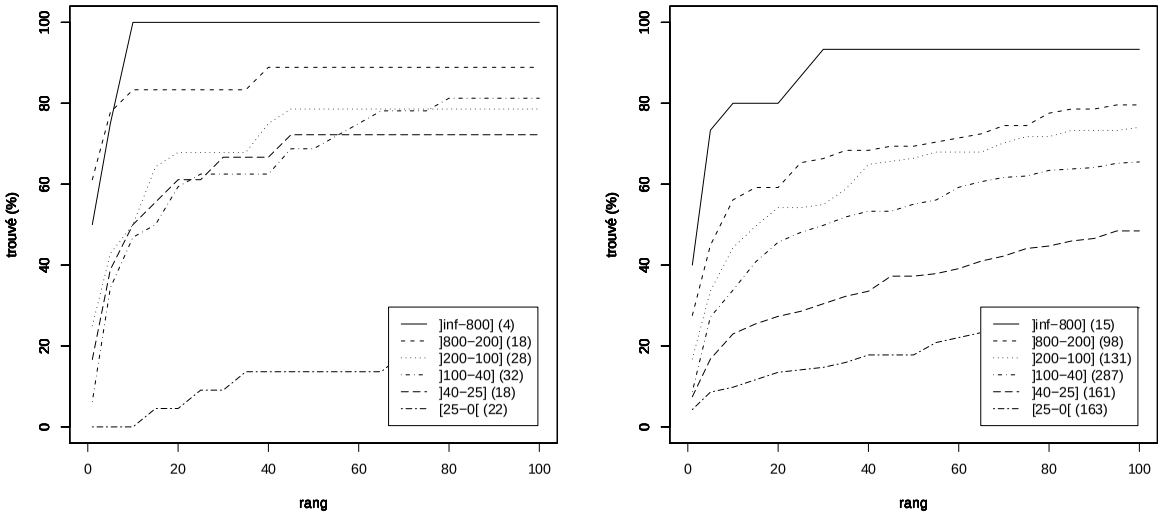


Figure 4.3 – Résultats de l'alignement anglais français, corpus *cancer du sein*. Liste [En-Fr-122] à gauche, [En-Fr-648] à droite. Influence de la fréquence (entre crochets, les intervalles de fréquences, entre parenthèses, l'effectif de chaque classe).

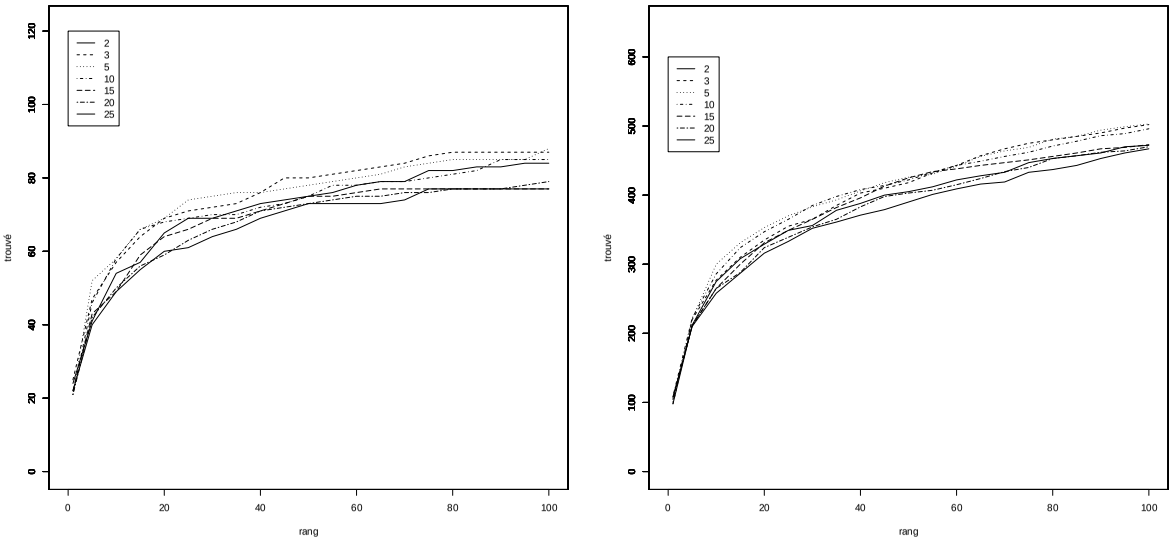


Figure 4.4 – Résultats de l'alignement anglais-français, corpus *cancer du sein*. Liste [En-Fr-122] (à gauche) et [En-Fr-648] (à droite). Comparaison des tailles de fenêtres contextuelles.

Nous avons mesuré la qualité des résultats en fonction de trois mesures d'association, le taux de vraisemblance (expérience de référence – TV), l'information mutuelle (IM) et l'information mutuelle au cube (IM3). Les résultats sont présentés dans la figure 4.5.

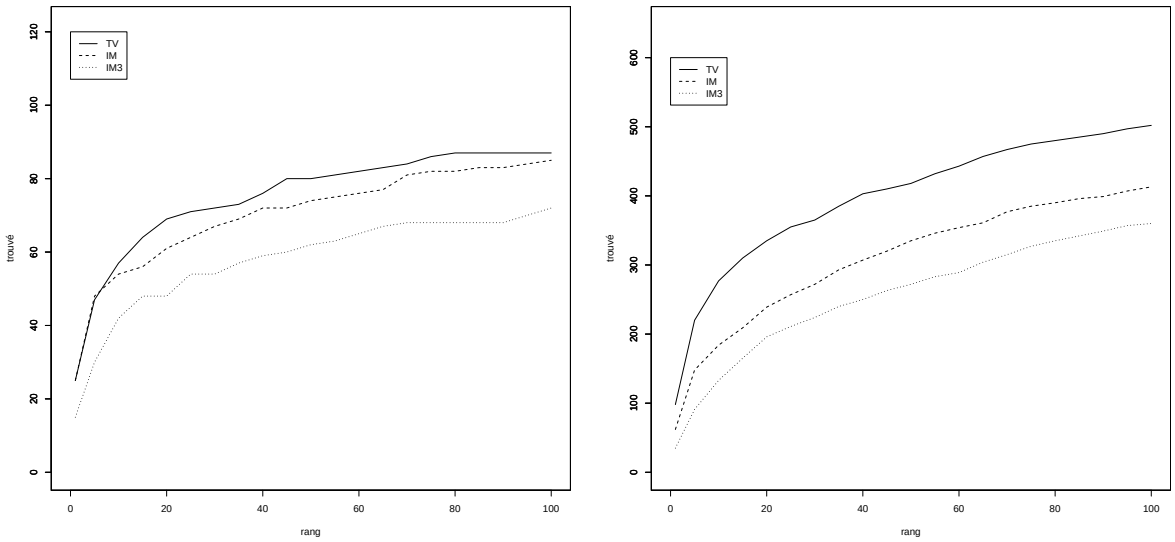


Figure 4.5 – Résultats de l'alignement anglais-français, corpus *cancer du sein*. Liste [En-Fr-122] (à gauche) et [En-Fr-648] (à droite). Comparaison des mesures d'associations.

Les résultats sont ici beaucoup plus contrastés : la dégradation de la qualité des résultats est nette au-delà du Top_{10}, en particulier pour l'information mutuelle au cube. Dans ce cas, les mesures d'association ont une influence très importante sur la qualité des résultats.

Les résultats utilisant les différentes mesures de similarité (à savoir, le cosinus et la mesure de Jaccard pondérée) sont présentés figure 4.6.

Les résultats en utilisant la mesure du cosinus sont là encore moins bons de façon assez contrastée. La perte de qualité est particulièrement nette pour la liste [En-Fr-122] avec laquelle nous observons une diminution de la qualité d'environ 14 points pour les Top_1 et Top_{10}, et plus de 17 points pour le Top_{20}. Cette comparaison indique que le choix de la mesure de distance a une influence considérable sur les résultats.

4.2.3 Discussions

Partant de l'ensemble des résultats que nous présentons dans cette section, nous pourrions conclure que la combinaison taux de vraisemblance / Jaccard pondérée est la meilleure. En réalité ce n'est pas toujours le cas : le choix des paramètres (incluant également la taille de la fenêtre contextuelle) dépend des corpus utilisés. Ainsi, pour le corpus *diabète et alimentation*, les résultats optimaux sont obtenus en utilisant une taille de fenêtre contextuelle de 25 mots et la combinaison taux de vraisemblance / cosinus (cf. expériences du chapitre 5). De plus, dans le cas du corpus *cancer du sein* mais en alignant du français vers l'anglais, les meilleurs résultats s'obtiennent en utilisant l'information mutuelle et le cosinus mais ces résultats sont moins bons que dans le sens anglais vers français (Morin, 2009). C'est un problème, car cela semble indiquer qu'il n'y a pas de combinaison de mesures optimales *a priori*. De fait,

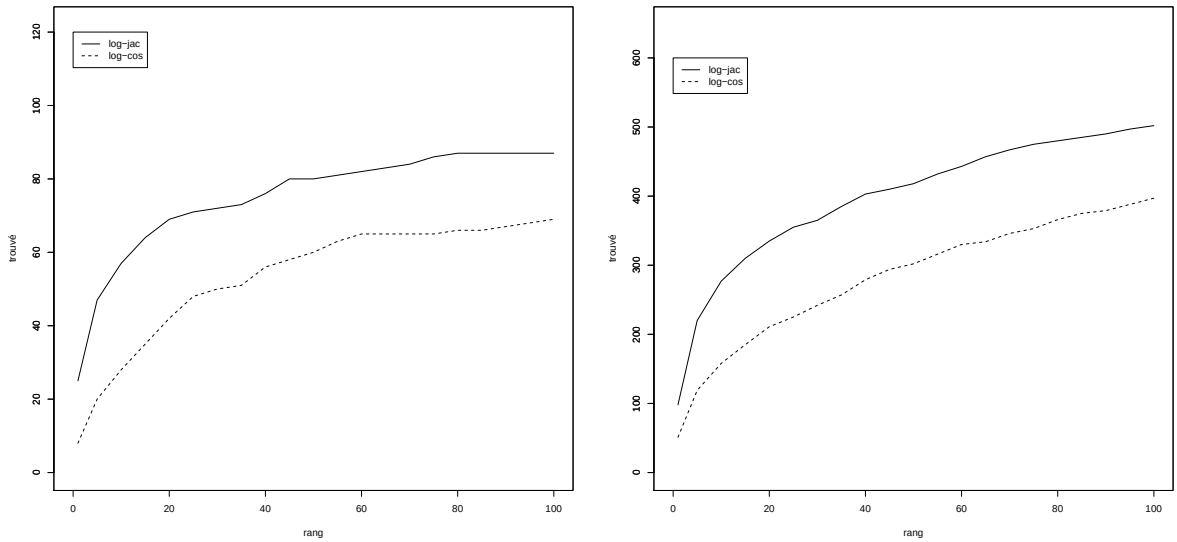


Figure 4.6 – Résultats de l'alignement anglais français, corpus *cancer du sein*. Liste [En-Fr-122] (à gauche) et [En-Fr-648] (à droite). Comparaison des mesures de similarité.

la littérature propose de nombreuses expériences semblables aux nôtres, et le choix des paramètres est obtenu empiriquement parmi l'ensemble des paramètres disponibles, en retenant la combinaison donnant les meilleurs résultats.

Par ailleurs, nous pouvons imaginer sélectionner les paramètres optimaux à partir de quelques jeux d'expériences : une première série en fixant tous les paramètres sauf la taille de fenêtre (par exemple), en regardant laquelle donne les meilleurs résultats, une seconde en choisissant la taille de fenêtre optimale ainsi qu'une mesure d'association, en faisant varier la mesure de similarité, enfin une dernière série en prenant la taille de fenêtre et la similarité optimale pour déterminer la mesure d'association optimale. Malheureusement, ces paramètres s'influencent les uns les autres et ce processus de sélection n'est pas efficace : la meilleure solution consiste donc à réaliser toutes les expériences correspondant à toutes les combinaisons de paramètres possibles pour déterminer la plus efficace. C'est un problème, d'abord en raison du coût calculatoire de chaque étape, mais aussi parce que dans un cas pratique, les bonnes traductions ne sont pas connues et il est impossible d'étalonner l'ensemble du processus sans une ressource de référence. Typiquement, il n'est pas en possible en l'état de proposer un système d'extraction lexicale bilingue vierge d'*a priori* sur les corpus utilisés. Il faudra l'étalonner à nouveau pour chaque nouvel usage (qui peut varier par exemple en fonction du type de discours, du domaine et des langues concernées par le corpus, mais aussi en fonction de la direction de traduction souhaitée). Cette problématique sera notamment l'objet de la discussion du chapitre 6.

Le problème s'aggrave encore en fonction des objectifs : certaines combinaisons de paramètres donnent de bons résultats pour des Top très faibles (ce qui est intéressant en pratique) mais de moins bons passé un certain seuil. À l'inverse, d'autres donnent des résultats moyens pour tous les Top : il est difficile de définir *a priori* une combinaison optimale dans ces cas. Ces deux problèmes (influences mutuelles des paramètres et difficulté à comparer la qualité des résultats) sont illustrés dans la figure 4.7.

La figure 4.7 montre que l'emploi de la combinaison information mutuelle locale (équation 3.2) et Jaccard pondérée donne de meilleurs résultats pour les Top de 1 à 20. Au-delà, c'est la combinaison

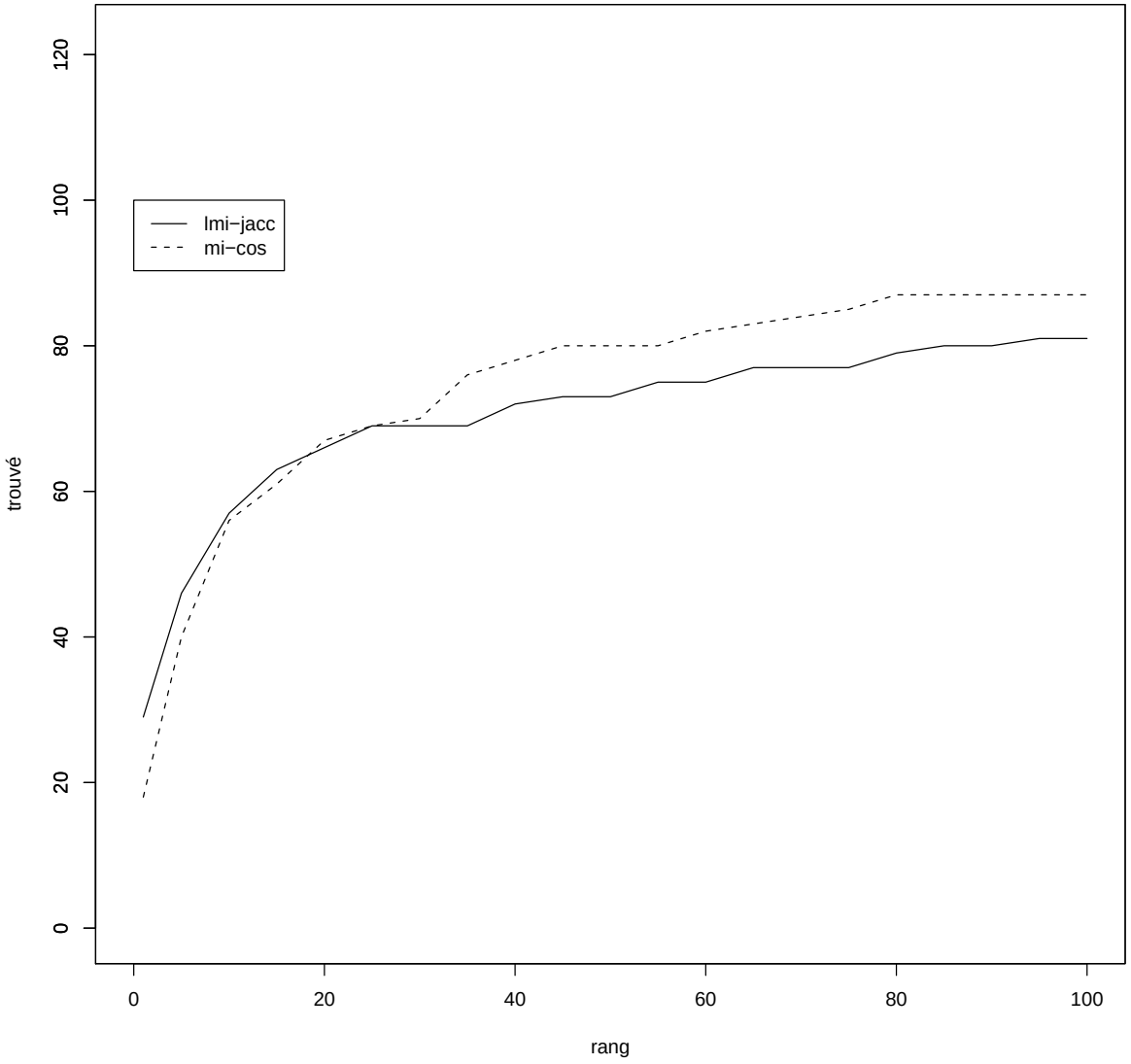


Figure 4.7 – Résultats de l'alignement français anglais, corpus *cancer du sein*. Liste [En-Fr-122]. Mesures utilisées : information mutuelle locale et Jaccard pondérée (courbe continue) ; information mutuelle et cosinus (courbe pointillée).

information mutuelle et cosinus qui donne les meilleurs résultats. Il est difficile dans ce cas de trancher pour sélectionner une combinaison optimale. Toutefois, la mesure d'information mutuelle locale donne dans la majorité des cas des résultats décevants, le résultat présenté ici étant en quelque sorte une exception (c'est pourquoi cette mesure n'a pas été prise en compte dans le processus d'étalonnage). Ce type de résultats est fréquent, dès lors une attitude pragmatique semble être de rigueur avec une décision empirique dépendante des données et des résultats escomptés.

4.3 Conclusion

Dans ce chapitre, nous avons détaillé notre implémentation de l'approche directe, en expliquant les mesures de similarité utilisées. Il est intéressant de noter que les mesures qui apparaissaient comme optimales pour le travail à réaliser (en particulier, le taux de vraisemblance, approximation fiable du test exact de Fisher) ne sont pas toujours les plus efficaces. Typiquement, dans certains cas, la mesure d'information mutuelle est la plus performante alors même que nous avions appuyé le fait qu'elle n'était pas adaptée pour mesurer la *significativité* d'une association mais seulement sa force. Nous avons montré que les bonnes combinaisons de paramètres dépendent du corpus, de la direction de traduction et des langues impliquées, nous en verrons d'autres exemples dans les chapitres suivants. C'est un des points que nous chercherons à éclaircir par la suite : existe-t-il une mesure plus adéquate que les autres pour la tâche que nous réalisons et peut-on la motiver théoriquement ? Dans le chapitre suivant, nous garderons ces questions en tête et présenterons quelques propositions d'améliorations de l'approche directe que nous implémenterons pour exposer une série d'expériences et de résultats.

CHAPITRE 5

Alignement multilingue en corpus comparables spécialisés

Ce chapitre introduit et éprouve de nouvelles propositions pour améliorer l'extraction lexicale bilingue à partir de corpus comparables. Les trois propositions présentées (section 5.1, 5.2 et 5.3) cherchent à renforcer la caractérisation des mots à traduire, elles sont donc directement liées à notre problématique : nous montrons comment nous parvenons à trouver automatiquement de nouveaux points de comparaison pour comparer un mot source et ses candidats à la traduction, par exemple en nous appuyant sur des points d'ancrage (section 5.2) mais également en exploitant différentes sources d'information, dans notre cas les informations apportées par un alignement anglais-japonais et un alignement français-japonais (section 5.3). Nous montrons également qu'il est possible de s'appuyer sur la fréquence d'un mot à traduire pour savoir *a priori* quelles tailles de fenêtre utiliser pour construire son vecteur de contexte.

5.1 Exploitation de la fréquence des termes

Dans le chapitre 3, nous avons relaté les travaux de G. Grefenstette concernant l'acquisition sémantique. Il relève que l'exploitation de relations syntaxiques est adéquate pour caractériser efficacement les mots de fortes fréquences alors que, à l'inverse, l'utilisation d'une fenêtre contextuelle permet de mieux caractériser les mots de fréquences plus faibles (Grefenstette, 1996). Dans notre cas, nous ne disposons pas d'outils pour détecter les relations syntaxiques mais, comme le soulignent Zweigenbaum et Habert (2006), elles peuvent être approchées en utilisant des fenêtres contextuelles de taille réduite (un ou deux mots avant et après le mot vedette). En d'autres termes, l'utilisation de fenêtres contextuelles de grande taille tend à relever des *dépendances paradigmatiques* (du type *diabète… insuline*) alors que l'utilisation de fenêtres de petite taille fait apparaître des *dépendances syntagmatiques* (du type *dépistage du cancer*), plus informatives, mais nécessitant un nombre de cooccurrences plus élevé pour caractériser efficacement un mot vedette.

