Points d'ancrage pour l'extraction lexicale bilingue à partir de petits corpus comparables spécialisés

Éléments de confiance pour la caractérisation des termes

Emmanuel Prochasson — **Emmanuel Morin**

Université de Nantes, LINA - UMR CNRS 6241
2, rue de la Houssinière, BP 92208
F-44 322 Nantes cedex 3
{emmanuel.prochasson,emmanuel.morin}@univ-nantes.fr

RÉSUMÉ. *Les recherches en extraction lexicale bilingue à partir de corpus comparables ont abouti à des résultats prometteurs pour les corpus très volumineux en utilisant une méthode d'alignement dite* directe. *Le changement d'échelle induit par des corpus d'une taille plus modeste conduit à l'obtention de résultats plus contrastés. Nous proposons d'introduire la notion de points d'ancrage sur laquelle nous faisons reposer une partie de l'alignement pour augmenter significativement les résultats de l'approche directe sur de tels corpus. Nous avons choisi de nous concentrer sur les translittérations et les mots savants comme points d'ancrage, sur un petit corpus comparable spécialisé. Nous montrons comment nous les avons exploités, ainsi que leur influence sur les candidats à la traduction.*

ABSTRACT. *Research on bilingual lexicon extraction from comparable corpora leads to promising results using large corpora (hundreds of billions of words) using the* direct *alignment method. However, when using smaller corpora (hundreds of thousands of words), results obtained are slightly lower. We propose to introduce some anchor points on which we can rely for the alignment process using the direct approach on small corpora and show how they influence the translation candidates.*

MOTS-CLÉS : *corpus comparables, approche directe, points d'ancrage, fouille de données, extraction de lexiques bilingues.*

KEYWORDS: *comparable corpora, direct approach, anchor points, text mining, bilingual lexicon extraction.*

1. Introduction

Le principal objectif de l'extraction lexicale bilingue est de compléter et de maintenir à jour des ressources linguistiques. Ces ressources seront des aides précieuses pour les traducteurs humains, qui pourront profiter de données actualisées et spécialisées. Les recherches en extraction lexicale à partir de corpus multilingues se sont largement concentrées sur les corpus *parallèles*, c'est-à-dire des ensembles de couples de documents en correspondance de traduction (Véronis, 2000). Les approches proposées s'appuient principalement sur des hypothèses fortes, notamment l'hypothèse que chaque élément (mot, phrase, paragraphe) d'un document *source* aura une correspondance, traduite, dans le document *cible*. Corollairement, l'organisation du document *cible* sera similaire à l'organisation du document *source* (ordre des phrases, des paragraphes, découpage en parties). Ainsi, en comparant la distribution d'un élément à aligner et les distributions des candidats dans le document traduit, il est possible de discriminer le candidat correct qui aura la distribution la plus similaire à l'élément à traduire. Il est également possible de s'appuyer sur des points d'ancrage, c'est-à-dire les éléments déjà traduits avec certitude pour aligner leurs voisins puisque, par hypothèse, des éléments voisins dans le document *source* seront voisins dans le document *cible* (Véronis, 2000).

L'emploi des corpus parallèles présente toutefois plusieurs faiblesses. La première concerne la disponibilité de ces corpus, qui nécessitent une traduction humaine coûteuse et rarement disponible entre couples de langues n'impliquant pas l'anglais. En outre, il est difficile de trouver des corpus parallèles en quantité suffisante, même entre langues largement utilisées comme le chinois, le français ou l'allemand. Le second point faible des corpus parallèles vient de leur nature même de traduction. L'organisation du document traduit est directement issue du document source et la traduction ne reflète pas alors l'emploi naturel d'une langue, mais uniquement les phénomènes propres au passage d'une langue à une autre. (Fung, 1995) ajoute que le résultat d'une extraction bilingue à partir de textes traduits est au mieux de l'ingénierie à rebours sur le lexique utilisé par le traducteur[1].

Face à ces inconvénients, les recherches se sont penchées sur l'exploitation de corpus *comparables*, c'est-à-dire des ensembles de textes dans des langues différentes qui présentent des traits communs (thème, période, auteur, etc.) sans toutefois être des traductions mutuelles. Ces corpus pallient les deux faiblesses des corpus parallèles. Ils sont plus largement disponibles et les textes qui les composent ont généralement été écrits indépendamment dans chaque langue. Les hypothèses posées pour les corpus parallèles ne sont plus valables avec les corpus comparables, puisque les documents ne sont pas des traductions. Il n'est donc pas possible de s'appuyer sur la structure des documents pour mener à bien l'extraction lexicale.