5.1.1 Observations

L'observation de Grefenstette (1996) se confirme sur le corpus *cancer du sein*. Bien que les meilleurs résultats globaux soient obtenus avec une taille de fenêtre de trois mots avant et après le mot vedette (cf. section 4.2), nous constatons que la qualité des résultats varie en fonction de la fréquence des mots à traduire. Ainsi, à partir de la liste d'évaluation de 648 termes, filtrés par fréquence, nous obtenons des

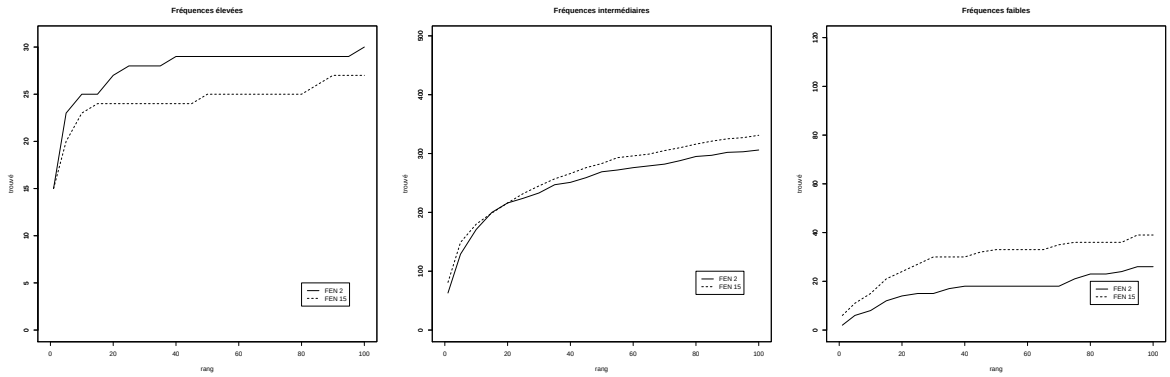


Figure 5.1 – Résultats de l'alignement anglais français, corpus *cancer du sein*. Liste [En-Fr-648]. Fréquences élevées (> 400 – à gauche), intermédiaires (au centre) et faibles (< 25 – à droite). Comparaison des résultats obtenus pour des tailles de fenêtres 2 et 15. En ordonnée, le nombre de traductions trouvées, en abscisse le rang des traductions.

résultats différents pour les mots de fréquence supérieure à 400 (31 mots), entre 400 et 25 (497 mots), et inférieure à 25 (120 mots). Les trois graphiques de la figure 5.1 soulignent cette différence.

Ces graphiques nous permettent d'observer plusieurs phénomènes, en particulier de confirmer l'observation de Grefenstette (1996) : l'utilisation de fenêtres contextuelles de taille 2 donne de meilleurs résultats pour les mots très fréquents que pour les mots rares (graphique de gauche) alors que l'utilisation de fenêtre de taille 15 donne de meilleurs résultats pour les mots rares (graphique de droite). Toutefois, les mots rares restent bien moins bien caractérisés que les mots fréquents, même en utilisant une taille de fenêtre optimale. Les mots de fréquences intermédiaires (entre 400 et 25 occurrences dans le corpus) sont sensiblement aussi bien traduits avec les deux tailles de fenêtres. Le graphique de la figure 5.2 nous permet d'affiner cette observation.

Le graphique de la figure 5.2 indique que l'utilisation d'une taille de fenêtre *intermédiaire*, de taille 10, donne des résultats sensiblement meilleurs que les petites et grandes tailles (2 et 15) pour les mots de fréquences intermédiaires. En d'autres termes, il semble possible de trouver une taille de fenêtre optimale pour un mot à caractériser, en fonction de sa fréquence, c'est-à-dire qu'il est possible de sélectionner automatiquement la taille de fenêtre optimale à partir de la fréquence du mot vedette.

5.1.2 Application

À partir des observations précédentes, nous proposons une table de correspondance entre la classe de fréquence du mot à traduire, et la fenêtre contextuelle à utiliser, décrite en table 5.1.

Cette table a été construite empiriquement à partir des observations précédentes, en cherchant à garder un équilibre entre les effectifs des classes de fréquences[1]. La plage des classes de fréquences diminue rapidement dans cette table car il y a peu de différences entre un mot construit sur 500 occurrences ou sur 600 : ils sont tous les deux caractérisés par un nombre important de cooccurrences, suffisantes pour lisser partiellement les artefacts de fréquence. À l'inverse, la différence est importante entre un mot caractérisé par 25 occurrences ou par 30 : les 5 occurrences supplémentaires peuvent faire varier de façon importante l'ensemble de ses voisins rencontrés et leurs associations. Dans notre cas, nous nous intéressons

[1] D'autres tables de correspondances donnent des résultats plus impressionants mais moins stables.

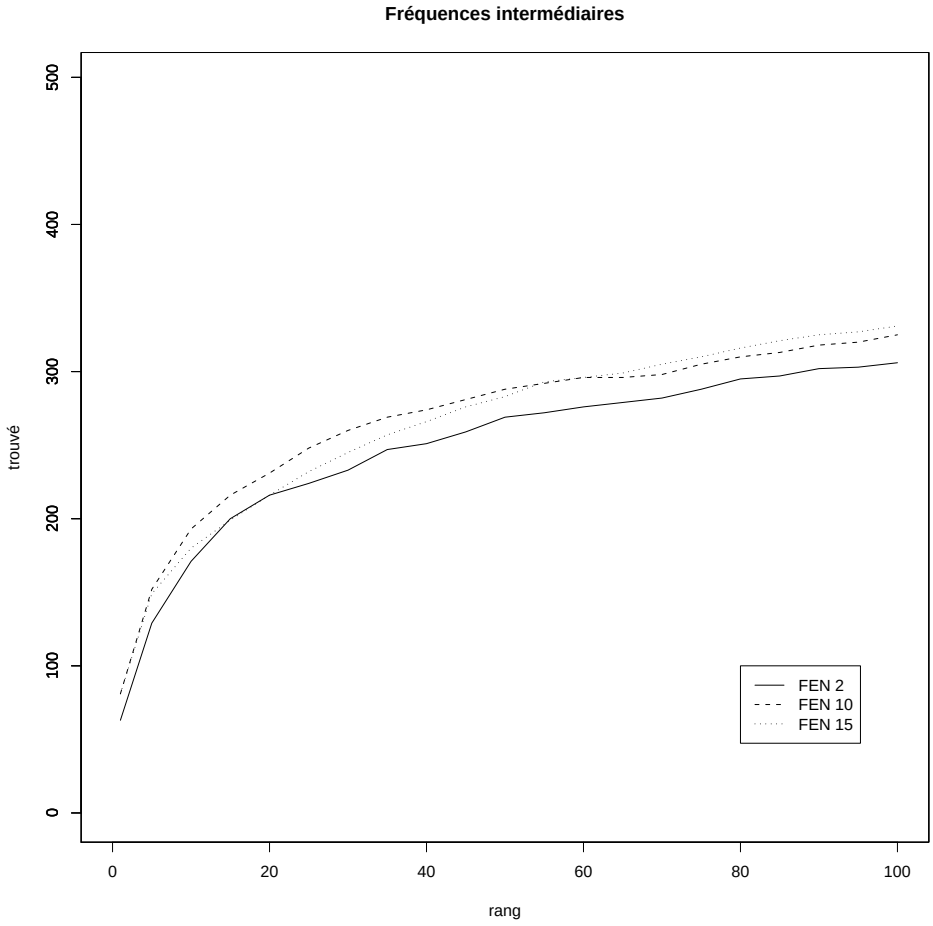


Figure 5.2 – Résultats de l'alignement anglais français, corpus *cancer du sein*. Liste [En-Fr-648]. Fréquences intermédiaires. Comparaison des résultats obtenus pour des tailles de fenêtre 2, 10 et 15.

Fréquence	Taille de fenêtre
]500–+∞[	2
]200–500]	3
]100–200]	5
]30–100]	10
]25–30]	15
[15–25]	20

Table 5.1 – Proposition de correspondances entre la classe de fréquence d'un mot et la fenêtre contextuelle optimale pour construire son vecteur de contexte.

aux mots *suffisamment fréquents* : les listes d'évaluation utilisées ne contiennent pas de mots dont la fréquence est inférieure à 20, ce pourquoi la table ne va pas plus en avant dans le découpage des classes de fréquences.

Aidé de cette table de correspondance, il est possible de sélectionner *a priori* la taille de fenêtre à utiliser pour construire le vecteur de contexte d'un mot source et la taille des vecteurs de contexte en langue cible à utiliser pour la comparaison. Dans notre cas, nous disposons déjà de l'ensemble des résultats d'alignement pour chaque taille de fenêtre (cf. section 4.2). Il nous suffit donc de sélectionner, pour un mot source et sa fréquence, les résultats de l'alignement obtenu avec la taille de fenêtre correspondante.

Nous obtenons les résultats présentés en table 5.2, en utilisant le taux de vraisemblance et la mesure de Jaccard pondérée (*Fenêtre variable*). Nous les comparons avec l'expérience donnant les meilleurs résultats (*Fenêtre fixe*), utilisant les mêmes mesures et une taille de fenêtre fixe de 3. Ces résultats sont obtenus avec la liste de référence [En-Fr-648].

	Fenêtre fixe (trouvé)	*Fenêtre variable* (trouvé) [Gain]
Top_1	13 % (83)	15 % (99) [19 %]
Top_5	27 % (175)	29 % (189) [10 %]
Top_{10}	34 % (223)	36 % (235) [5 %]
Top_{20}	41 % (263)	43 % (277) [7 %]

Table 5.2 – Résultats obtenus avec une fenêtre de taille variable en fonction de la fréquence du mot source, comparé avec l'expérience témoin. Liste de référence [En-Fr-648].

Nous constatons que l'utilisation d'une taille de fenêtre variant en fonction de la fréquence du mot source améliore la qualité des résultats, avec notamment un gain de 19 % pour le Top_1 et de 10 % pour le Top_{10}. Ces résultats confirment qu'une table de correspondance telle que celle présentée dans cette section peut être exploitée efficacement pour améliorer les résultats.

En utilisant la liste [En-Fr-122], nous trouvons des résultats encourageants bien que moins impressionants, ils sont consignés dans le tableau 5.3.

	Fenêtre fixe (trouvé)	*Fenêtre variable* (trouvé) [Gain]
Top_1	20,5 % (25)	22,1 % (27) [8 %]
Top_5	38,5 % (47)	38,5 % (47) [0 %]
Top_{10}	46,7 % (57)	46,7 % (57) [0 %]
Top_{20}	56,6 % (69)	55,7 % (68) [-1,5 %]

Table 5.3 – Résultats obtenus avec une fenêtre de taille variable en fonction de la fréquence du mot source, comparé avec l'expérience témoin. Liste de référence [En-Fr-122].

Cette expérience montre un gain intéressant pour le Top_1, mais un gain nul ou négatif pour d'autres rangs. Ces résultats peuvent s'expliquer en particulier par la taille de la liste, beaucoup plus réduite, qui fait apparaître des dégradations locales des résultats, potentiellement comblées par l'introduction de nouvelles bonnes traductions avec une liste plus volumineuse telle que la liste [En-Fr-648].

Ces résultats ne se retrouvent toutefois pas dans le cas de l'alignement français-japonais ou anglais-japonais : l'utilisation d'une taille de fenêtre variable dégrade considérablement la qualité des résultats. Nous avons observé les résultats obtenus par classe de fréquence sur le corpus *diabète et alimentation* pour l'alignement anglais-japonais[2], ils sont présentés dans la figure 5.3.

[2]Les résultats sont comparables pour l'alignement français-japonais.

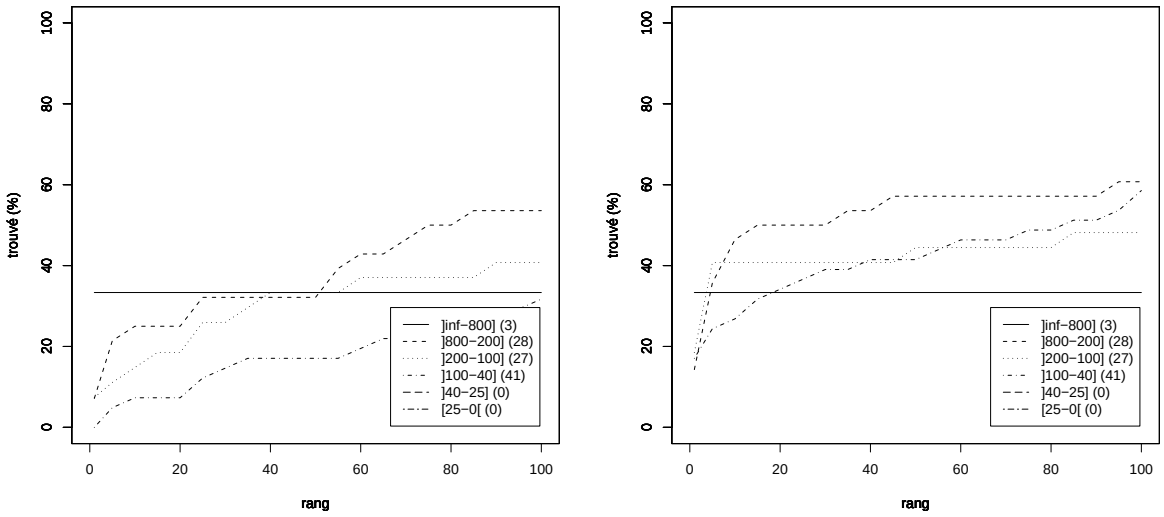


Figure 5.3 – Résultats par fréquence pour l'alignement anglais japonais, corpus *diabète et alimentation*. Liste [En-Jp-99]. Résultats obtenus pour une taille de fenêtre de 3 (à gauche) et de 25 (à droite).

La figure de gauche indique les résultats de l'alignement anglais-japonais par classe de fréquences pour une taille de fenêtre contextuelle de 3, celle de droite montre les résultats pour une taille de fenêtre de 25. Ces figures indiquent que, dans ce cas, l'hypothèse formulée précédemment n'est plus vraie : pour l'alignement anglais-japonais, les grandes tailles de fenêtre donnent des résultats meilleurs ou équivalents pour toutes les classes de fréquences, cette observation se confirme pour l'alignement français-japonais.

5.1.3 Analyses et discussion

Cette proposition montre un avantage significatif dans le cas de l'alignement français-anglais. Elle semble valider une hypothèse prometteuse, puisqu'elle permet de répondre à une des questions que nous avons posées au chapitre 4, en particulier en section 4.2. Nous y avions expliqué qu'il était difficile de prédire quelle taille de fenêtre était optimale pour un corpus donné. Nous avions effectivement noté que, dans le cas du corpus *cancer du sein*, les meilleurs résultats étaient obtenus pour de petites tailles de fenêtre alors que dans le cas du corpus *diabète et alimentation*, les plus grandes tailles de fenêtre donnaient les meilleurs résultats.

Cette proposition offre une première réponse : la taille de la fenêtre contextuelle à utiliser pour constituer les vecteurs de contexte dépend de la fréquence du mot à traduire. C'est d'autant plus intéressant que cette fréquence est une valeur absolue, indépendante de la taille du corpus utilisé (un mot qui apparait 500 fois dans un corpus très volumineux sera vraisemblablement aussi bien caractérisé qu'un mot qui apparait 500 fois dans un corpus plus modeste).

Toutefois, ce n'est pas la seule réponse possible. Pour des raisons de coût calculatoire, il est délicat de constituer *a priori*, comme nous l'avons fait, l'ensemble des vecteurs de contexte pour chaque taille de fenêtre. Notre table de correspondance s'applique donc uniquement aux paramètres que nous avons exploités, à savoir des tailles de fenêtre de 2, 3, 5, 10, 15, 20 et 25. Il est probablement possible d'affiner cette table en découpant plus finement les classes de fréquences et les tailles de fenêtre correspondantes.

Une approche plus subtile consisterait à combiner les alignements obtenus pour différentes tailles de fenêtre pour chaque classe. Par exemple, les candidats à la traduction des mots de fréquences comprises entre 500 et 200 pourraient être obtenus en combinant les résultats obtenus pour les tailles de fenêtre 2 et 3 (éventuellement en pondérant l'importance de chaque alignement avec la fréquence du mot à traduire).

Les résultats obtenus pour l'alignement anglais-japonais sont difficiles à justifier. Ils peuvent peut-être s'expliquer par plusieurs raisons. La première concerne le couple de langues impliqué. La réflexion de Zweigenbaum et Habert (2006), concernant la capacité des tailles de fenêtres contextuelles réduites à « absorber » les dépendances syntagmatiques n'est peut-être pas valide dans le cas du japonais. Une connaissance plus précise de la langue japonaise et de sa grammaire sera nécessaire pour confirmer cette impression. Cette différence peut aussi provenir de la différence dans la construction même des corpus : le corpus *cancer du sein* a été construit uniquement à partir de publications scientifiques, alors que le corpus *diabète et alimentation* regroupe des publications, mais aussi des documents courts disponibles sur internet, de conseils de médecins... Il est potentiellement *moins comparable* que le corpus *cancer du sein*. Cette notion de *comparabilité* est toutefois floue et difficile à évaluer qualitativement et quantitativement, nous y reviendrons dans le chapitre 6. La section suivante cherche à améliorer spécifiquement l'alignement avec la langue japonaise, notamment en s'appuyant sur une terminologie connue *a priori* pour renforcer la caractérisation des termes à traduire.

5.2 Points d'ancrage

Nous l'avons montré dans le chapitre 1, l'alignement de lexique bilingue à partir de corpus comparables repose sur les vecteurs de contexte qui caractérisent l'environnement des termes à aligner. Les vecteurs de contexte doivent être construits de manière à représenter, d'une façon aussi discriminante que possible, un terme donné. Ainsi, et si la généralisation de l'hypothèse de Firth[3] dans un cadre multilingue est avérée, les contextes d'un terme t et de ses traductions $T(t)$ seront similaires, alors que les contextes de t et des autres éléments, n'étant pas des traductions, seront différents. Il est alors crucial que ces vecteurs de contexte soient les plus représentatifs d'un terme donné. Cette question est encore plus critique dans le cas de corpus comparables de tailles modestes. Dans le cas des corpus volumineux (millions de mots), les artefacts de fréquence sont lissés par la masse de matériau textuel. À l'inverse, dans les petits corpus (centaines de milliers de mots), l'orientation d'un seul document peut influencer largement les fréquences de certains termes. Par exemple, un document peut introduire des termes peu significatifs, mais de fréquence similaire ou supérieure à certains termes plus significatifs qui, par le hasard de la distribution, sont peu représentés dans le corpus.

5.2.1 Propriétés

Nous proposons dans cette partie d'étudier et d'exploiter des points d'ancrage dans les vecteurs de contexte, c'est-à-dire des éléments de confiance que nous pouvons utiliser pour augmenter la force de discrimination des vecteurs de contexte. Cette notion est proche de celle introduite initialement dans les corpus parallèles (voir section 1.3.2), c'est-à-dire des éléments lexicaux ou structurels alignés avec confiance et sur lesquels les méthodes peuvent s'appuyer pour réduire l'espace de recherche dans le but d'aligner leurs voisins. Dans notre cas, les points d'ancrage ne sont plus utilisés au niveau du corpus mais au niveau des vecteurs de contexte.

En pratique, ces points d'ancrage doivent avoir plusieurs propriétés :

[3] cf. section 1.4.1.1.

1. Ils doivent être identifiables automatiquement.
2. Ils doivent être pertinents, relativement aux thèmes des documents à aligner. En d'autres termes, ils doivent couvrir une sous-partie (connue) du vocabulaire spécialisé.
3. Ils doivent être peu polysémiques.

La première propriété, motivée par l'application, permet d'exploiter ces points d'ancrage dans le cadre d'un processus automatique : à partir de critères d'identification, il faut être capable de les extraire et de les intégrer au processus de reconnaissance sans intervention manuelle lourde. La deuxième propriété assure que les points d'ancrage sont représentatifs : Habert *et al.* (1997) remarquent que *en règle générale, on constate que plus les noms sont techniques et fréquents, meilleure est leur description*. Les éléments respectant la deuxième propriété sont *fréquents*, car liés aux thématiques du corpus ; ils sont par ailleurs *techniques*, car représentatifs de la terminologie du domaine des documents. La troisième propriété assure que l'utilisation des points d'ancrage ne rajoutera pas de nouvelles ambiguïtés. Ces trois propriétés assurent que les points d'ancrage ainsi sélectionnés sont des éléments de confiance : il est nécessaire que chaque point d'ancrage sélectionné soit fiable, puisque nous nous appuierons dessus de façon importante. Il ne faut donc pas qu'ils introduisent plus de bruit dans les vecteurs de contexte. Par ailleurs, cet objectif implique que les points d'ancrage soient des paires de traductions, sans quoi ils ne franchiront pas l'étape de transfert des vecteurs sources dans l'espace cible. Ils seront inexploités dans la langue source et déséquilibreront les vecteurs de contexte en langue cible. Nous émettons l'hypothèse que ces points d'ancrage sont des éléments discriminants dans la caractérisation des contextes des termes.

Nous avons présenté en section 2.3 le phénomène des translittérations japonaises qui respectent les trois propriétés énoncées précédemment. Ces translittérations sont notre premier choix de points d'ancrage. Nous étudions également dans la section 5.2.3 le cas des composés savants, c'est-à-dire les mots construits sur des racines grecques et latines spécifiques.

5.2.2 Les translittérations comme points d'ancrage

L'étude du chapitre 2 sur les translittérations fait apparaître qu'elles sont de bons candidats pour être des points d'ancrage en accord avec les propriétés énoncées en introduction de cette section. En effet, elles sont faciles à identifier automatiquement étant écrites dans un syllabaire quasiment dédié à cet usage et elles peuvent être alignées avec leurs correspondances dans les langues *source*. Elles sont par ailleurs révélatrices d'un vocabulaire particulier et sont employées dans un contexte particulier. Enfin, elles sont peu polysémiques dans le cadre d'un vocabulaire spécialisé étant généralement forgées pour correspondre à un unique mot en langue source.

Nous souhaitons extraire automatiquement les translittérations à partir du corpus *diabète et alimentation* pour les exploiter en tant que points d'ancrage dans le processus d'alignement. En effet, les méthodes de détection des translittérations retournent des couples de candidats (l'extraction des séquences en katakana est triviale et n'a que peu d'intérêt si elles ne sont pas alignées avec leur correspondance en langue *source*). Ainsi, le processus de détection permet de construire automatiquement des listes de paires de points d'ancrage utilisables par la suite.

Nous avons utilisé un outil réalisant la détection automatique de translittérations entre l'anglais et le japonais (Tsuji *et al.*, 2005). Bien que plus simple que l'algorithme proposé par Knight et Graehl (1997) (voir section 2.6.2), cet outil, basé sur des *chaînes de Markov* donne des résultats satisfaisants. Il génère un ensemble de correspondances potentielles pour une entrée en katakana ou en anglais. Les résultats doivent alors être comparés avec un vocabulaire existant pour sélectionner les candidats les plus probables. Cet outil a d'abord été conçu pour que les résultats soient comparés avec des requêtes sur

le Web. Puisque nous travaillons sur des corpus comparables, l'ensemble du vocabulaire cible est plus réduit, mais pas forcément aussi bien représenté. Une paire de translittérations est détectée entre un mot anglais m_a et un mot japonais m_j lorsque l'une des conditions suivantes est réalisée :

- m_j existe dans le corpus japonais et a été généré par l'outil comme une correspondance potentielle de m_a, ;
- m_a existe dans le corpus anglais et a été généré par l'outil comme une correspondance potentielle de m_j.

La relation ainsi définie est symétrique.

Nous avons dans un premier temps utilisé un outil dédié à la détection de translittérations entre le français et le japonais (modèle à base de règles) mais avons obtenu de mauvais résultats, en particulier un grand nombre de faux-positifs. Nous avons toutefois mis en évidence dans la section 2.4.4 que, bien que les translittérations directes entre le français et le japonais soient rares, beaucoup de translittérations japonaises issues de l'anglais peuvent être alignées avec des mots français, en raison de relations de cognats fréquentes entre ces deux langues. Aussi avons-nous choisi de réaliser la détection de translittérations entre le français et le japonais en utilisant l'outil dédié à la détection anglais-japonais. Avant traitement, les termes français à comparer sont lissés pour faire disparaître les signes diacritiques spécifiques (mais leur trace est conservée dans les couples alignés). Nous obtenons 356 translittérations japonaises et 1 312 paires de candidats français-japonais. Notons que, en raison de fautes d'orthographe dans le corpus français, certaines translittérations japonaises sont correctement alignées avec plusieurs mots français (par exemple コレステロール – `ko-re-su-te-ro-o-ru` est aligné avec *cholestérol*, *cholesterol* et *choléstérol*, l'orthographe correcte étant *cholestérol*).

À la fin du processus d'identification et d'alignement des translittérations, nous obtenons 589 paires de translittérations pour le couple anglais-japonais et 526 pour le couple français-japonais à partir du corpus « *diabète et alimentation* ». Ces paires seront utilisées comme points d'ancrage et seront également ajoutées en complément des dictionnaires bilingues, de manière à ce que ces points d'ancrage puissent franchir l'étape de traduction des vecteurs de contexte.

5.2.3 Composés savants

Nous nous sommes également penchés sur les *composés savants*. Il s'agit de mots, en français et en anglais, construits à partir de racines spécifiques (Namer, 2005). Claveau (2007), s'intéressant à la traduction automatique de termes biomédicaux, observe que « *les termes biomédicaux sont construits sur les mêmes racines grecques et latines, et leurs dérivations très régulières* » (p. 2). Ces composés dénotent un vocabulaire spécialisé notamment dans le domaine médical (Lovis *et al.*, 1997 ; Namer et Zweigenbaum, 2004). Ce sont donc des points d'ancrage pertinents dans le cas d'un corpus spécialisé sur le *diabète et l'alimentation* tel que le nôtre. En outre, ils peuvent être facilement identifiés à partir de leur morphologie.

En ce qui concerne la détection des composés savants, nous nous sommes appuyés sur une liste de 606 racines et affixes médicaux utilisés en anglais[4]. Le processus d'extraction est trivial : en compilant une expression régulière par préfixe et par suffixe, il cherche les mots anglais correspondants dans les dictionnaires bilingues utilisés pour l'alignement. Les mots extraits sont conservés ainsi que leurs traductions en japonais pour obtenir des paires de traductions utilisées comme points d'ancrage dans les vecteurs de contexte. La liste des affixes a été conçue pour l'anglais, mais elle peut facilement être traduite en français en accord avec la remarque de Claveau (2007). Nous nous sommes inspirés de ce travail pour écrire quelques règles simples de conversion. La terminaison `-y` (comme dans *psychology*)

[4] `www.medo.jp/a.htm`

est par exemple transformée en `-ie` en français (*psychologie*). Certains affixes retournent beaucoup de paires de traductions qui ne correspondent pas nécessairement à des racines grecques ou latines (typiquement le préfixe `a-`). De plus, les mots correspondants extraits ne sont pas toujours formés à partir de ces préfixes (par exemple, le `a-` de *armoire*). Tous les affixes générant plus de 1 000 correspondances sur les ressources ont été écartés pour retirer les moins pertinents. Ils sont toutefois assez rares et 12 seulement ont été écartés pour l'anglais, 17 pour le français. Nous avons ainsi obtenu 17 210 composés savants en anglais correspondant à 60 341 traductions (les ressources linguistiques comprennent des traductions multiples pour un seul élément source). Nous avons également obtenu 8 254 composés savants français, soit 24 240 traductions. Ces différences de résultats proviennent principalement de la nature des dictionnaires bilingues utilisés dans chaque couple de langues. À l'inverse des translittérations, la détection des composés savants ne permet pas d'inférer leur traduction par un processus automatique. C'est la raison pour laquelle l'étape de détection est réalisée directement sur les dictionnaires bilingues de façon à obtenir des paires de points d'ancrage traductions.

5.2.4 Exploitation des points d'ancrage

Nous avons choisi de modifier l'approche directe en accordant plus d'importance aux points d'ancrage lors du calcul de l'association entre la tête d'un vecteur et ses éléments. L'objectif étant que la comparaison des vecteurs se fasse en priorité sur les points d'ancrage, puis sur les éléments moins significatifs. Après avoir calculé l'association de façon classique (cf. section 1.5.2), nous rehaussons le score des points d'ancrage et diminuons le score des autres éléments de manière à ce que les sommes des scores initiaux et finaux soient identiques (voir équations 5.1 à 5.3). Dans ces équations, PA est l'ensemble des points d'ancrage extraits ($|PA|_l$ le nombre de points d'ancrage trouvé dans le vecteur de contexte l et $|\neg PA|_l$ le cardinal des autres éléments), $assoc_j^l$ est la mesure d'association de l'élément j dans le vecteur de contexte du mot l.

$$assoc\_pondérée_j^l := assoc_j^l + \beta,\ si\ j\ \in\ PA \tag{5.1}$$

$$assoc\_pondérée_j^l := assoc_j^l - décalage_l,\ si\ j\ \notin\ PA \tag{5.2}$$

$$décalage_l := \frac{|PA|_l}{|\neg PA|_l} \times \beta \tag{5.3}$$

Le paramètre β permet de calibrer l'importance donnée aux points d'ancrage. Ce paramètre est ajouté de façon absolue au score de chaque point d'ancrage et non de façon proportionnelle par rapport au score initial. En effet, nous souhaitons rehausser le score de tous les points d'ancrage de manière à ce qu'ils soient tous pris en compte de façon significative par les mesures de similarité, plus ou moins indépendamment de leurs scores initiaux.

Le choix du paramètre β influence grandement les résultats de l'alignement : s'il est trop faible, il ne rendra pas les points d'ancrage suffisamment importants dans le calcul de la similarité ; s'il est trop élevé, il risque d'écraser le poids des autres éléments et leurs potentiels de discrimination (et ne faire plus reposer l'alignement que sur les éléments de confiance). Il doit être calibré en fonction :

- du nombre de points d'ancrage susceptibles d'être présents dans un vecteur de contexte ;
- du type des points d'ancrage utilisés ;
- de la confiance qui doit leur être accordée ;

– des mesures d'association utilisées, puisqu'elles ne retournent pas toutes les mêmes intervalles de valeurs.

Pour mieux comprendre le choix de cette modification et son application, reprenons l'image des motifs d'association présentés dans les sections 1.5.2 à 1.5.4. La figure 5.4 présente graphiquement la façon dont sont exploités les points d'ancrage. La première étape consiste à les identifier dans les vecteurs puis à les incrémenter de la valeur β.

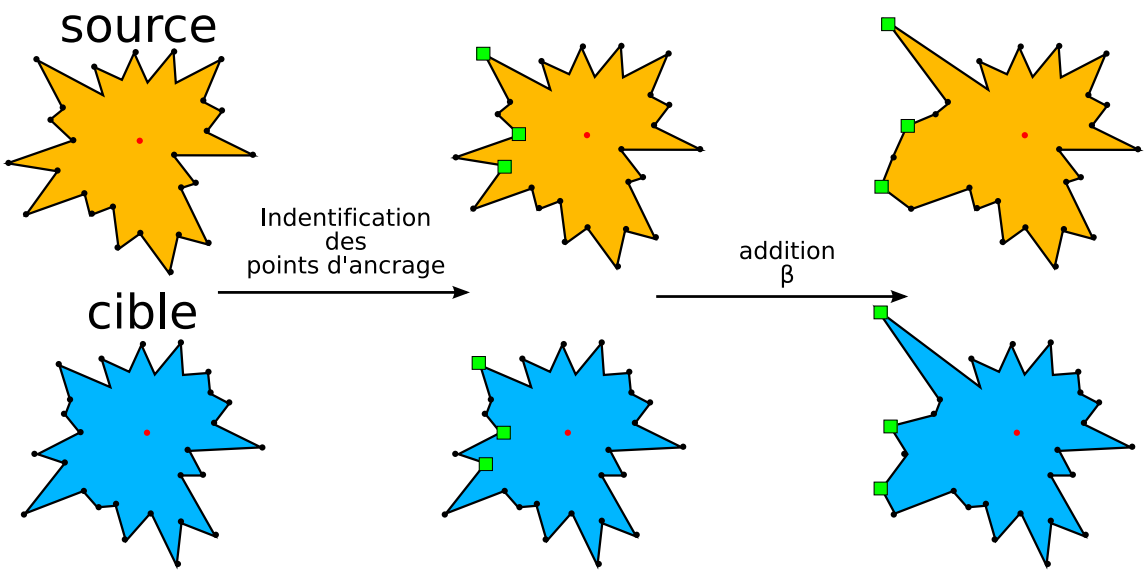


Figure 5.4 – Exploitation des points d'ancrage.

Cette figure illustre le choix que nous avons fait : il ne s'agit pas seulement de rendre les motifs de deux vecteurs en relation de traductions *plus similaires* (et les expériences révèlent d'ailleurs que ce n'est pas toujours le cas, voir section 5.2.7), il s'agit aussi de rendre les motifs de vecteurs qui ne sont pas traductions *plus différents*. En effet, les vecteurs qui ne contiennent pas les points d'ancrage identifiés précédemment ne seront pas déformés de la sorte et seront différenciés des vecteurs les contenant. Nous nous appuyons donc bien sur les points d'ancrage pour *discriminer* les vecteurs de contexte.

Nous avons réalisé plusieurs expériences pour évaluer l'efficacité et l'impact des points d'ancrage dans l'alignement lexical bilingue :

– (*a*) approche directe standard ;
– (*b*) en utilisant les translittérations détectées automatiquement ;
– (*c*) en utilisant les composés savants extraits automatiquement.

Toutes les expériences sont réalisées avec les mêmes paramètres : le taux de vraisemblance et le cosinus, en utilisant une taille de fenêtre de 25. Ce sont les paramètres qui donnent les meilleurs résultats dans le cas de l'expérience témoin.

5.2.5 Résultats

Le tableau 5.4 synthétise les résultats obtenus pour les expériences *a*, *b* et *c* pour les Top 1 et 10, pour l'alignement anglais-japonais et français-japonais (entre crochets, le gain obtenu).

Les résultats pour l'expérience de contrôle (exp. *a*) sont comparables aux résultats obtenus par Chiao et Zweigenbaum (2002) discutés en section 1.5.5 mais sont moins bons que les résultats obtenus pour

	a	*b*	*c*
anglais-japonais (Top_1)	17,1 %	20,2 % [18,2 %]	20,2 % [18,2 %]
anglais-japonais (Top_{10})	36,3 %	39,3 % [8,2 %]	40,4 % [11,2 %]
français-japonais (Top_1)	20,4 %	20,4 % [0,0 %]	22,4 % [10,0 %]
français-japonais (Top_{10})	36,7 %	37,8 % [2,8 %]	38,8 % [5,6 %]

Table 5.4 – Résultats de l'alignement anglais-japonais et français-japonais ($\beta = 8$) ; *a* : expérience témoin ; *b* : utilisation des translittérations ; *c* : utilisation des composés savants.

l'alignement avec le corpus *cancer du sein* pour une liste d'évaluation construite de la même façon (sections 4.2 et 5.1). Dans le cas de l'anglais, le gain obtenu en nous appuyant sur les points d'ancrage est important : à hauteur de 18 % en utilisant les translittérations (exp. *b*) et les composés savants (exp. *c* – Top_1). Le gain est moins important pour l'alignement français-japonais : il est nul pour le Top_1 en utilisant les translittérations et atteint 10 % en utilisant les composés savants. La qualité plus faible des résultats avec le français peut s'expliquer par la moins bonne qualité des listes de points d'ancrage. En particulier, les translittérations ont été extraites avec un outil dédié à la détection entre l'anglais et le japonais, de plus les translittérations entre le français et le japonais sont plus rares.

5.2.6 Discussion

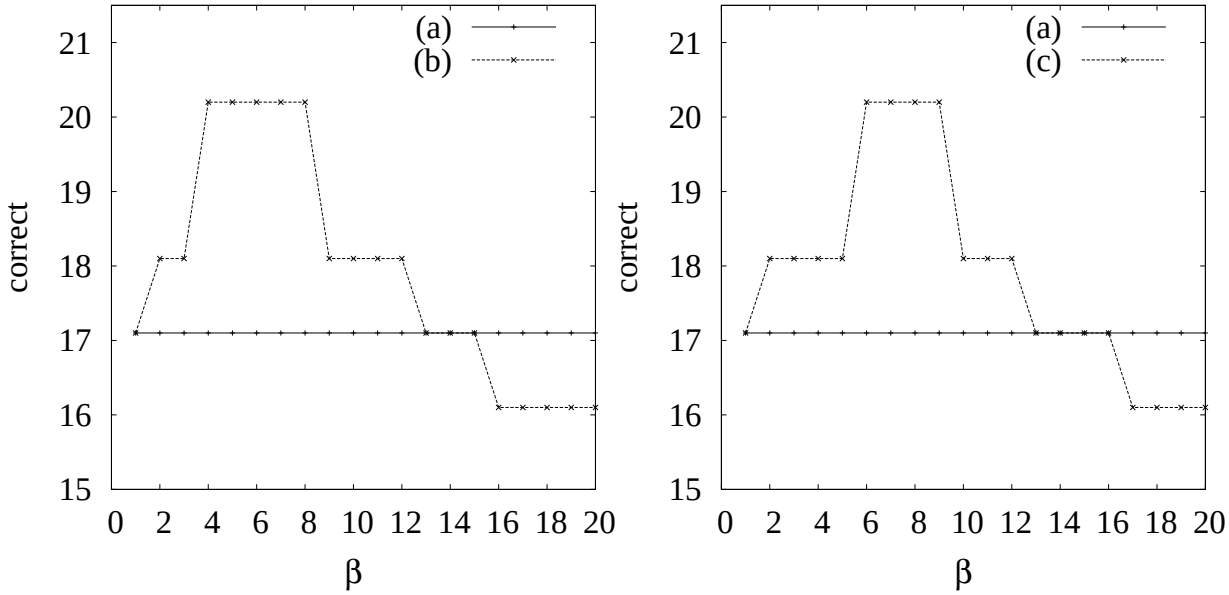
Les résultats que nous avons présentés dans le tableau 5.4 sont ceux obtenus avec le paramètre β le plus favorable. Ils montrent que l'utilisation de points d'ancrage peut contribuer à l'amélioration des résultats de l'alignement. Nous avons toutefois observé un phénomène intéressant relatif à la variation du paramètre β et des listes de mots utilisés comme points d'ancrage. Les figures 5.5(a) et 5.5(b) indiquent les résultats obtenus pour l'alignement anglais, par rapport à l'expérience témoin, pour le Top_1 en faisant varier le paramètre β de 0 à 20 (le résultat de l'expérience témoin est constant et ne dépend pas du paramètre β).

Ces figures montrent que dans le cas des points d'ancrage sélectionnés (fig. 5.5(a) et 5.5(b)), les résultats varient selon une cloche autour du paramètre β le plus favorable. Ce phénomène se reproduit de façon similaire avec l'alignement français-japonais, mais également en utilisant d'autres mesures d'association ou de similarité (bien qu'elles donnent des résultats sensiblement moins bons) ou d'autres tailles de fenêtre. Cette observation est intéressante car elle confirme notre hypothèse : certains mots ont un statut différent dans l'alignement. Leur mise en évidence influence fortement la qualité des résultats. De plus, au-delà des Top 1 et 10, l'utilisation des points d'ancrage a une influence sur l'ensemble des traductions candidates, nous le montrons dans la section suivante.

5.2.7 Influence des points d'ancrage

La figure 5.6 compare les résultats obtenus entre l'expérience témoin et l'utilisation des composés savants, dans le cas de l'alignement français-japonais ($\beta = 8$). En effet, le tableau 5.4 semble indiquer que l'apport des points d'ancrage dans le cas de l'alignement français-japonais n'est pas aussi significatif que dans le cas de l'alignement anglais-japonais. Cette figure présente l'évolution des positions des traductions correctes dans les listes de candidats obtenues à la fin du processus d'alignement (en ordonnée), ainsi que de leurs scores de similarité (en abscisse).

Les triangles vides représentent les traductions n'étant plus obtenues avec l'utilisation des composés savants alors que les triangles noirs indiquent les nouvelles traductions obtenues, indisponibles dans le



(a) expérience *b* (translittérations) (b) expérience *c* (composés savants)

Figure 5.5 – Influence du paramètre β, comparé à l'expérience témoin. Alignement anglais-japonais.