1. « *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.* » – p. 173.

La méthode d'extraction dite *approche directe* (Fung, 1998), traditionnellement associée à l'extraction de lexiques bilingues à partir de corpus comparables, donne de bons résultats pour des corpus volumineux (centaines de millions de mots). Ces résultats chutent significativement en utilisant des petits corpus (centaines de milliers de mots). Nous proposons dans cet article une nouvelle hypothèse pour améliorer les résultats de l'approche directe à partir de petits corpus comparables spécialisés en s'appuyant sur un vocabulaire commun et identifiable auquel nous conférons un statut particulier dans le processus d'extraction.

Après avoir présenté l'approche directe et précisé les enjeux de la caractérisation des contextes des termes en section 2, nous définissons la notion de point d'ancrage sous-jacente à ce travail et montrons son intégration avec l'approche directe en section 3. Cette notion est illustrée à travers deux exemples de vocabulaire spécialisé : les translittérations et les composés savants. Nous présentons les différentes expériences réalisées visant à préciser l'influence de ces points d'ancrage dans le processus d'alignement en section 4. Enfin, la section 5 dresse le bilan de ce travail.

2. Extraction de lexiques bilingues à partir de corpus comparables

2.1. *Caractérisation des contextes des termes*

Les premières recherches en extraction lexicale à partir de corpus comparables se sont naturellement éloignées des approches proposées pour les corpus parallèles. Elles ont proposé d'identifier des caractéristiques qui seront proches entre un terme i et sa traduction $t(i)$ mais éloignées entre un terme i et d'autres mots qui n'en sont pas des traductions. (Fung, 1995) propose d'abord de s'intéresser à la *productivité* des termes, c'est-à-dire au nombre de voisins directs rencontrés dans le corpus ; l'hypothèse posée étant que ces productivités (productivités à gauche et à droite) sont proches entre un terme i et sa traduction $t(i)$, mais très différentes pour des termes non-traduction. À la même période, (Rapp, 1995) propose une approche comparant les motifs de relations d'association entre un terme i et l'ensemble de ses voisins (l'association entre deux termes indique dans quelle mesure ils apparaissent plus souvent que *par chance*, voir section 2.2.2). Il propose l'hypothèse qu'un mot i et sa traduction $t(i)$ devaient avoir des motifs d'association comparables.

Ces deux approches réalisent le bond conceptuel vers l'extraction à partir de corpus comparables : elles proposent de s'appuyer non plus sur un terme i à traduire, mais sur son *contexte*, c'est-à-dire sur l'ensemble des mots avec lesquels i cooccurre. Elles font ainsi écho à la proposition de (Firth, 1957) : « *on reconnaît un mot à ses fréquentations* » [2]. Le processus d'extraction à partir de corpus comparables repose en grande partie sur la capacité à caractériser les contextes des termes à aligner. L'approche directe propose de s'appuyer sur des ressources linguistiques existantes pour

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

comparer les contextes des termes. Elle propose aussi une structure de données pour enregistrer ces contextes.

2.2. *Approche directe*

L'approche directe a été introduite par (Fung, 1998), s'inspirant des modèles de contextes pour les corpus monolingues ainsi que de méthodes utilisées en *recherche d'information*. Les concepts de *requêtes* et de *documents* devenant respectivement ceux de *terme* et de *contexte de terme*. Le principe de cette approche est synthétisé dans le schéma de la figure 1.