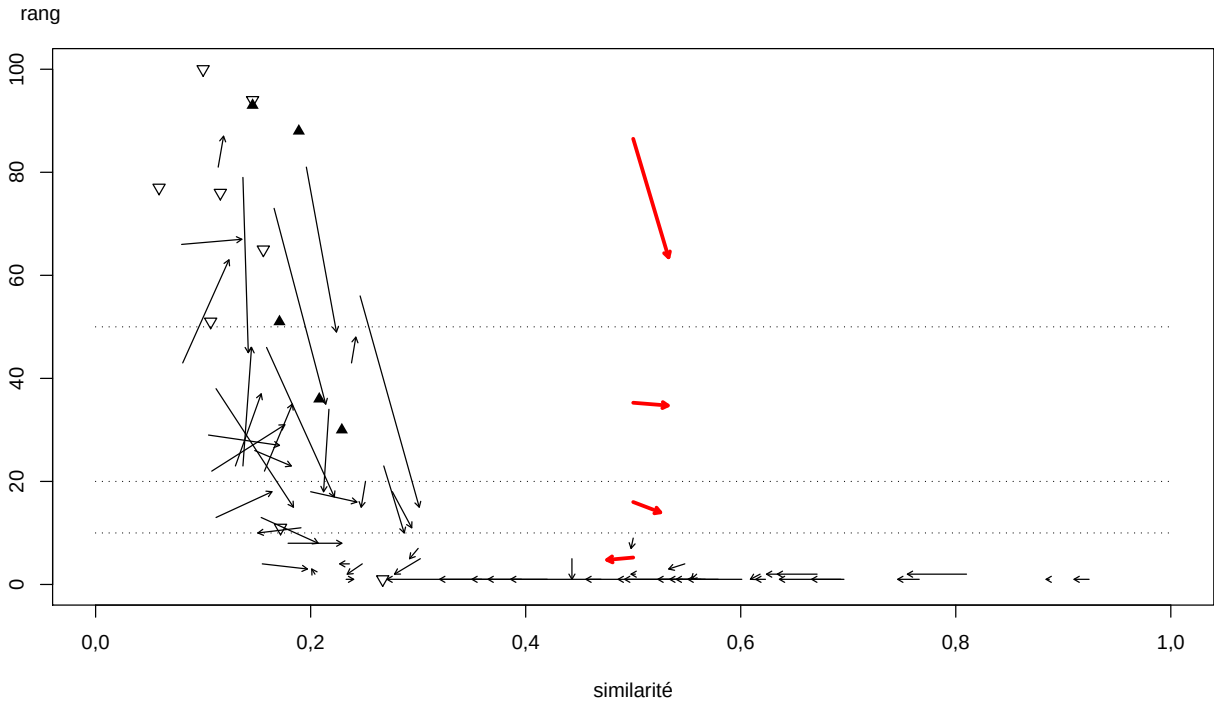


Figure 5.6 – Rangs et scores des traductions correctes pour l'alignement français-japonais, avec et sans utilisation des points d'ancrage (composés savants – $\beta = 8$).

cas de l'expérience témoin. Les flèches fines représentent le déplacement d'une traduction entre l'expérience témoin (début de la flèche) et l'expérience utilisant des points d'ancrage (pointe de la flèche). Enfin, les quatre flèches plus épaisses sont la somme des flèches fines pour chaque zone délimitée par des pointillés.

Cette figure montre d'abord que le nombre de traductions introduites est proche du nombre de traductions disparues. Elles correspondent à des traductions instables, étant très sensibles aux différents paramètres utilisés (taille de la fenêtre, mesure d'association et de similarité...). Les flèches épaisses permettent de mieux comprendre l'influence des points d'ancrage. Elles indiquent en effet que, en moyenne, les traductions correctes obtiennent un meilleur rang dans les résultats de l'alignement. C'est particulièrement visible pour les traductions initialement mal classées (Top_{50} à Top_{100}). Leur rang est largement amélioré comme l'indique la somme des vecteurs pour cette zone. Ce constat est valable pour les autres zones, même s'il est moins flagrant. Dans tous les cas, en moyenne, l'utilisation des points d'ancrage améliore le classement des traductions correctes dans la liste des candidats obtenus. Toutefois, les traductions initialement correctement alignées (Top_{10} ou inférieur) sont peu reclassées (elles ne sont toutefois pas désavantagées même si leur indice de similarité moyen baisse). Ces observations viennent compléter les résultats présentés : elles montrent une tendance au réarrangement des candidats à la traduction vers des positions plus avantageuses, quelque soit leur rang initial, malgré une amélioration des Top 1 et 10 peu importante.

5.3 Alignement multi-sources

5.3.1 Hypothèse

Une autre proposition d'amélioration s'inspire à la fois des travaux de Chiao (2004), sur l'hypothèse de symétrie distributionnelle (cf. 1.6.2) et de travaux en traduction automatique sur l'exploitation de sources sources. Un des problèmes qui se pose dans la traduction automatique est la désambiguïsation lexicale. Typiquement, le mot *livre* en français se traduit en anglais par *book*, lorsqu'il s'agit d'un ouvrage écrit, et par *pound* lorsqu'il s'agit d'une unité de mesure. Dans certains cas, l'ambiguïté peut être levée en étudiant le contexte : dans la phrase *j'ai lu un livre*, il est peu probable qu'on se réfère à l'unité de mesure. Dans d'autre cas, l'ambiguïté persiste, même pour un humain, par exemple dans *je me suis fait voler deux livres*. Och et Ney (2001) proposent de s'appuyer sur des traductions déjà connues d'un document pour désambiguïser ces situations. Par exemple si l'on veut traduire un document français en anglais, et que l'on dispose déjà de sa traduction en espagnol, l'ambiguïté *book/pound* sera levée si la traduction de *livre* dans le document espagnol est *libro* (ouvrage écrit) ou *libra* (unité de mesure). Ils proposent donc de renforcer la traduction de chaque langue à partir des traductions des autres langues. Crego *et al.* (2009) exploitent cette idée et propose même un cheminement optimal pour traduire chaque document, c'est-à-dire, l'ordre des langues dans lequel traduire chaque document pour améliorer la traduction dans d'autres langues. Ils constatent, dans le cas du corpus EUROPARL, que l'espagnol et le suédois sont les langues les plus utiles pour traduire du français vers l'anglais. Ces propositions (et la nôtre) font notamment écho au travaux de Dagan *et al.* (1991) : ils montrent qu'il est possible de tirer avantage de l'information apportée par plusieurs langues dans le traitement d'une langue tierce.

Notre proposition consiste à s'appuyer sur les résultats des alignements du français vers le japonais, et de l'anglais vers le japonais pour améliorer la qualité des deux, à partir d'une liste d'équivalence français-anglais. Pour un mot i, connu en français et en anglais par exemple, *insuline*/*insulin*, il s'agit d'analyser le rang des candidats à la traduction, obtenu pour *insuline* et *insulin* pour les réorganiser en utilisant la moyenne harmonique de leur rang, équation 1.1 rappelée ici :

$$MH(r_{en}, r_{fr}) = \frac{1}{\frac{1}{2}(\frac{1}{r_{en}} + \frac{1}{r_{fr}})} = \frac{2r_{en}r_{fr}}{r_{en} + r_{fr}} \tag{5.4}$$

Le principe est d'une part d'éliminer les éléments que l'on ne retrouve pas dans les candidats à la traduction pour chaque langue et d'autre part de renforcer la position des bons candidats, que l'on trouve dans les deux alignements français-japonais et anglais-japonais. Des méthodes de combinaisons de sources d'informations de ce type se retrouve en RI et tentent d'exploiter trois phénomènes (Vogt et Cottrell, 1998) :

- l'effet de *chœur*[5] se produit lorsque différentes approches font toute apparaître un candidat bien classé à une même requête, ce qui laisse penser que ce candidat est pertinent ;
- l'effet d'*écrémage*[6] se produit lorsque différentes approches donnent des résultats différents. Cet effet est exploité en récupérant les candidats les mieux classés de chaque approche, en espérant que les meilleurs candidats seront trouvés par au moins une approche ;
- l'effet *candidat de l'ombre*[7] se produit lorsque seule une approche classe étonnamment bien un candidat correct.

Dans notre cas, nous exploiterons en particulier l'effet de chœur : si un bon candidat est bien classé pour les alignements français-japonais et anglais-japonais, il sera retenu en bonne position dans la combinaison des deux candidats en utilisant la moyenne harmonique. Nous exploiterons indirectement l'effet d'écrémage : si un candidat est bien classé dans un alignement et mal dans un autre, il obtiendra un rang moyen. Toutefois la moyenne harmonique a tendance à favoriser les candidats bien classés (le candidat moyen ne sera pas à équidistance des rangs précédents[8]). En revanche, nous ne tirons partie de l'effet « candidat de l'ombre » que pour les candidats corrects qui se retrouvent dans les deux alignements d'origine : si un candidat apparaît bien classé pour un alignement mais est manquant pour un autre, il sera ignoré.

5.3.2 Observation

Nous avons dans un premier temps conçu une nouvelle liste d'évaluation [En-Fr-Jp-89] contenant des entrées équivalentes en anglais et en français (chaque entrée de la liste correspond à un ensemble de traduction en anglais, français et japonais). C'est une projection de la liste [En-Jp-99] sur la partie française du corpus. Il est intéressant de noter que certains termes présents dans la liste [En-Jp-99] ne se retrouvent pas dans le corpus français.

Nous avons observé le rang et la présence ou non des traductions entre les alignements anglais-japonais et français-japonais. Ces éléments sont comparés en figure 5.7.

Cette figure s'interprète de la même manière que la figure 5.6. Les triangles noirs (respectivement, triangles blancs) correspondent aux traductions trouvées pour l'alignement anglais-japonais (resp. français-japonais) mais pas pour l'alignement français-japonais (resp. anglais-japonais). Les flèches correspondent à l'évolution du rang et de la similarité des traductions obtenues : le début de la flèche indique la position d'une traduction obtenue pour l'alignement anglais-japonais, la pointe de la flèche indique la position de la traduction dans l'alignement français-japonais.

[5] *Chorus effect.*
[6] *Skimming effect.*
[7] *Dark horse effect.*
[8] Par exemple, si un candidat est classé en première position pour un alignement et en dixième position pour un autre, son rang calculé par la moyenne harmonique sera de 1,8 environ.

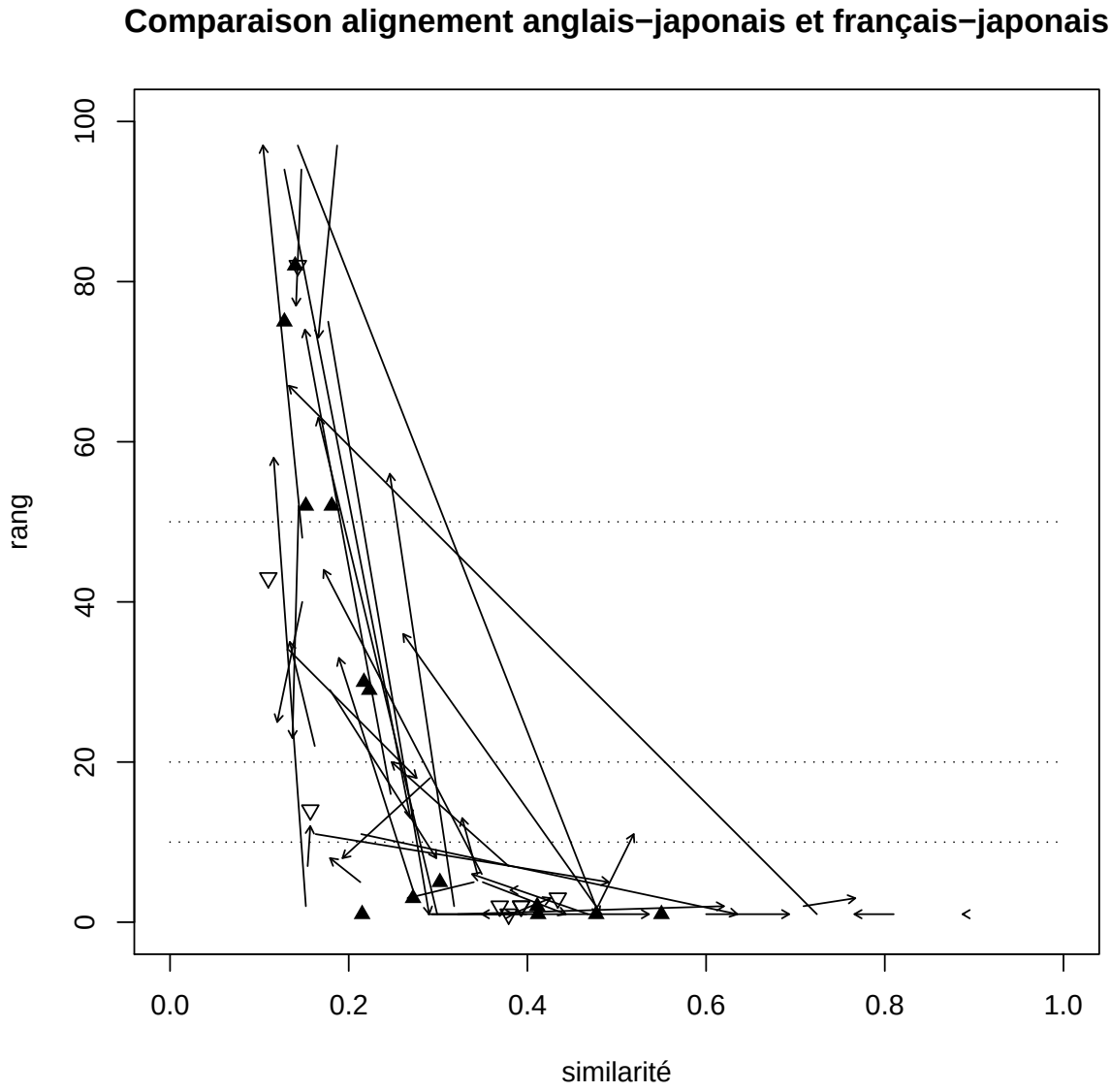


Figure 5.7 – Comparaison du rang et de la similarité des traductions obtenues pour l'alignement anglais-japonais et français-japonais.

Cette figure indique que de nombreuses traductions sont obtenues avec les deux alignements mais que leurs rangs et similarités varient beaucoup, sans tendance visible : dans certains cas, l'alignement anglais-japonais donne de meilleurs résultats que l'alignement français-japonais, mais dans d'autres cas, l'alignement français-japonais l'emporte. Notre objectif est de tirer parti de ces différences de résultats en prenant le meilleur de chaque alignement dans le but d'obtenir un alignement français-anglais-japonais de meilleure qualité.

5.3.3 Expérience

Nous avons testé cette proposition sur le corpus trilingue *diabète et alimentation*. À partir des résultats obtenus dans les conditions optimales avec les alignements anglais-japonais et français-japonais, nous recombinons le rang des candidats à la traduction en utilisant la *moyenne harmonique* présentée précédemment. Nous reprenons les résultats obtenus pour les paramètres optimaux pour l'expérience témoin, à savoir les mesures du taux de vraisemblance et du cosinus et une taille de fenêtre de 25 pour la constitution des vecteurs de contexte.

La table 5.5 consigne les résultats obtenus pour l'alignement anglais-japonais, français-japonais (témoins) et l'alignement multi-sources en utilisant les listes d'évaluation [En-Fr-Jp-89]. Le gain pour l'alignement multi-sources est calculé sur le meilleur résultat des autres alignements.

	Anglais-Japonais	Français-Japonais	Multi-sources [Gain]
Top_1	14,6 % (13)	15,7 % (14)	21,3 % (19) [35,7 %]
Top_5	30,3 % (27)	28,0 % (25)	37,0 % (33) [22,0 %]
Top_{10}	34,8 % (31)	32,6 % (29)	40,4 % (36) [16,1 %]
Top_{20}	40,4 % (36)	39,3 % (35)	46,1 % (41) [13,9 %]

Table 5.5 – Résultats de l'alignement multi-sources comparés aux résultats initiaux.

Les résultats présentés dans cette table sont bons : ils dépassent dans tous les cas les meilleurs résultats pour les autres alignements pris séparément : le gain pour le Top_1 est élevé et les résultats obtenus pour le Top_{10} avec l'alignement multi-sources correspondent aux résultats obtenus pour le Top_{20} avec les autres expériences.

5.3.4 Discussion

Cette approche est intéressante dans un cadre applicatif : si l'on dispose d'un lexique bilingue suffisamment complet pour un couple de langues, il est possible d'inférer efficacement les traductions vers une troisième langue à partir d'une combinaison d'alignement. En cela, cette méthode est conceptuellement proche de la désambiguisation multi-sources (Dagan *et al.*, 1991 ; Och et Ney, 2001). De plus, cette proposition semble particulièrement pertinente dans le cas des petits jeux de données utilisés. Autant il reste toujours délicat de constituer des corpus volumineux pour une langue donnée avec les contraintes que nous avons imposées sur le type de discours et la thématique, autant il est envisageable de multiplier les sources de données, en créant plusieurs « petit » corpus dans de nombreuses langues. C'est la combinaison des informations apportés par chacune de ces langues qui permet d'améliorer les résultats avec cette approche, et non une meilleure exploitation de chaque langue prise séparément.

Le défaut de cette méthode est qu'elle nécessite non seulement une variété de jeux de données source, mais également une bonne connaissance des relations entre ces jeux de données. En d'autre terme, elle fonctionne dans le cas que nous présentons ici parce que nous exploitons des traductions connues entre

l'anglais et le français. Par ailleurs, cette proposition ne va être efficace que sur les éléments qui ont effectivement été trouvés dans les alignements sources (français-japonais et anglais-japonais dans notre cas) : les traductions manquantes ne sont pas prise en compte dans l'étape de combinaison (les triangles noirs et blancs de la figure 5.7). Les traductions trouvées dans les alignements sources sont toutefois signicativement réorganisées vers des rangs plus élevés.

5.4 Conclusion

Nous avons présenté trois propositions visant à renforcer la caractérisation terminologique multilingue dans le but d'extraire un lexique multilingue à partir de corpus comparables. La première proposition exploite la fréquence des termes pour savoir quelle approche utiliser pour les caractériser en faisant varier la taille de la fenêtre contextuelle des vecteurs. Elle s'est avérée efficace pour la caractérisation des termes, mais également pour caractériser un lexique plus générique. Elle semble toutefois inutile dans le cas de l'alignement avec la langue japonaise qui présente des caractéristiques très différentes du français ou de l'anglais. Le cas du japonais a été traité en particulier avec les deux autres propositions.

Le corpus *diabète et alimentation* est moins volumineux que le corpus *cancer du sein* : l'alignement est de moins bonne qualité que dans le cas anglais-français, ce qui s'explique à la fois par la nature du corpus (sa taille, la façon dont il a été construit) et par les couples de langues impliqués. Nous avons donc choisi d'améliorer la caractérisation des termes en nous appuyant sur une terminologie déjà connue. Nous pensons en effet que ces éléments terminologiques, détectables automatiquement, sont suffisamment fiables pour être utilisées comme points d'ancrage dans les motifs d'association. L'expérience tend à vérifier ce point : il est possible d'améliorer la qualité des résultats dans le cas difficile d'un corpus peu volumineux et impliquant des langues très différentes.

La dernière approche part du constat que de nombreuses traductions communes sont obtenues avec l'alignement français-japonais et anglais-japonais, mais qu'elles ne sont pas classées de la même façon dans ces deux cas. En exploitant les indices donnés par ces deux alignements, nous sommes parvenus à améliorer de façon significative la qualité des résultats dans le cas du japonais.

Dans ce chapitre, nous avons donc montré différentes manières d'affiner la caractérisation des termes dans un cadre multilingue dans le but d'obtenir automatiquement des couples de traduction à partir de corpus comparables. Le chapitre suivant s'éloigne de cette problématique et se penche sur la notion même de corpus comparables et la façon dont ils sont exploités pour l'extraction lexicale.

CHAPITRE 6

Discussion : incomparabilité des corpus comparables

Ce chapitre présente quelques réflexions issues de l'ensemble des travaux réalisés pendant ce travail de doctorat. Cette discussion est motivée par un ensemble d'observations et par l'intuition acquise pendant ces trois années d'étude des corpus comparables. La vision développée ici s'applique aux corpus comparables *tels que nous les utilisons dans cette étude*, c'est-à-dire en particulier sur leur exploitation dans le cas de l'extraction lexicale bilingue. Dans ce cadre, la caractérisation des paires de traductions se fait sur les points de comparaisons que l'on peut extraire automatiquement des deux parties des corpus comparables.

6.1 Statistique des corpus comparables

L'extraction de lexique bilingue à partir de corpus comparables se fait, nous l'avons vu tout au long de cette étude, en extrayant le contexte des mots, en façonnant ce contexte de manière à évaluer l'importance de chacun de ses éléments, puis en comparant les *motifs d'association* ainsi construits. En d'autres termes, il s'agit de trouver des points de comparaison discriminants pour chacun des mots à traduire. Certains de ces points de comparaison donneront des informations sur la *similarité* entre deux mots (les informations que l'on retrouve de manière similaire entre deux motifs) et d'autres donneront des informations sur la dissimilarité entre deux mots (les informations absentes d'un motif à l'autre, ou très différentes dans leur pondération).

Le motif d'un mot est forgé en utilisant des mesures d'associations que nous avons présentées au chapitre 3. Nous revenons tout d'abord sur ces associations pour montrer pourquoi elles sont parfois exploitées de façon erronée et pourquoi elles peuvent être inadéquates dans le cas général.

6.1.1 Retour sur les mesures statistiques utilisées

Nous avons présenté dans le chapitre 3 différentes mesures d'association utilisées pour extraire des relations sémantiques entre les mots d'un corpus, notamment pour regrouper des éléments en relation de synonymie ou d'antonymie. Nous avons par ailleurs constaté, tout au long des expériences présentées dans ce document, qu'il était difficile de prévoir quelles mesures d'association permettaient d'obtenir les meilleurs résultats dans le cas de l'extraction lexicale bilingue.

Ce résultat est surprenant : dans le cas de la détection de collocations (section 3.2), le taux de vraisemblance, introduit tardivement en statistique (Dunning, 1993), semble obtenir les faveurs d'une partie

de la communauté scientifique, pour sa capacité à être une approximation du test exact de Fisher (Evert, 2008) et parce qu'il combine, dans une seule valeur numérique, l'*effet* d'une association (la force de la relation entre deux mots w_1 et w_2) et sa *significativité*, c'est-à-dire la confiance que l'on peut accorder à une telle relation. À l'inverse, l'information mutuelle ponctuelle, que nous utilisons fréquemment dans nos expériences sur les corpus comparables a montré qu'elle était capable de fournir les meilleurs résultats dans certaines configurations alors même qu'elle est intrinsèquement naïve : en reprenant les notations du chapitre 3, les deux tables de contingence suivantes (table 6.1) retourneront une même valeur d'information mutuelle ponctuelle, bien que l'une contienne des valeurs beaucoup plus importantes que l'autre.

	j	$\neg j$	
i	1	5	6
$\neg i$	4	30	34
	5	35	40

	j	$\neg j$	
i	50	250	300
$\neg i$	200	1 500	1 700
	250	1 750	2 000

Table 6.1 – Tables de contingence équivalentes pour le calcul de l'information mutuelle ponctuelle.

Dans les deux cas, le score d'information mutuelle ponctuelle sera de $1,333$. Le taux de vraisemblance retourne lui des résultats plus contrastés, puisque la table de gauche obtiendra un score de $0,075$ alors que la table de droite obtiendra un score de $1,88$, correspondant à une association très faible dans un cas, à une association significative dans l'autre.

6.1.2 Effectifs et fréquences

Fort de cette observation, il apparaît naturel de s'appuyer sur la significativité de l'association entre deux mots d'un corpus, ne serait-ce que pour ne pas sur-pondérer les couples de mots rares. Toutefois, cette significativité est généralement calculée sur la *fréquence* de la cooccurrence dans un corpus isolé (d'où la correction de l'information mutuelle ponctuelle en information mutuelle locale, présentée en section 3.2). En réalité, dans le cas d'un corpus isolé, la fréquence n'est pas calculée, seul l'effectif est pris en compte, ce qui est cohérent puisque la taille du corpus ne variant pas, l'effectif est directement proportionnel à la fréquence (rappelons que la fréquence est le rapport entre l'effectif d'une cooccurrence et le nombre de cooccurrences du corpus).

Dans le cas de corpus comparables, il s'agit de comparer ces associations non plus au sein d'un seul corpus, mais entre deux corpus ayant des caractéristiques potentiellement différentes. Il est maladroit d'utiliser les mesures d'association « telles quelles », puisque rien n'indique que les effectifs seront similaires d'un corpus à l'autre. Il semble que ce soit un mauvais usage de ces outils, conçus pour des comparaisons au sein de corpus isolés. Il est nécessaire dans le cas de corpus distincts d'ajouter l'information de la fréquence (c'est-à-dire de rapporter chaque valeur de la table de contingence par le nombre de cooccurrences du corpus), et donc de modifier les mesures d'association en conséquence.

De toute façon, rien n'indique que les fréquences des mots, et par extension, les fréquences des *cooccurrences* soient elles-même similaires dans les parties sources et cibles d'un corpus comparable. Nous avons repris les corpus utilisés par Morin (2009) qui sont des échantillons du corpus *Cancer du sein* complet[1]. Une des premières observations de ce travail est que les résultats de l'approche directe varient sensiblement en fonction de l'échantillon de corpus anglais utilisé. Ces résultats sont consignés dans la table 6.2.

[1] Jusque là, nous n'utilisions que l'un de ces échantillons, de manière à avoir un corpus équilibré entre la partie française et anglaise.

Échantillon	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Top_1	59	66	69	59	58	67	52	64	70	62	67	64	67	60
Top_5	136	134	134	147	134	163	137	151	160	149	130	147	138	143

Table 6.2 – Alignement français-anglais, corpus *Cancer du sein* (échantillonné). Mesures du taux de vraisemblance et Jaccard pondéré, liste d'évaluation [En-Fr-648]. Nombre de traductions correctes trouvées.

Pour le Top_1, ces résultats varient entre 52 et 70, avec une moyenne de 63,1 et un écart-type de 5,0 : les résultats sont significativement variables d'une expérience à l'autre alors même que le corpus d'origine est *homogène*, dans le sens où tous les documents le composant ont été sélectionnés de la même façon et que les échantillons ont été découpés aléatoirement. Les paramètres de chaque expérience étant les mêmes, la différence de résultat ne peut provenir que des différences entre les corpus. Notre deuxième observation porte sur l'effectif de certains termes disponibles dans chaque échantillon. Ils sont regroupés dans le tableau 6.3.

Terme	1	2	3	4	5	6	7
cancer	3 285	3 566	3 238	3 848	3 591	3 763	3 035
breast	3 304	3 370	3 187	3 905	3 396	3 528	2 972
tumor	930	707	657	595	699	764	604
tumour	574	396	476	919	404	802	865
abnormality	33	26	36	48	16	21	71
	8	9	10	11	12	13	14
cancer	3 895	4 181	3 836	3 360	3 659	3 726	3 786
breast	3 918	4 108	3 722	3 270	3 941	3 696	3 443
tumor	438	615	518	861	535	620	993
tumour	835	579	783	693	736	540	346
abnormality	21	31	41	23	23	48	20

Table 6.3 – Effectif de quelques termes par échantillon de la partie anglaise du corpus *Cancer du sein*.

Les informations brutes de cette table sont analysées dans le tableau 6.4 qui présente l'étendue, la moyenne et le taux de variation par rapport à la moyenne (équation 6.1) pour les effectifs de chacun des termes du tableau 6.3. La mesure de variation que nous proposons calcule l'écart maximal par rapport à la moyenne, rapporté à la moyenne[2]. Les indices de dispersion classiques (écart-type, variance) ont le défaut de ne pas être comparables dans ce cas.

$$Variation(e_i) = \frac{Max((Moyenne(e_i) - e_{min}), (e_{max} - Moyenne(e_i)))}{Moyenne(e_i)} \tag{6.1}$$

Les valeurs du tableau 6.4 indiquent que les effectifs de chacun des termes varient de façon significative d'un échantillon à l'autre. Pour les mots très fréquents (*cancer* et *breast*[3]), la variation est de l'ordre de 16 %. Pour les mots plus rares (*abnormality*), la variation est supérieure à la moyenne : dans ce cas, alors que la moyenne des effectifs est de 33, dans certains corpus, cet effectif n'est que de 16

[2]Nous aurions pu calculer cet indice de plusieurs autres façons, par exemple avec le rapport de la moitié de l'étendue sur la médiane ou le rapport de l'étendue et de la moyenne : les résultats sont proches quelque soit l'indice utilisé. Celui que nous avons choisi indique dans quelles proportions les valeurs e_i de la liste s'éloignent de la moyenne.

[3]Notons que les effectifs de ces deux termes semblent correlés d'un échantillon à l'autre, ce qui n'est pas surprenant puisque les documents traitent tous du cancer du sein, *breast cancer* en anglais.

Terme	Étendue	Moyenne	Variation
cancer	1 146	3 626, 4	16, 3 %
breast	1 136	3 554, 3	16, 4 %
tumor	555	681, 1	45, 8 %
tumour	573	639, 1	45, 9 %
abnormality	55	32, 7	117, 0 %

Table 6.4 – Indice de position et de dispersion pour les effectifs de différents termes.

(soit environ 50 % de la moyenne), la valeur la plus extrême étant de 71, c'est-à-dire plus de 200 % de la moyenne. Ces observations amènent à penser que les corpus comparables ne sont pas des ressources *homogènes* : un échantillon (suffisamment conséquent) tiré au sort dans un plus grand corpus comparable n'aura pas les mêmes propriétés que le corpus d'origine ou que d'autres échantillons. C'est un point clé dans l'exploitation des corpus comparables puisque nous faisons l'hypothèse que la distribution des cooccurrences est comparable entre les différentes parties d'un corpus comparable.

6.1.3 Disparité des cooccurrences

Si les fréquences des termes varient, les fréquences des cooccurrences devraient varier également. Le terme *cancer* qui n'apparait que 3 035 fois dans l'échantillon 7 ne pourra cooccurer autant de fois avec *breast* que dans l'échantillon 9 où il apparaît 4 181 fois (là encore, le nombre de cooccurrences étant très proche d'un corpus à l'autre, la fréquence d'une cooccurrence est directement proportionnelle à son effectif). Le tableau 6.5 compare les cooccurrences les plus fréquentes pour les cinq termes étudiés précédemment, pour chaque échantillon de corpus, pour une taille de fenêtre contextuelle de 2. Pour chacun des cinq termes étudiés précédemment, nous précisons les deux cooccurrences les plus fréquentes et leur effectif. La valeur entre parenthèses indique le nombre de cooccurrences du terme par rapport au terme principal (par exemple, pour le premier échantillon, dans la colonne *cancer*, *breast* (70,2 %) indique que *breast* apparait avec 70,2 % des occurrences de *cancer*).

Cette table montre d'une part que, pour les mots fréquents (*cancer* et *breast*), les cooccurrences les plus fréquentes sont relativement stables lorsqu'ordonnées : la première cooccurrence de *cancer* est toujours *breast*, la première cooccurrence de *breast* est toujours *cancer*. Toutefois, même dans ce cas, la deuxième cooccurrence est la plupart du temps *cell*, mais parfois *patient* (échantillon 8, 9 et 13) ou *risk* (échantillon 10). Cette stabilité se dégrade pour les termes moins fréquents (*tumor* et *tumour*) et ne se retrouve que marginalement pour le terme *abnormality*. Au-delà de l'ordre des cooccurrences principales, c'est surtout la variation du nombre de ces cooccurrences qui nous intéresse : ces valeurs varient significativement en fonction des échantillons. Ainsi, le nombre de cooccurrences de *cancer* et *breast* évolue entre 2 172 et 3 106. Ces valeurs sont encore plus dispersées pour les termes les moins fréquents. Par exemple, le nombre de cooccurrences entre *tumor* et *cell* évolue entre 71 et 276.

6.1.4 Inadéquation des mesures d'association

Tous les résultats présentés jusqu'ici dans ce chapitre étaient relativement attendus : il est en somme assez normal que l'on ne retrouve pas un nombre exact d'occurrences et de cooccurrences entre deux corpus différents même s'ils traitent des mêmes sujets. Toutefois, c'est une propriété qui est tout de même exploitée pour l'extraction de lexiques bilingues à partir de corpus comparables puisque le nombre de cooccurrences est utilisé pour pondérer les mesures d'association, étant révélateur de leur *significativité*.

Éch.	*cancer*	*breast*	*tumor*	*tumour*	*abnormality*
1	breast 2307 (70,2 %) cell 576 (17,5 %)	cancer 2307 (69,8 %) patient 471 (14,2 %)	cell 276 (18,2 %) breast 113 (8,5 %)	cell 105 (29,6 %) size 49 (12,1 %)	breast 5 (15,1 %) change 3 (9,0 %)
2	breast 2606 (73,0 %) cell 563 (15,7 %)	cancer 2606 (77,3 %) cell 436 (12,9 %)	cell 131 (15,4 %) size 73 (14,1 %)	size 61 (18,5 %) tumour 56 (10,3 %)	study 4 (15,3 %) breast 4 (15,3 %)
3	breast 2309 (71,3 %) cell 488 (15,0 %)	cancer 2309 (72,4 %) cell 460 (14,4 %)	cell 154 (16,5 %) growth 65 (13,4 %)	cell 79 (23,4 %) primary 64 (9,8 %)	endometrial 12 (33,3 %) sign 4 (11,1 %)
4	breast 2897 (75,2 %) cell 452 (11,7 %)	cancer 2897 (74,1 %) cell 396 (10,1 %)	size 88 (12,5 %) cell 66 (8,2 %)	size 115 (14,7 %) grade 76 (11,0 %)	number 10 (20,8 %) recurrent 8 (16,6 %)
5	breast 2616 (72,8 %) cell 598 (16,6 %)	cancer 2616 (77,0 %) cell 508 (14,9 %)	cell 164 (15,5 %) size 115 (12,8 %)	size 63 (23,4 %) patient 52 (16,4 %)	endometrial 7 (43,7 %) patient 2 (12,5 %)
6	breast 2675 (71,0 %) cell 638 (16,9 %)	cancer 2675 (75,8 %) cell 533 (15,1 %)	cell 233 (12,7 %) size 82 (9,6 %)	cell 102 (30,4 %) breast 77 (10,7 %)	chromosome 4 (19,0 %) genetic 4 (19,0 %)
7	breast 2172 (71,5 %) cell 478 (15,7 %)	cancer 2172 (73,0 %) cell 391 (13,1 %)	cell 134 (10,5 %) size 50 (9,8 %)	cell 91 (22,1 %) breast 85 (8,2 %)	test 22 (30,9 %) cardiac 16 (22,5 %)
8	breast 3105 (79,7 %) patient 440 (11,2 %)	cancer 3105 (79,2 %) patient 489 (12,4 %)	cell 99 (11,7 %) breast 38 (10,5 %)	size 98 (22,6 %) patient 88 (8,6 %)	endometrial 3 (14,2 %) mammographic 3 (14,2 %)
9	breast 3106 (74,2 %) patient 507 (12,1 %)	cancer 3106 (75,6 %) patient 508 (12,3 %)	cell 154 (14,8 %) breast 44 (10,5 %)	cell 86 (25,0 %) breast 61 (7,1 %)	cgh 7 (22,5 %) cell 5 (16,1 %)
10	breast 2917 (76,0 %) risk 416 (10,8 %)	cancer 2917 (78,3 %) risk 426 (11,4 %)	cell 71 (14,0 %) size 59 (12,0 %)	size 110 (13,7 %) patient 94 (11,3 %)	endometrial 11 (26,8 %) rate 5 (12,1 %)
11	breast 2417 (71,9 %) cell 615 (18,3 %)	cancer 2417 (73,9 %) cell 493 (15,0 %)	cell 163 (10,9 %) tumor 64 (8,6 %)	size 76 (18,9 %) tumour 60 (7,4 %)	include 3 (13,0 %) endometrial 3 (13,0 %)
12	breast 3003 (82,0 %) cell 551 (15,0 %)	cancer 3003 (76,1 %) cell 536 (13,6 %)	cell 121 (12,7 %) growth 52 (12,5 %)	patient 94 (22,6 %) size 92 (9,7 %)	endometrial 3 (13,0 %) patient 3 (13,0 %)
13	breast 2760 (74,0 %) patient 479 (12,8 %)	cancer 2760 (74,6 %) patient 506 (13,6 %)	cell 98 (10,9 %) size 72 (10,0 %)	breast 59 (15,8 %) size 54 (11,6 %)	breast 7 (14,5 %) complex 6 (12,5 %)
14	breast 2696 (71,2 %) cell 590 (15,5 %)	cancer 2696 (78,3 %) cell 457 (13,2 %)	cell 229 (17,0 %) size 152 (15,0 %)	patient 59 (23,0 %) cell 52 (15,3 %)	cgh 4 (20,0 %) cell 3 (15,0 %)

Table 6.5 – Observation des cooccurrences les plus fréquentes pour chaque échantillon du corpus *Cancer du sein*, pour une fenêtre contextuelle de taille 2.

Evert (2008) écrit (l'emphase est nôtre) :

> « *In practical applications, MI was found to have a tendency to assign inflated scores to low-frequency word pairs with* $E \ll 1$ *[...]. In order to counterbalance this low-frequency bias of MI, various heuristic modifications have been suggested. The most popular one* ***multiplies the denominator with*** O ***in order to increase the influence of observed cooccurrence frequency compared to the expected frequency***. »

La valeur O correspond exactement au nombre de cooccurrences observées, la valeur E correspond au nombre de cooccurrences espérées sous l'hypothèse nulle (cf. chapitre 3). Or, même si le rapport O/E entre deux valeurs devait être identique (revoir l'exemple du tableau 6.1), il sera pondéré par O, qui lui même n'est pas comparable d'un corpus à l'autre.

C'est précisément le problème qui se pose pour l'exploitation de corpus comparables déséquilibrés. Morin (2009) constate une forte variation des résultats de l'alignement au fil du déséquilibre du corpus. Il utilise les 14 échantillons de la partie anglaise du corpus *Cancer du sein* qu'il concatène pour obtenir un déséquilibre de 1 : 1 (partie française et anglaise de même taille) à 1 : 14 (partie anglaise 14 fois plus volumineuse que la partie française). Il constate qu'en utilisant le taux de vraisemblance, les résultats se dégradent au fil du déséquilibre. À l'inverse, lorsqu'il utilise l'information mutuelle ponctuelle, la qualité des résultats s'améliore. Il conclut donc qu'il est possible d'exploiter des corpus comparables déséquilibrés dans certaines conditions. Nous allons plus loin : l'augmentation *homogène* de la taille d'un corpus augmente le nombre d'occurrences et de cooccurrences, et donc la fiabilité des observations consignées dans les vecteurs de contexte. Il est alors pertinent d'utiliser des ressources déséquilibrées, c'est-à-dire de prendre autant de ressources que disponibles, à condition de les traiter en conséquence. Si les résultats se dégradent en utilisant le taux de vraisemblance, c'est qu'il utilise l'information du nombre de cooccurrences pour pondérer la significativité du score : en augmentant le déséquilibre des corpus, on augmente d'un côté seulement le nombre de cooccurrences, ce qui entraîne également un déséquilibre des associations, donc des points de comparaisons, donc une moins bonne qualité d'alignement. À l'inverse, puisque l'information mutuelle n'évalue pas cette significativité, l'augmentation du nombre de cooccurrences augmente uniquement la *précision* du score d'association calculé : les vecteurs de contexte construits sur un plus grand nombre d'observations sont peuplés de façon plus précise, permettant de meilleures comparaisons et donc une meilleure qualité d'alignement. La faiblesse de l'information mutuelle ponctuelle devient ici une qualité.

En accord avec les remarques des deux précédents paragraphe, nous pouvons illustrer ces situations par un exemple concret. Dans la partie française du corpus *cancer du sein*, pour une fenêtre contextuelle de taille 2, *cancer* et *sein* cooccurrent 1 788 fois : c'est la valeur observée O ($O_1 = 1\,788$). Dans la partie anglaise (contenant tous les échantillons du corpus), *cancer* et *breast* cooccurrent 35 547 fois ($O_2 = 35\,547$). Le terme *cancer* apparaît 3 235 fois dans le corpus français, soit une fréquence d'apparition dans les fenêtres contextuelles de $3\,235/2\,120\,000 \approx 0,15\,\%$ (2 120 000 correspond au nombre de contextes évalués sur le corpus et non au nombre de mots du corpus). La fréquence de *cancer* dans les fenêtre contextuelle du corpus anglais est de $50\,769/29\,680\,000 \approx 0,17\,\%$. Ainsi, *sein*, qui apparaît 2 927 fois a $2\,927 \times 0,15 \approx 439$ chances d'apparaître avec *cancer* sous l'hypothèse que les mots sont distribués aléatoirement, c'est la valeur espérée E_1. D'autre part, *breast* a $49\,760 \times 0,17 \approx 8\,459$ d'apparaître dans le contexte de *cancer*, toujours sous l'hypothèse nulle, c'est la valeur E_2.

Dans le corpus français, le rapport sera de $O_1/E_1 = 1\,788/439 \approx 4,07$. Pour le corpus anglais, il sera de $O_2/E_2 = 35\,547/8\,459 \approx 4,20$. L'information mutuelle ponctuelle calculera une valeur de $log(O_1/E_1) = log(4,07) \approx 1,40$ pour le premier cas et de $log(O_2/E_2) = log(4,20) \approx 1,43$ pour le deuxième cas. Le taux de vraisemblance sera, sur le corpus français de $2 \cdot (O_1 log(O_1/E_1) - (O_1 - E_1)) \approx$

2 323. Sur le corpus anglais : $2 \cdot (O_2 log(O_2/E_2) - (O_2 - E_2)) \approx 47\,888$. Nous voyons bien dans cette exemple que l'information mutuelle ponctuelle calcule des valeurs d'association très proches entre les cooccurrences française et anglaise alors que taux de vraisemblance est largement déséquilibré.

Les mesures *intelligentes* de l'association ne donnent de bons résultats que lorsque les nombres d'occurrences et de cooccurrences sont comparables, ce qui est sensiblement le cas dans un corpus équilibré, mais qui s'avère faux dans un corpus déséquilibré. Ainsi, nous avons obtenu précédemment des résultats concluants en utilisant le taux de vraisemblance, sur des corpus très contraints (en particulier en terme d'équilibre), mais dans un cadre général, cette mesure s'avère inadéquate, par rapport à des mesures plus naïves, ne mesurant que l'*effet* des associations. Toutefois, l'information mutuelle ponctuelle a d'autres faiblesses puisqu'elle donne des résultats beaucoup moins bons que ceux obtenus avec le taux de vraisemblance dans le cas de corpus équilibrés : elle place au même niveau des cooccurrences rares et fréquentes, c'est-à-dire notamment qu'elle est susceptible de construire des motifs d'associations proches pour des éléments incomparables, comme nous l'avons montré en section 6.1.1. Une proposition pour contourner ce problème est non pas de s'appuyer sur la fréquence des cooccurrences, mais sur la fréquence des mots comparés, par exemple en ne comparant deux à deux que les motifs des mots ayant des fréquences proches, en fixant un seuil de variation. Ainsi, le motif de *cancer* ne sera pas comparé au motif de *abnormality* puisqu'ils appartiennent à des classes de fréquences très différentes. Dans ce cas, nous conservons l'indice de la fréquence sans lui donner une importance quantitative dans la comparaison des contextes puisque ce sont des valeurs trop instables pour être utilisées ainsi. La fréquence n'est plus exploitée que qualitativement pour éviter de comparer des motifs incomparables.

Nous avons appliqué cette proposition sur le corpus *cancer du sein* (dans sa configuration équilibrée), en ne comparant que les éléments dont la fréquence ne diffère pas de plus de 15 %. Les résultats sont consignés dans le tableau 6.6 :

- l'expérience *Témoin* est celle qui donne les meilleurs résultats dans le cadre général (étalonnée au chapitre 5 –taux de vraisemblance, Jaccard pondéré) ;
- l'expérience IM/COS correspond aux résultats obtenus en utilisant les mesures d'information mutuelle et de cosinus non modifiées ;
- la dernière colonne, *Filtre fréquence*, correspond aux meilleurs résultats obtenus en filtrant les candidats par leur fréquence, dans ce cas avec la combinaison de l'information mutuelle et du cosinus.

Ainsi, nous pouvons comparer ces résultat aux meilleurs résultats obtenus précédemment, mais aussi aux résultats obtenus avec les mêmes mesures, sans filtre de fréquences. Toutes les expériences sont réalisées avec la liste [En-Fr-122] et une fenêtre de taille 3, le gain est calculé par rapport à l'expérience témoin.

	Témoin	*IM/COS*	*Filtre fréquence* [Gain]
Top_1	20,5 % (25)	18,0 % (22)	23,0 % (28) [12,0 %]
Top_5	38,5 % (47)	35,2 % (43)	40,1 % (49) [4,2 %]

Table 6.6 – Effet du filtrage par fréquence des candidats à la traduction. Liste [En-Fr-122], taille de fenêtre 3.

Les résultats présentés dans le tableau sont intéressants non seulement parce qu'ils indiquent un gain pour le Top_1, mais surtout parce que la mesure du taux de vraisemblance, qui donnait jusque là les meilleurs résultats pour cette configuration, est dépassée par l'information mutuelle. Ces résultats indiquent non seulement qu'il est possible de s'appuyer sur la fréquence des mots à traduire pour élaguer l'espace de recherche (offrant un gain en qualité, mais aussi en temps de calcul) mais également que l'hypothèse développée dans ce chapitre se confirme sur cet exemple. Toutefois, cette expérience n'est

qu'une illustration : nos recherches n'ont pas portées en profondeur sur la notion de filtrage par fréquence et il est probable que la méthode soit largement améliorable en utilisant des filtres plus subtils, en accord notamment avec les données présentées dans le tableau 6.5.

6.2 Un autre regard sur les corpus comparables

Les observations précédentes et la discussion sur les mesures d'association nous invitent à revoir la façon dont les corpus comparables sont exploités, ainsi qu'à compléter la notion relativement floue de comparabilité.

6.2.1 Des ressources hétérogènes

Les corpus comparables, comme tout corpus correctement constitué[4], sont des ressources hétérogènes[5]. C'est un point que nous avons soulevé rapidement dans le premier chapitre, concernant le problème des fréquences nulles. La *loi de Zipf*[6] dit grossièrement que la fréquence d'un mot est inversement proportionnelle à son rang lorsque classé par fréquence. Corollairement, cette loi indique qu'il y aura très peu de mots avec une fréquence très forte et beaucoup de mots avec des fréquences très faibles : un sous-échantillon d'un corpus verra disparaître certains mots de fréquence faible, d'où le problème des fréquences nulles[7]. Une telle loi se vérifie sur la partie anglaise du corpus *cancer du sein*, la figure 6.1 présente la fréquence de chaque mot du corpus (en ordonnée) en fonction de son rang (échelle logarithmique, en abscisse et en ordonnée).

Augmenter la taille d'un corpus ne modifiera pas cette distribution : l'augmentation apportera vraisemblablement de nouveaux mots peu fréquents et peu de mots très fréquents, déjà présents dans le corpus d'origine. Autrement dit : la loi de Zipf se vérifie sur des sous-parties du corpus pour peu que leur taille soit suffisamment importante. Toutefois, il est délicat d'extrapoler les fréquences des mots d'un corpus à partir de l'observation d'un échantillon du corpus : ce n'est pas parce qu'un mot apparait x fois dans un quart du corpus qu'il apparaîtra $4 \times x$ fois dans l'ensemble du corpus. La loi des grands nombres – qui, sommairement, statue que la moyenne d'une variable aléatoire tend à se stabiliser au-delà d'un grand nombre de tirages – ne s'applique pas sur les corpus (même volumineux), car les mots n'y apparaissent pas indépendamment (Kilgarriff, 2005). C'est en ce sens que nous parlons de *ressources hétérogènes*. Nous avons montré à plusieurs reprises dans ce chapitre des exemples d'hétérogénéité, en particulier en observant des exemples d'effectifs et de cooccurrences sur des échantillons issus d'un même corpus.

Cette hétérogénéité nous pousse à revoir la notion de comparabilité ou plutôt, à revoir la façon dont les informations que nous voulons exploiter sont distribuées dans le corpus. Nous avons montré précédemment qu'un déséquilibre d'information n'était pas nécessairement un problème et qu'il était possible d'en tirer parti, puisque ce déséquilibre impliquait également une meilleure représentation de l'information.

[4] Revoir le cas échéant le début du chapitre 1, concernant la *représentativité des corpus*.

[5] La notion d'hétérogénéité ici n'est pas en opposition avec la notion d'*homogénéité* présentée précédemment, qui ne s'appliquait qu'à la sélection des documents constituant le corpus, *homogènes* car tous sélectionnés avec des contraintes précises et communes.

[6] Cette loi a été généralisée par Benoît Mandelbrot en 1965, au cadre général des données ordonnées (Mandelbrot, 1965).

[7] Le problème des fréquences nulles s'exprime généralement dans le sens inverse, il est impossible d'avoir un corpus suffisamment représentatif pour faire apparaître tous les points d'observations. Dans ce cas, le sous-corpus correspond aux corpus sur lesquels sont réalisées les observations, le corpus correspond à tous les cas corrects pouvant apparaître en réalité.

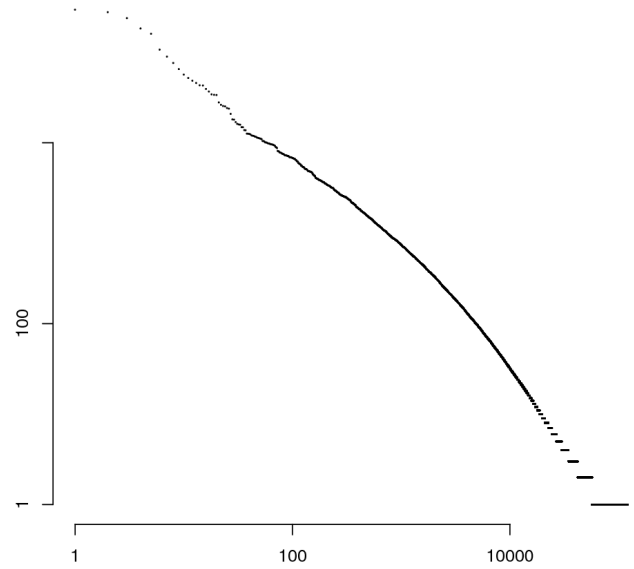


Figure 6.1 – Distribution de Zipf sur le corpus *cancer du sein*.

6.2.2 Comparabilité des corpus comparables

D'après la définition de Bowker et Pearson (2002), les corpus comparables ne sont « que » des ensembles de textes, dans des langues différentes, n'étant pas en relation de traduction. La *comparabilité* d'un corpus se mesure aux traits communs entre les différentes parties de ce corpus. Nous distinguons deux types de critères pour évaluer cette comparabilité (Déjean et Gaussier, 2002) :

- les critères qualitatifs : genre, auteur, type de discours ;
- les critères quantitatifs : la quantité et la similarité de certains traits linguistiques (typiquement, le vocabulaire commun).

Cette notion reste toutefois floue et le choix de critères de comparabilité va dépendre principalement de l'usage du corpus. Ils vont donc être définis lors de sa construction. Dans notre cas, nous voulons aligner la terminologie de certains domaines de spécialité. Il est donc apparu naturel de s'intéresser à des textes médicaux et d'imposer certaines restrictions (notamment le type de discours) pour garantir une certaine représentativité de cette terminologie. Nous avons utilisé des critères qualitatifs pour obtenir un corpus, comparable quantitativement, c'est-à-dire contenant un certain usage d'une terminologie. Pour d'autres usages, par exemple, l'analyse contrastive multilingue (voir chapitre 1) les corpus sont construits sur des bases différentes. Ainsi Lewis (2005) construit un corpus de discours politiques français, anglais britannique et irlandais pour étudier les connecteurs adversatifs.

Dans le cadre de l'extraction lexicale bilingue et de la façon dont nous exploitons les corpus comparables, ils doivent avoir les propriétés suivantes :

- ils doivent contenir un vocabulaire commun, vocabulaire que nous cherchons à extraire et à aligner ;
- ils doivent contenir des usages identiques du vocabulaire commun, usages que l'on pourra compa-

rer d'un corpus à l'autre ;

Ce sont a priori les deux seules propriétés dont nous avons besoin, qui pourront être la conséquence des critères qualitatifs énoncés précédemment. Nous proposons alors de voir les corpus comparables comme des ensembles de textes contenant des *pépites* de comparabilités, c'est-à-dire des exemples d'emplois identiques de termes, entre le corpus cible et le corpus source. *Identique* doit ici se comprendre dans le sens d'*un même usage*, d'une même signification, et non pas dans le sens de passage de documents en relation de traduction. La comparabilité s'évalue alors sur le volume de données comparables entre le corpus source et le corpus cible. Un corpus parallèle reste donc *fortement comparable*, un corpus déséquilibré peut être très comparable, à condition que la terminologie employée dans les corpus source et cible soit équivalente (un même sens pour chacun des termes en relation de traduction). La comparabilité diminue lorsqu'un terme source a une traduction cible ayant plusieurs sens.

Concrètement, nous proposons de définir la comparabilité des corpus comparables en terme de *densité de comparabilité*, c'est-à-dire en terme de *volume d'information comparable* et de *volume d'information incomparable*. La figure 6.2 schématise cette nouvelle définition.

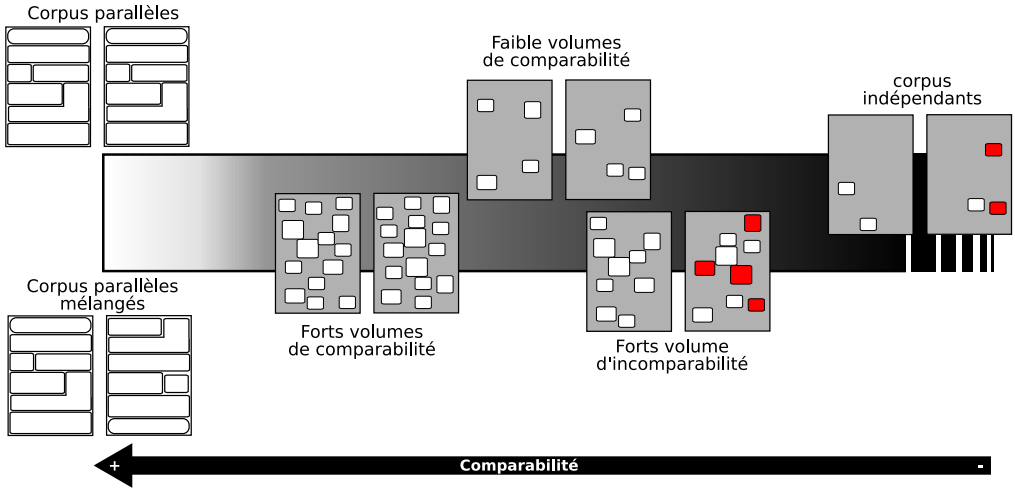


Figure 6.2 – Comparabilité des corpus multilingues.

La figure 6.2 représente plusieurs types de corpus. Dans cette figure, les zones blanches aux coins arrondis représentent les *pépites* comparables. Les zones grises indiquent les parties neutres des corpus, ne contenant pas d'informations comparables alors que les zones rouges aux coins arrondis représentent les parties *incomparables*, c'est-à-dire contenant des informations susceptibles de dégrader les résultats de l'alignement si exploitées. La *densité* de comparabilité dépend donc du volume d'informations comparables, pénalisée par le volume d'informations neutres et par le volume d'informations incomparables. Cette définition a l'avantage de faciliter la construction des corpus : il suffit de relever des documents traitant d'un sujet précis (ce qui est exactement la façon dont nous avons procédé pour construire les corpus utilisés dans nos travaux), sans avoir à se soucier d'équilibre entre les parties ou de bruit introduit par des documents non pertinents. Il reste important d'avoir des parties suffisamment conséquentes pour avoir une bonne représentation des mots à traduire (c'est-à-dire, un volume d'informations comparables suffisant), puisque nous avons montré que les mots de faibles fréquences tendent à avoir des vecteurs de contexte faiblement peuplés, et donc une représentation instable entre le corpus source et le corpus cible.

A priori, rien n'empêche d'ajouter des documents non-reliés aux corpus, c'est-à-dire des documents ne contenant pas d'occurrence des mots à traduire. Enfin, il doit être envisageable d'ajouter des docu-

ments *incomparables*, c'est-à-dire des documents contenant des occurrences des mots à traduire, mais avec des sens différents. L'enjeu consiste alors à être capable de déterminer ces différents sens, et surtout à repérer les *pépites* de comparabilité, pour savoir quels points d'observation sont pertinents et lesquels ne le sont pas. Précisons que cette comparabilité doit s'évaluer selon l'usage des corpus : le corpus *cancer du sein* est comparable lorsqu'il s'agit d'extraire la terminologie propre à la thématique du cancer du sein mais n'est pas adéquat dans d'autres cas. Les *zones de comparabilité* sont donc liées à ce que l'on veut comparer d'un corpus à l'autre.

Cette thèse n'est pas en contradiction avec celle de Morin *et al.* (2007), qui proposent qu'un corpus comparable correctement constitué est au moins aussi efficace qu'un corpus comparable moins bien constitué mais plus volumineux. D'une part, cela reste sans doute vrai : un corpus comparable correctement constitué contiendra une plus grande densité de parties comparables qu'un corpus volumineux mais moins contraint. D'autre part, leur proposition s'appuie sur l'approche par traduction directe pour leur démonstration, c'est-à-dire une approche s'appuyant sur la comparaison des fréquences de cooccurrences des termes. Or, nous avons démontré que cette fréquence était instable, même dans le cas de corpus fortement comparables, mais que ce phénomène est aggravé dans le cas de corpus moins comparables. Autrement dit, l'hypothèse de Morin *et al.* (2007) est vraie en utilisant les approches classiques d'extraction lexicale bilingue.

L'enjeu est, à notre sens, de trouver de nouvelles méthodes (plutôt des améliorations de méthodes actuelles) pour exploiter des corpus contenant une plus faible densité de parties comparables. Cette densité ne devrait pas être un obstacle tant que le volume d'information comparable est stable. Dans un premier temps, il faudrait être en mesure de trouver des résultats équivalents en utilisant un corpus comparable témoin et le même corpus auquel auront été ajoutés des documents non reliés (n'introduisant pas d'homonymes des mots déjà en relation de traduction) voire même en ajoutant des documents *incomparables*, ajoutant des sens nouveaux à des graphies déjà connues. Une première étape a déjà été franchie, en exploitant des corpus comparables déséquilibrés (Morin, 2009), c'est-à-dire contenant des effectifs de cooccurrences incomparables. La solution a été de s'affranchir des mesures d'associations s'appuyant sur ces effectifs pour la pondération et des résultats prometteurs ont été obtenus en utilisant l'information mutuelle ponctuelle.

6.2.3 Exploitation de corpus à faible densité de comparabilité

L'exploitation de corpus à faible densité de comparabilité nécessite de résoudre plusieurs problèmes. L'un des premiers consiste à tirer profit des zones comparables, ce qui est fait naturellement dans l'approche directe utilisée tout au long de cette étude, à condition d'utiliser les paramètres adéquats. La figure 6.3 schématise un corpus déséquilibré en reprenant les représentations utilisées dans la figure 6.2.

Cette figure montre que les deux parties d'un corpus déséquilibré ont, certes, des volumes d'informations comparables différents, mais que leurs densités peut rester comparables. Il est donc *a priori* très intéressant d'utiliser des corpus comparables déséquilibrés : ils permettent d'avoir plus d'informations (c'est-à-dire, entre autres, plus de points de comparaison à exploiter, une meilleure représentation de certains mots à traduire…) sans nécessairement les dénaturer.

Les zones *neutres* (en gris dans les figures 6.2 et 6.3) n'entrent a priori pas en jeu dans l'alignement. Si aucun des mots sources que l'on cherche à traduire se situe dans ces zones, leur vecteur de contexte ne seront pas exploités. Les candidats cibles construits sur ces zones n'auront en principe rien en commun avec les mots sources à traduire ; ils seront facilement écartés lors de l'étape de comparaison. En d'autres termes, les résultats obtenus à partir d'un corpus comparable donné et les résultats obtenus à partir de ce même corpus augmenté de documents très différents devraient être identiques.

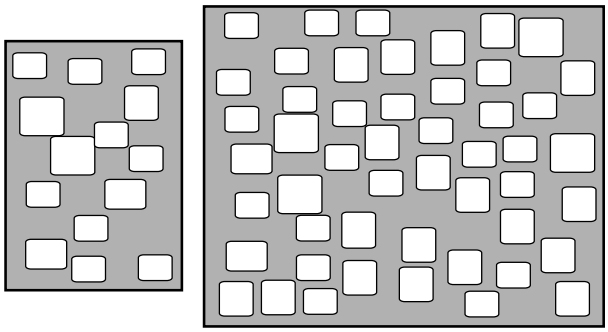


Figure 6.3 – Représentation d'un corpus déséquilibré.

Les zones *incomparables* posent plus de problèmes (en rouge sur la figure 6.2). Nous en rencontrons dans le corpus *cancer du sein*. Dans ce corpus, *cancer* a vraisemblablement toujours le même sens et n'est probablement jamais employé dans un sens figuré, puisqu'il s'agit de registre scientifique. D'autres termes n'ont pas cette propriété, par exemple *history*, correspond généralement au terme français *antécédent*, par exemple dans l'extrait « *[..] each patient underwent a baseline assessment, including a complete* ***medical history*** *and physical examination, complete blood counts (CBC), and blood chemistry tests.* ». Ici, *medical history* se traduit par *antécédent médicaux*. Toutefois, *history* est aussi employé dans l'état de l'art de certains documents, pour référer à l'évolution des techniques, comme dans l'extrait « *Radiation-induced lung cancer after radiotherapy for breast cancer has already been found to have* ***a long natural history****, with much less hazard during the first decade than later.* ». Dans ce cas, l'expression *a long natural history* est idiomatique et se traduit par *une longue histoire naturelle*. Cet exemple démontre que, même pour ce corpus soigneusement construit, il est possible de trouver des traces d'incomparabilités : *antécédent* devra être aligné avec *history* parce que certaines occurrences d'*history* dans le corpus cible sont les traductions d'*antécédent*, mais il faut écarter les autres sens qui ne feront vraisemblablement qu'ajouter des points de comparaisons non-pertinents. Il est donc nécessaire de séparer les différents sens que peut avoir un mot donné avant de construire son vecteur de contexte, ce qui est précisément une des tâches de l'acquisition sémantique, que nous avons présentée au chapitre 3. Cette étape devrait permettre d'isoler les différentes *pépites* de comparabilité entre les deux corpus en isolant les usages inadéquats de certains termes.

Toutefois, il faut s'attendre à une dégradation des résultats dans ce cas, par rapport à l'exploitation de corpus plus comparables. D'une part, l'acquisition sémantique nécessite des jeux de données conséquents pour être efficace. Il faut un nombre suffisant de représentations de chaque usage d'un mot pour que le partitionnement des différents sens soit efficace. Par ailleurs, cette étape risque d'introduire à nouveau du bruit en raison de partitionnements peu efficaces. Enfin, s'il faut écarter tous les usages inadéquats d'un mot, il doit rester suffisamment d'exemples d'usages adéquats pour que les vecteurs de contexte construits puissent être significatifs et puissent constituer des points de comparaisons pertinents.

Pour résumer, nous avons donc différentes méthodes pour exploiter les corpus comparables en fonction de leur comparabilité :

– L'approche directe classique, telle qu'utilisée dans cette étude, pour les corpus dont les volumes de comparabilité sont proches.
– L'approche directe classique, en utilisant les mesures d'associations adéquates, pour des corpus dont les volumes d'incomparabilité sont faibles et les volumes de comparabilité différents (cas des corpus déséquilibrés).

– L'approche directe précédée d'un partitionnement lexical sémantique pour des corpus à fort volume d'incomparabilité.

Reste toutefois un problème, celui de l'évaluation de la comparabilité, pour être capable de choisir *a priori* la méthode la mieux adaptée.

6.2.4 Évaluation de la comparabilité

Le degré de comparabilité entre deux corpus, en respectant notre définition, va se définir par rapport, d'une part, aux volumes d'informations comparables entre les parties source et cible, d'autre part, aux volumes d'informations incomparables entre ces deux parties. Ainsi, un corpus parallèle aura une forte comparabilité (des volumes très proches d'informations comparables entre les deux) alors que des corpus non-reliés seront très peu comparables, puisque contenant peu d'informations comparables, et un certain nombre d'informations incomparables.

L'évaluation des quantités d'informations comparables peut se faire en appliquant l'approche directe sur des mots dont les traductions sont connues en mesurant la qualité de l'alignement. Cela sous-entend de ne s'appuyer que sur des mots bien caractérisés, c'est-à-dire dont les contextes sont bien connus, couverts par les ressources linguistiques et dont les traductions sont elles mêmes bien caractérisées. La comparaison des vecteurs de contexte de ces paires de mots doit donner un bon indice quant à la comparabilité des corpus : s'ils sont proches, alors les corpus sont comparables (pour ce vocabulaire).

L'évaluation des quantités d'informations incomparables peut se faire en plusieurs étapes. Elle est tout d'abord donnée par l'évaluation des quantités d'informations comparables : s'il y en a peu, c'est-à-dire, si nous avons trouvé un vocabulaire commun entre les deux parties du corpus, mais que leurs contextes sont très différents – ce qui implique des usages différents des mots – alors les corpus sont incomparables. Elle peut aussi être évaluée au sein de chaque corpus en observant, pour un vocabulaire donné, les différentes partitions retournées par l'acquisition sémantique. Les usages franchement différents d'un mot pourtant jugé pertinent par rapport aux thématiques du corpus sont de bons indices de leur incomparabilité. Par exemple, un corpus où l'usage de *drug* avec le sens de *médicament* est en concurrence avec l'usage au sens de *drogue* sera révélateur de zone d'incomparabilité. Dans le cas du corpus *cancer du sein*, l'usage de *history* au sens d'*antécédent* est largement plus répandu que l'usage au sens de *passé* : l'incomparabilité est faible.

L'évaluation de la comparabilité peut donc se faire en étudiant un vocabulaire suffisamment bien représenté sur les critères suivants :

1. la quantité de vocabulaire commun entre les deux parties du corpus ;
2. la quantité de vocabulaire commun dont les contextes sont proches d'un corpus à l'autre (lorsque les contextes peuvent être comparés) ;
3. la quantité de vocabulaire commun dont les contextes sont différents d'un corpus à l'autre (lorsque les contextes peuvent être comparés) ;
4. la quantité de vocabulaire commun utilisé avec des sens différents au sein d'un même corpus ;

Le critère de comparabilité doit augmenter avec l'augmentation des critères 1 et 2, et diminuer avec l'augmentation des critères 3 et 4. En respectant ces critères, l'évaluation de la comparabilité d'un corpus comparable déséquilibré se fera sur le vocabulaire commun entre les deux parties et ses différents usages. Des corpus indépendants auront peu de vocabulaire représentatif en commun et surtout le risque que ce vocabulaire ait des usages différents au sein d'un même corpus ou entre les deux corpus (critères 3 et 4).

6.3 Conclusion

Cette discussion propose de revoir la notion de comparabilité, notamment au regard des usages faits des corpus comparables, dans le cas de l'extraction lexicale bilingue, mais également au regard des mesures statistiques employés dans l'approche directe. Ces mesures nécessitent, pour être efficaces, le respect de certaines hypothèse que l'on ne retrouve pas dans les corpus comparables.

La notion de comparabilité ne doit donc pas se voir seulement à travers les contraintes externes des documents utilisés pour constituer les corpus (tel que le thème ou le registre), mais aussi à travers des contraintes internes. Une première contrainte concerne le vocabulaire commun disponible entre les deux corpus : il doit être suffisamment abondant, en quantité et en qualité. D'une part, pour qu'un corpus puisse être exploité pour l'extraction lexicale, il faut un nombre suffisant de mots en commun, mots que nous allons chercher à aligner d'un corpus à l'autre. D'autre part, il faut que les éléments de ce vocabulaire commun soient suffisamment bien représentés pour être comparables. Nous avons montré que les mots très fréquents étaient représentés d'une façon relativement stable (bien que la façon dont ils sont comparés puisse s'avérer inadéquate) mais que les mots plus rares étaient sujets à une plus grande instabilité, rendant difficile leur exploitation en utilisant les méthodes par comparaison de contexte.

Au-delà du vocabulaire commun et de sa représentativité, nous ajoutons une contrainte sur les usages faits de ce vocabulaire commun, en pointant en particulier le risque de cas d'incomparabilité pour des termes ayant un sens sensiblement différent en fonction du contexte des documents, voire au sein même des documents. La notion de comparabilité se mesure donc à plusieurs niveaux : des contraintes externes sur la sélection des documents (ce qui est traditionnellement fait) mais également des contraintes internes liées à un premier degré au vocabulaire, à un second degré à l'emploi de ce vocabulaire.

Ces réflexions se veulent être un point de départ à un nouveau regard sur les corpus comparables et la comparabilité ; nous avons cherché les problèmes de l'approche actuelle et comment les contourner, par exemple en expliquant le succès de certaines méthodes et en proposant quelques pistes pour exploiter des corpus *moins comparables*. Les définitions proposées ici ont aussi pour objectif de rendre la conception de corpus comparables plus simple et leur exploitation plus robuste, en particulier en retirant la contrainte d'équilibre entre les parties d'un corpus, injustifiée, ou en permettant d'introduire des documents non-pertinents, à la condition de respecter quelques contraintes dans leur exploitation.

Conclusion générale

Nos recherches ont porté sur les corpus comparables, définis comme des ensembles de textes dans des langues différentes n'étant pas en relation de traduction. Nous nous sommes en particulier intéressés aux corpus comparables spécialisés issus du domaine médical, dans le but de comprendre et d'améliorer le processus d'extraction lexical bilingue, c'est-à-dire le processus consistant à reconnaître un vocabulaire commun entre les différents sous-corpus et à l'aligner pour constituer automatiquement des lexiques. Dans notre cas, nous avons exploité un corpus portant sur le thème *diabète et alimentation* et un autre traitant du *cancer du sein*. Les documents de ces deux corpus sont par ailleurs tous de type *scientifique*, c'est-à-dire écrits par des experts à destination d'autres experts. Les contraintes de construction de ces corpus nous ont permis de nous intéresser en particulier à la terminologie de ces domaines, pour chercher à constituer des ressources lexicales spécialisées. C'est une problématique intéressante puisque la terminologie des langues de spécialité est généralement moins bien connue que le lexique général : l'extraction lexicale bilingue se veut une aide pour le lexicographe et le terminologue, dans le but de les assister pour constituer des ressources linguistiques plus précises et pertinentes.

Le corpus *cancer du sein* regroupe des documents en français et en anglais alors que le corpus *diabète et alimentation* contient des documents en français, anglais et japonais. Ce dernier corpus nous a permis d'étudier le cas difficile de l'alignement impliquant des langues très différentes, comme le français et le japonais. Par ailleurs, ces deux corpus sont de tailles relativement modestes, de l'ordre de centaines de milliers de mots alors qu'on rencontre fréquemment des corpus beaucoup plus volumineux dans la littérature. Nous dressons ici un bilan des points principaux abordés dans ce document.

Exploitation des corpus multilingues

Les premiers chapitres se sont concentrés sur la définition et l'exploitation des corpus multilingues. Nous avons notamment présenté différents types de corpus multilingues : les corpus parallèles et les corpus comparables. Nous avons introduit différentes façons de les exploiter en insistant sur l'approche directe, support de nos expériences. Nous avons également repris des éléments utilisés dans un cadre monolingue, en particulier la *caractérisation sémantique*, pour montrer que c'est une problématique proche de la nôtre. Nous avons pu revenir plus en détail sur la notion d'*association* entre mots, utilisée dans un cadre monolingue pour la détection de collocations et dans un cadre multilingue pour la comparaison des vecteurs de contexte. Nous avons mis en évidence quelques propriétés des différentes mesures d'association présentées, ce qui nous a permis par la suite de justifier ou d'invalider le choix d'une mesure *a priori*, problématique généralement laissée de côté dans la littérature. Nous avons enfin détaillé un exemple de processus d'extraction lexicale bilingue à partir de corpus comparables, en présentant notre implémentation de l'approche directe sur le corpus *cancer du sein*. Cette description a notamment permis d'étalonner notre chaîne de traitement, pour présenter des résultats de référence pour notre étude, auxquels nous opposons les résultats obtenus avec nos différentes propositions.

Propositions

Nous avons présenté, à travers les expériences réalisées, plusieurs méthodes pour améliorer la caractérisation terminologique multilingue et ainsi l'extraction lexicale bilingue.

L'approche exploitant la fréquence des mots à traduire s'appuie sur l'observation des résultats de l'approche directe et met en évidence une corrélation entre la fréquence d'un mot et la taille de fenêtre optimale utilisée pour le caractériser. Cette approche exploite les dépendances syntagmatiques et paradigmatiques. Elle répond à la problématique de la caractérisation terminologique multilingue en mettant en évidence des méthodes pour améliorer cette caractérisation. Cette approche s'est par ailleurs révélée efficace non seulement pour l'alignement de termes, mais aussi pour l'alignement de mots moins caractéristiques. L'atout de cette approche est de permettre la sélection et la combinaison *a priori* de certains paramètres de l'approche directe, à partir d'une propriété facile à mesurer des mots à traduire.

L'approche utilisant les points d'ancrage consiste à transformer les motifs d'association des termes à aligner, en donnant plus de poids à une partie de leur contexte. Les points d'ancrage sont sélectionnés automatiquement en raison de leur fiabilité pour décrire un terme. Ils recouvrent eux-mêmes une terminologie propre à la thématique des documents exploités et ont l'avantage de ne pas être ambigus. Ces points d'ancrage se sont révélés efficaces pour l'alignement vers le japonais. Dans ce cas, nous disposons de jeux de données relativement restreints, à quoi s'ajoutent les difficultés liées au traitement du japonais et à la distance entre cette langue, le français et l'anglais. L'exploitation des points d'ancrage a permis une amélioration significative des résultats de l'approche directe grâce à une amélioration de la caractérisation des termes à traduire.

L'approche multi-sources, utilisée pour l'alignement vers le japonais, consiste à combiner les points de comparaisons construits pour plusieurs langues sources pour l'alignement vers une langue cible. En exploitant les résultats des alignements français-japonais et anglais-japonais, nous avons montré une amélioration conséquente de la qualité des traductions obtenues pour la terminologie japonaise. Cette méthode, très simple, permet de découvrir la terminologie d'une langue à partir de la terminologie, connue, d'autres langues et peut vraisemblablement s'étendre facilement à d'autres couples. Cette approche ne modifie pas la caractérisation des termes, mais cherche à exploiter au mieux les points de comparaison disponibles en prenant le meilleur des différentes caractérisations terminologiques.

Ces propositions se placent dans le cadre de l'alignement multilingue en utilisant des corpus comparables spécialisés de tailles modestes, en accord avec les matériaux textuels à notre disposition et avec l'hypothèse énoncée dans Morin *et al.* (2007). Nous avons développé et éprouvé ces approches en gardant en tête le problème du choix *a priori* des paramètres optimaux pour l'alignement. Nous avons obtenu un premier succès, en choisissant une taille de fenêtre contextuelle à partir de la fréquence des mots dont on cherche une traduction. Nous avons proposé au chapitre 6 une discussion née du choix des mesures d'association à utiliser pour la constitution des vecteurs de contexte.

Comparabilité des corpus comparables

La discussion du chapitre 6 propose de réviser la notion de *comparabilité*, traditionnellement vue comme *le nombre de traits communs entre les documents constituants les sous-corpus* ou *la quantité de vocabulaire commun disponible entre les sous-corpus*. Nous émettons l'idée que cette comparabilité doit plutôt être opposée à l'*incomparabilité*, qui elle se mesure aux nombre d'éléments incomparables. Il s'agit d'éléments communs entre les différentes parties d'un corpus, mais utilisés avec un sens différent.

Cette discussion découle d'une observation en corpus et d'une réflexion sur le choix des mesures d'association, notamment pour expliquer pourquoi la contrainte d'équilibre entre les différentes parties

d'un corpus était superflue. L'hypothèse principale de cette partie est qu'il doit être possible d'exploiter des corpus structurellement différents, avec des résultats similaires. Nous pensons qu'un corpus comparable donné, et le même corpus augmenté de données non reliées devraient pouvoir être exploités avec la même efficacité en utilisant des méthodes classiques. En allant plus loin, un corpus comparable donné et le même augmenté de données *incomparables* devraient eux aussi pouvoir être exploités avec des résultats semblables, pour peu que les zones d'incomparabilité puissent être détectées et traitées comme telles *a priori*. Ce travail de dépistage de données incomparables semble pertinent même dans le cas de corpus fortement comparables puisque, nous l'avons montré, même dans des corpus très contraints (comme le corpus *cancer du sein*) il est possible de trouver des éléments d'incomparabilité.

Ouverture

Les corpus comparables restent des objets d'études relativement nouveaux et la communauté de chercheurs s'y intéressant est naissante. Beaucoup reste à faire pour les caractériser et les exploiter. Nous avons essayé de faire les deux en proposant des améliorations de l'approche directe puis en discutant la notion de *comparabilité*.

Au vue des résultats présentés, de nos observations, et en lien avec la discussion du chapitre 6, certaines barrières paraissent difficiles à franchir en utilisant les méthodes classiques. Ainsi, bien qu'il reste sans doute beaucoup de place pour améliorer la recherche de paires de traduction à partir de corpus comparables, il nous paraît plus astucieux de chercher à multiplier les jeux de données que de chercher à extraire plus d'informations à partir d'un jeu de données en particulier : les termes peu fréquents sont trop délicats à caractériser avec les méthodes présentées dans cette étude. Ils sont trop peu représentés pour définir suffisamment de points de comparaison discriminants. Toutefois, la discussion du chapitre 6 ouvre des perspectives pour l'exploitation de corpus comparables moins contraints, donc plus faciles à obtenir. Cela peut se faire par exemple en exploitant des alignements multi-sources, comme nous l'avons fait au chapitre 5, mais aussi en créant plus de jeux de données hétérogènes. L'exploitation de ces différentes ressources permettrait de valider ou d'invalider des candidats à la traduction, mais aussi d'obtenir des jeux de ressources mieux représentatifs d'un certain vocabulaire.

Il peut s'agir d'augmenter le volume des corpus, en gardant à l'esprit la notion d'hétérogénéité développée précédemment, ou peut-être plus efficacement de multiplier les exemples d'alignement pour révéler les plus pertinents, par exemple, les plus stables entre différents jeux de ressources ou de paramètres. Autrement dit, nous ne prônons pas l'approche consistant à augmenter le *volume* ou la *qualité* des données, mais plutôt une approche consistant à augmenter la *variété* des données.

Bibliographie

ABDUL-RAUF, S. et SCHWENK, H. (2009). On the use of comparable corpora to improve SMT performance. *In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*, pages 16–23.

AGAFONOV, C., GRASS, T., MAUREL, D. et ROSSI-GENSANE, N. (2006). La traduction multilingue des noms propres dans prolex. *In* CLAS, A., directeur de la publication : *La traduction des noms propres et Langue, traduction et mondialisation : interactions d'hier, interactions d'aujourd'hui*, volume 51, numéro 4, pages 622–636. Presses de l'Université de Montréal.

AIJMER, K., ALTENBERG, B. et JOHANSSON, M. (1996). *Languages in Contrast*. Lund University Press.

BERRY, M. W., DUMAIS, S., O'BRIEN, G., BERRY, M. W., DUMAIS, S. T. et GAVIN (1995). Using Linear Algebra for Intelligent Information Retrieval. *SIAM Review*, 37:573–595.

BLANK, I. (2000). Terminology extraction from parallel technical texts. *In* VÉRONIS, J., directeur de la publication : *Parallel Text Processing*, pages 337–252. Kluwer Academic Publishers.

BODENREIDER, O. et ZWEIGENBAUM, P. (2000). Stratégies d'identification des noms propres à partir de nomenclatures médicales parallèles. *Traitement automatique des langues*, 41(3).

BONHOMME, P. et ROMARY, L. (1995). Projet de concordances parallèles lingua : gestion de textes multilingues pour l'apprentissage des langues. *In Quizième Journées Internationales IA 95*.

BOWKER, L. et PEARSON, J. (2002). *Working with Specialized Language : A Practical Guide to Using Corpora*. Routledge, London/New York.

BROWN, P. F., COCKE, J., DELLA PIETRA, S. A., DELLA PIETRA, V. J., JELINEK, F., LAFFERTY, J. D., MERCER, R. L. et ROOSSIN, P. S. (1990). A statistical approach to machine translation. *Computational Linguistic*, 16(2):79–85.

BROWN, P. F., PIETRA, S. A. D., PIETRA, V. J. D. et MERCER, R. L. (1993). The mathematics of statistical machine translation : Parameter estimation. *Computational Linguistics*, 19(2):263–310.

CHIAO, Y.-C. (2004). *Extraction lexicale bilingue à partir de textes médicaux comparables : application à la recherche d'information translangue*. Thèse en Informatique, Université Pierre et Marie Curie, Paris VI.

CHIAO, Y.-C. et ZWEIGENBAUM, P. (2002). Looking for candidate translational equivalents in specialized, comparable corpora. *In Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 1208–1212.

CLAVEAU, V. (2007). Inférence de règles de réécriture pour la traduction de termes biomédicaux. *In Actes de la conférence Traitement Automatique des Langues Naturelles (TALN'07)*, pages 111–120.

CREGO, J. M., MAX, A. et YVON, F. (2009). Plusieurs langues (bien choisies) valent mieux qu'une : traduction statistique multi-source par renforcement lexical. *In Actes de la conférence Traitement Automatique des Langues Naturelles (TALN'09)*, pages 0–10.

CRETTEZ, J.-P. et LORETTE, G. (1998). *Reconnaissance de l'écriture manuscrite*, chapitre 0, pages 1–15. Technique de l'Ingénieur.

DAGAN, I., ITAI, A. et SCHWALL, U. (1991). Two languages are more informative than one. *In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL'91)*, pages 130–137.

DAGAN, I., LEE, L., et PEREIRA, F. C. N. (1999). Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34(1-3):43–69.

DAILLE, B. (1994). *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. Thèse de doctorat, Université Paris 7.

DAILLE, B. (2002). Terminology mining. *In* PAZIENZA, M. T., directeur de la publication : *Information Extraction in the Web Era, Lecture Notes in Artificial Intelligence (LNAI)*, pages 29–44. Springer.

DAILLE, B., GAUSSIER, É. et LANGE, J.-M. (1994). Towards automatic extraction of monolingual and bilingual terminology. *In Proceedings of the 15th conference on Computational Linguistics (COLING'94)*, pages 515–521, Morristown, NJ, États-Unis d'Amérique.

DEERWESTER, S., DUMAIS, S., FURNAS, G. W., LANDAUER, T. K. et HARSHMAN, R. (1990). Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 41(6):391–407.

DUNNING, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74.

DÉJEAN, H. et GAUSSIER, E. (2002). Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *In* VÉRONIS, J., directeur de la publication : *Lexicometrica, Alignement lexical dans les corpus multilingues*, pages 1–22.

DÉJEAN, H., SADAT, F. et Éric GAUSSIER (2002). An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. *In Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 218–224.

ESSEN, U. et STEINBISS, V. (1992). Cooccurrence smoothing for stochastic language modeling. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1:161–164.

EVERT, S. (2004). *The Statistics of Word Cooccurrences*. Thèse de doctorat, Université de Stuttgart.

EVERT, S. (2008). Corpora and collocations. *In* LÜDELING, A. et KYTÖ, M., direction de la publication : *Corpus Linguistics. An International Handbook*, chapitre 58. Mouton de Gruyter, Berlin.

FIRTH, J. (1957). *A synopsis of linguistic theory 1930-1955*. Studies in Linguistic Analysis, Philological. Longman.

FUNG, P. (1995a). Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. *In* YAROVSKY, D. et CHURCH, K., direction de la publication : *Proceedings of the 3rd Workshop on Very Large Corpora (VLC'95)*, pages 173–183.

FUNG, P. (1995b). A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. *In Proceedings of the 33rd Meeting of the Association for Computational Linguistics (ACL'95)*, pages 236–243.

FUNG, P. (1998). A statistical view on bilingual lexicon extraction : From parallel corpora to non-parallel corpora. *In* FARWELL, D., GERBER, L. et HOVY, E. H., direction de la publication : *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA'98)*, pages 1–17.

GALE, W. A. et CHURCH, K. W. (1991). Identifying word correspondence in parallel texts. *In HLT '91 : Proceedings of the workshop on Speech and Natural Language*, pages 152–157.

GAO, W., WONG, K.-F. et LAM, W. (2004). Phoneme-based transliteration of foreign names for OOV problem. *In International Joint Conference on Natural Language Processing (IJCNLP'04)*, pages 110–119.

GAUSSIER, E., RENDERS, J.-M., MATVEEVA, I., GOUTTE, C. et DÉJEAN, H. (2004). A Geometric View on Bilingual Lexicon Extraction from Comparable Corpora. *In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04)*, pages 526–533.

GOEURIOT, L., GRABAR, N. et DAILLE, B. (2008). Characterization of scientific and popular science discourse in french, japanese and russian. *In Proceedings of the 6th edition of Language Resources and Evaluation Conference (LREC'08)*.

GREFENSTETTE, G. (1994a). Corpus-derived first, second and third-order word affinities. *In Acts, 6th International Congress of the European Association for Lexicography (EURALEX'94)*, pages 279–290.

GREFENSTETTE, G. (1994b). *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publisher.

GREFENSTETTE, G. (1996). Evaluation techniques for automatic semantic extraction : comparing syntactic and window based approches. *In* BOGURAEV, B. et PUSTEJOVSKY, J., direction de la publication : *Corpus processing for lexical acquisition*, chapitre 11, pages 205–216. MIT Press.

GREFENSTETTE, G. (1999). The world wide web as a resource for example-based machine translation tasks. *In Proceedings of Aslib Conference on Translating and the Computer (ASLIB'99)*.

GRUNDY, V. (1996). L'utilisation d'un corpus dans la rédaction du dictionnaire bilingue. *In* BÉJOINT, H. et THOIRONS, P., direction de la publication : *Les dictionnaires bilingues*, pages 127–149. Duculot, Louvain-la-Neuve.

HABERT, B., NAZARENKO, A. et SALEM, A. (1997). *Les linguistiques de corpus*. Armand Colin, Paris.

HABERT, B. et ZWEIGENBAUM, P. (2002). Régler les règles. *Traitement Automatique des Langues*, 43(3):83–105.

HAKUSUISHA, directeur de la publication (1989). *French-Japanese Scientific Dictionary*. 4th edition édition.

HAMMING, R. (1950). Error-detecting and error-correcting codes. *Bell System Technical Journal*, 29(2):147–160.

HARRIS, Z. (1988). *Language and information*. Columbia University Press, New-York, NY, États-Unis d'Amérique.

HARUNO, M. et YAMAZAKI, T. (1996). High-performance bilingual text alignment using statistical and dictionary information. *In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL'96)*, pages 131–138.

HINDLE, D. (1990). Noun classification from predicate-argument structures. *In Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics (ACL'90)*, pages 268–275.

HOFMANN, T. (1999). Probabilistic latent semantic analysis. *In* LASKEY, K. et PRADE, H., direction de la publication : *In Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI'99)*, pages 289–296.

HUMBLEY, J. (2006). La traduction des noms d'institutions. *In* CLAS, A., directeur de la publication : *La traduction des noms propres et Langue, traduction et mondialisation : interactions d'hier, interactions d'aujourd'hui*, volume 51, numéro 4, pages 671–689. Presses de l'Université de Montréal.

ITO, M. (2007). Loan words and foreign words in 90 magazines and 70 magazines. *In The 51st Annual Meeting of the Mathematical Linguistic Society of Japan.*

JACQUEMIN, C. et BOURIGAULT, D. (2003). Term extraction and automatic indexing. *In* MITKOV, R., directeur de la publication : *Handbook of Computational Linguistics*, pages 599–615. Oxford University Press.

JAGTMAN, M. (1994). COMOLA : A computer system for the analysis of interlanguage data. *Second Language Research, 10*, pages 49–83.

KAGEURA, K. (2003). The dynamics of morphemes in Japanese terminology. *Journal of Natural Language Processing*, 10(4):125–144.

KATZ, J. J. et FODOR, J. A. (1963). The structure of semantic theory. *Language*, 39:170–210.

KAY, M. et RÖSCHEISEN, M. (1988). Text-translation alignment. Rapport technique, Xerox Palo Alto Research Center.

KILGARRIFF, A. (2005). Language is never ever ever random. *Corpus Linguistics and Linguistic Theory*, 1(2):263–276.

KNIGHT, K. et GRAEHL, J. (1997). Machine transliteration. *In* COHEN, P. R. et WAHLSTER, W., direction de la publication : *Proceedings of the 3rd Annual Meeting of the Association for Computational Linguistics (ACL'97) and 8th Conference of the European Chapter of the Association for Computational Linguistics (EACL'97)*, pages 128–135.

KNOWLES, F. (1996). L'informatisation de la fabrication des dictionnaires bilingues. *In* BÉJOINT, H. et THOIRONS, P., direction de la publication : *Les dictionnaires bilingues*, pages 151–168. Duculot, Louvain-la-Neuve.

KOEHN, P. et KNIGHT, K. (2002). Learning a translation lexicon from monolingual corpora. *In Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition*, pages 9–16. Association for Computational Linguistics.

KOTANI, T. et KORI, A. (1990). *Dictionary of Technical Terms*. Kenkyusha.

LANGÉ, J.-M. et GAUSSIER, E. (1995). Alignement de corpus multilingues au niveau des phrases. *Traitement Automatique des Langues*, 36(1–2):67–80.

LEVENSHTEIN, V. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707–710.

LEWIS, D. (1972). General semantics. *In* DAVIDSON, D. et HARMAN, G., direction de la publication : *Semantics of natural language*, pages 169–218. Dordrecht, Reidel.

LEWIS, D. (2005). Corpus comparables et analyse contrastive : l'apport d'un corpus français/anglais de discours politiques à l'analyse des connecteurs adversatifs. *In* WILLIAMS, G., directeur de la publication : *La linguistique de corpus*, pages 179–190. Presse Universitaire de Rennes.

L'HOMME, M.-C. (2004). *La terminologie : principes et techniques*. Presses de l'Université de Montréal.

LOVIS, C., BAUD, R., MICHEL, P. A., SCHERRER, J. R. et RASSINOUX, A. M. (1997). Building medical dictionaries for patient encoding systems : A methodology. *Lecture Notes in Computer Science*, 1211:373–380.

MANDELBROT, B. (1965). Information theory and psycholinguistics. *In* WOLMAN, B. et NAGEL, E., direction de la publication : *Scientific psychology*. Basic Books.

MANNING, C. D. et SCHÜTZE, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, États-Unis d'Amérique.

MORIN, E. (2009). Apport d'un corpus comparable déséquilibré à l'extraction de lexiques bilingues. *In Actes de la 16ème Conférence Traitement Automatique des Langues Naturelles (TALN'09)*.

MORIN, E. et DAILLE, B. (2004). Extraction terminologique bilingue à partir de corpus comparables d'un domaine spécialisé. *Traitement Automatique des Langues (TAL)*, 45(3):103–122.

MORIN, E. et DAILLE, B. (2008). An effective compositional model for lexical alignment. *In Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP'08)*, pages 95–102.

MORIN, E., DAILLE, B., TAKEUCHI, K. et KAGEURA, K. (2007). Bilingual terminology mining – using brain, not brawn comparable corpora. *In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL'07)*, pages 664–671.

NAKAMURA-DELLOYE, Y. (2007). *Alignement automatique de textes parallèles français-japonais*. Thèse de doctorat, Université Paris 7.

NAMER, F. (2005). Morphosémantique pour l'appariement de termes dans le vocabulaire médical : approche multilingue. *In Actes de la conférence Traitement Automatique des Langues Naturelles (TALN'05)*, pages 63–72.

NAMER, F. et ZWEIGENBAUM, P. (2004). Acquiring meaning for french medical terminology : contribution of morphosemantics. *In* FIESCHI, M., COIERA, E. et LI, Y.-C. J., direction de la publication : *Studies in Health Technology and Informatics*, volume 107, pages 535–539.

OCH, F. J. et NEY, H. (2001). Statistical multi-source translation. *In Proceedings of MT Summit*, pages 253–258.

PASTOR, G. C., MITKOV, R., AFZAL, N. et MOYA, L. G. (2008). Translation universals : do they exist ? a corpus-based and nlp approach to convergence. *In Proceedings of the Workshop on Comparable Corpora, Language Resources and Evaluation Conference (LREC'08)*.

PEARSON, J. (1998). *Terms in Context*. John Benjamins publishing company.

PEARSON, J. (1999). Comment accéder aux éléments définitoires dans les textes spécialisés. *Terminologies nouvelles*, 19:21–28.

PEKAR, V., MITKOV, R., BLAGOEV, D. et MULLONI, A. (2006). Finding translations for low-frequency words in comparable corpora. *Machine Translation*, 20(4):247–266.

PIENEMANN, M. (1992). COALA - a computational system for interlanguage analysis. *Second Language Research, 8*, pages 59–92.

PROCHASSON, E., KAGEURA, K., MORIN, E. et AIZAWA, A. (2008). Looking for transliterations in a trilingual english, french and japanese specialised comparable corpus. *In Proceedings of the 1st Workshop on Building and Using Comparable Corpora, Language Resources and Evaluation Conference (LREC'08)*, pages 83–86.

PROCHASSON, E. et MORIN, E. (2009a). Influence des points d'ancrage pour l'extraction lexicale bilingue à partir de corpus comparables spécialisés. *In Conférences sur le Traitement Automatique des Langues Naturelles (TALN'09)*, page 10.

PROCHASSON, E. et MORIN, E. (2009b). Points d'ancrage pour l'extraction lexicale bilingue à partir de petits corpus comparables spécialisés. *Traitement Automatique des Langues (TAL)*, pages 283–304.

PROCHASSON, E., MORIN, E. et KAGEURA, K. (2009). Anchor points for bilingual lexicon extraction from small comparable corpora. *In Machine Translation Summit 2009*, page 8.

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

RAPP, R. (1999). Automatic Identification of Word Translations from Unrelated English and German Corpora. *In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 519–526.

ROBITAILLE, X., SASAKI, Y., TONOIKE, M., SATO, S. et UTSURO, T. (2006). Compiling french-japanese terminologies from the web. *In Proceedings of 11st Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*.

ROMARY, L. et BONHOMME, P. (2000). Parallel alignment of structured documents. *In* VÉRONIS, J., directeur de la publication : *Parallel Text Processing*, pages 201–218. Kluwer Academic Publisher.

SADAT, F., YOSHIKAWA, M. et UEMURA, S. (2003). Learning bilingual translations from comparable corpora to cross-language information retrieval : Hybrid statistics-based and linguistics-based approach. *In Proceedings of the Sixth Internation Workshop on Information Retrieval with Asian Languages (IRAL'03)*, pages 57–64.

SAGER, J. C. (1990). *A practical course in terminology processing*. John Benjamins, Amsterdam, Pays-Bas.

SAGER, N. (1986). Analyzing language in restricted domains. sublanguage description and processing. *In* GRISHAM, R. et KITTREDGE, R., direction de la publication : *Sublanguages : Linguistic Phenomenon, computational tool*, pages 1–18. Lawrence Erlbaum Associates.

SALTON, G. et LESK, M. E. (1968). Computer evaluation of indexing and text processing. *In Journal of the Association for Computational Machinery*, volume 15(1), pages 8–36.

SHAO, L. et NG, H. T. (2004). Mining new word translations from comparable corpora. *In Proceedings of the 20th International Conference on Computational Linguistics (COLING'04)*, pages 618–624.

SHAROFF, S. (2006). Open-source corpora : using the net to fish for linguistic data. *International Journal of Corpus Linguistics*, 11(4):435–462.

SHAROFF, S., BABYCH, B. et HARTLEY, A. (2006). Using comparable corpora to solve problems difficult for human translators. *In Proceedings of the joint COLING-ACL Conference*.

SHERIF, T. et KONDRAK, G. (2007). Bootstrapping a stochastic transducer for Arabic-English transliteration extraction. *In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 864–871.

SIMARD, M., FOSTER, G. F. et ISABELLE, P. (1992). Using cognates to align sentences in bilingual corpora. *In Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'92)*, pages 67–81.

SINCLAIR, J. (1996). Preliminary recommendations on corpus typology. Rapport technique, Expert Advisory Group on Language Engineering Standards (EAGLE).

SPROAT, R., TAO, T. et ZHAI, C. (2006). Named entity transliteration with comparable corpora. *In ACL-44 : Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 73–80.

TANIMOTO, T. (1958). An elementary mathematical theory of classification. Rapport technique, IBM Research.

TAO, T., YOON, S.-Y., FISTER, A., SPROAT, R. et ZHAI, C. (2006). Unsupervised named entity transliteration using temporal and phonetic correlation. *In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP'06)*, pages 250–257.

TAYLOR, P., BLACK, A. et CALEY, R. (1998). The architecture of the the festival speech synthesis system. *In 3rd International Workshop on Speech Synthesis*.

TILLMANN, C. et NEY, H. (2003). Word reordering and a dynamic programming beam search algorithm for statistical machine translation. *Computational Linguistics*, 29(1):97–133.

TSUJI, K., DAILLE, B. et KAGEURA, K. (2002). Extracting French-Japanese word pairs from bilingual corpora based on transliteration rules. *In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'02)*, pages 499–502.

TSUJI, K., SATO, S. et KAGEURA, K. (2005). Evaluating the effectiveness of transliteration and search engines in bilingual proper name identifications. *In The 11th Annual Meeting of the Association for Natural Language Processing (ANLP'05)*, pages 352–355.

TUTIN, A. et GROSSMANN, F. (2002). Collocations régulières et irrégulières : esquisse du phénomène collocatif. *Revue française de linguistique appliquée*, 7(1).

VIRGA, P. et KHUDANPUR, S. (2003). Transliteration of proper names in cross-lingual information retrieval. *In Proceedings of the ACL workshop on Multi-Lingual Named Entity Recognition*.

VOGT, C. et COTTRELL, G. W. (1998). Predicting the performance of linearly combined IR systems. *In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98)*, pages 190–196.

VÉRONIS, J., directeur de la publication (2000). *Parallel Text Processing*. Kluwer Academic Publishers.

VÉRONIS, J. et LANGLAIS, P. (2000). Evaluation of parallel text alignment systems. *In* VÉRONIS, J., directeur de la publication : *Parallel Text Processing*, pages 369–388. Kluwer Academic Publishers.

WITTEN, I. H. et BELL, T. C. (1991). The zero-frequency problem : Estimating the probabilities of novel events in adaptive text compression. *IEEE Transations on Information Theory*, 37(4):1085–1094.

YASER, A.-O. et KNIGHT, K. (2002). Translating named entities using monolingual and bilingual resources. *In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL'02)*, pages 400–408.

YOON, S.-Y., KIM, K.-Y. et SPROAT, R. (2007). Multilingual transliteration using feature based phonetic method. *In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL'07)*, pages 112–119.

ZANETTIN, F. (1998). Bilingual corpora and the training of translators. *Meta*, 43:616–630.

ZWEIGENBAUM, P. et HABERT, B. (2006). Faire se rencontrer les parallèles : regards croisés sur l'acquisition lexicale monolingue et multilingue. *Revue de sociolinguistique en ligne GLOTTOPOL*, 8:22–44.

Liste des tableaux

— Corps du document —

Liste des figures

— Corps du document —

Crédit photo

La photo de la Pierre de Rosette, présentée au chapitre 1 est disponible sous licence *Creative Commons Attribution ShareAlike 2.5*. Elle a été publiée par l'utilisateur Hans Hillewaert le 21 novembre 2007 sur *Wikimedia Commons*[a]. Elle peut être utilisée, modifiée et redistribuée sous les termes de la même licence. L'image du chapitre 2 est un extrait du manga *GTO* réalisé par Tōru Fujisawa et édité par Kōdansha.

[a] `http://commons.wikimedia.org`.

Table des matières

— Corps du document —

Alignement multilingue en corpus comparables spécialisés

Caractérisation terminologique multilingue

Emmanuel PROCHASSON

Résumé

Les corpus comparables rassemblent des documents multilingues n'étant pas en relation de traduction mais partageant des traits communs. Notre travail porte sur l'extraction de lexique bilingue à partir de ces corpus, c'est-à-dire la reconnaissance et l'alignement d'un vocabulaire commun multilingue disponible dans le corpus. Nous nous concentrons sur les corpus comparables spécialisés, c'est-à-dire des corpus constitués de documents révélateurs de la terminologie utilisée dans les langues de spécialité. Nous travaillons sur des corpus médicaux, l'un deux couvre la thématique du diabète et de l'alimentation, en français, anglais et japonais ; l'autre couvre la thématique du cancer du sein, en anglais et en français. Nous proposons et évaluons différentes améliorations du processus d'alignement, en particulier dans le cas délicat de la langue japonaise. Nous prolongeons ce manuscrit par une réflexion sur la nature des corpus comparables et la notion de comparabilité.

Mots-clés : corpus comparables, langue de spécialité, alignement multilingue

Multilingual alignment from specialised comparable corpora

Multilingual terminology characterisation

Abstract

Comparable corpora are sets of documents written in different languages, which are not translations of each other but share common features, such as the topic or the discourse type. Our work concerns bilingual lexicon extraction from such corpora, in other word, the process of finding translation pairs among the common multilingual vocabulary available in comparable corpora. We focus on specialised comparable corpora, for they are likely to reveal the terminology proper to specialised language. We work on corpora made of medical documents: one of them covers the topic of diabetes and feeding, in French, English and Japanese; the other one covers the topic of breast cancer, in French and English. We propose several improvements for the classical alignment process, especially concerning the delicate case of the Japanese language, distant from French and English. We conclude this thesis with thoughts concerning the nature of comparable corpora and the question of comparability.

Keywords: comparable corpora, specialised language, multilingual aligment

Discipline : Informatique
Spécialité : Traitement Automatique du Langage Naturel

Laboratoire : LABORATOIRE D'INFORMATIQUE DE NANTES ATLANTIQUE.
CNRS UMR 6241. 2, rue de la Houssinière, BP 92 208 – 44 322 Nantes, CX 3.