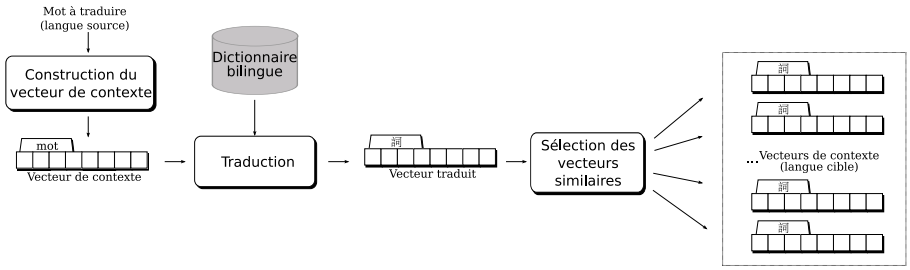


Figure 1. *Principe de l'approche directe*

L'approche directe repose sur la caractérisation et la comparaison des contextes des termes à aligner en utilisant une structure de données particulière, les *vecteurs de contexte*. Ces vecteurs stockent 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 contexte source s'obtient à l'aide d'un dictionnaire bilingue. L'algorithme de cette approche 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. Si le dictionnaire propose plusieurs traductions pour un élément, nous ajoutons au vecteur de contexte de i l'ensemble des traductions proposées, pondérées par leur fréquence dans le corpus cible. Les éléments pour lesquels aucune traduction n'est disponible ne sont pas transférés en langue cible ;

– **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 d'*extraction* ou d'*alignement* lexical.

2.2.1. *Construction des vecteurs de contexte*

La qualité des vecteurs de contexte conditionne grandement la qualité des résultats de l'alignement, en accord avec la remarque de la section 2.1 sur l'importance d'une caractérisation efficace des contextes. 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 mots avant, n mots après l'unité considérée) ou variable (phrase, paragraphe, etc.). Par exemple, (Déjean et Gaussier, 2002) considèrent tous les mots dans les phrases qui précèdent et suivent l'unité étudiée, ainsi que dans la phrase la contenant.

Un deuxième paramètre à calibrer soigneusement est le type des mots que nous souhaitons voir apparaître dans les vecteurs. Il est par exemple peu utile 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 prétraiter les corpus pour en extraire les unités pleines, de les lemmatiser pour regrouper les formes fléchies, de les étiqueter pour ne conserver que les unités pertinentes (substantifs, adverbes, etc.) et de filtrer les mots fonctionnels (articles, auxiliaires, conjonctions, etc.).

D'autres indices peuvent être utilisés, par exemple, la fréquence minimale d'un terme dans le corpus pour qu'il soit pris en compte 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 les plus significatifs.

2.2.2. *Mesures d'association*

Une mesure d'association évalue le degré de dépendance statistique de deux variables aléatoires. Si deux variables sont corrélées, 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 automatique des langues naturelles, ces mesures peuvent être utilisées 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 des choses[3]. Un mot d'une collocation apparaîtra en effet plus souvent avec les autres mots de la collocation qu'avec d'autres

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

mots. Les mesures d'association peuvent notamment se calculer à partir d'une table de contingence (*cf.* tableau 1), indiquant pour chaque couple de termes le nombre de leurs cooccurrences et le nombre de cooccurrences de l'un sans l'autre. Nous présentons la formule du *Taux de vraisemblance* (Dunning, 1993, équation 1) ; $occ(i, j)$ est le nombre de cooccurrences des éléments i et j dans une fenêtre donnée ; $\neg i$ représente toutes les unités considérées sauf i.

	j	$\neg j$
i	$a = occ(i, j)$	$b = occ(i, \neg j)$
$\neg i$	$c = occ(\neg i, j)$	$d = occ(\neg i, \neg j)$

Tableau 1. *Table de contingence pour un couple i et j*

$$\begin{aligned}\lambda(i,j) \quad = \quad & a\log(a) + b\log(b) + c\log(c) + d\log(d) + \\ & (a+b+c+d)\log(a+b+c+d) - \\ & (a+b)\log(a+b) - (a+c)\log(a+c) - \\ & (b+d)\log(b+d) - (c+d)\log(c+d)\end{aligned} \qquad [1]$$

Dans le cadre de nos recherches, nous utilisons les 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. La comparaison des vecteurs se fait donc sur les associations de leurs éléments.

2.2.3. *Traduction des vecteurs*

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

L'utilisation de dictionnaires bilingues pose aussi des problèmes lorsqu'un mot possède plusieurs traductions, qu'il s'agisse de traductions synonymes ou d'un terme source polysémique. Dans ce cas, et comme il est difficile d'évaluer dans des ressources *plates* comme les dictionnaires quelles traductions sont les plus pertinentes (les différentes traductions sont le plus souvent non ordonnées), 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 deuxième. Cette démarche suppose que les entrées des ressources sont classées par ordre d'importance, ce qui n'est pas toujours le cas.

2.2.4. *Résultats de l'approche directe*

Il est difficile de comparer les résultats entre les différentes études publiées sur l'extraction lexicale bilingue à partir de corpus comparables, en raison des différences entre corpus utilisés (en particulier leurs contraintes de construction et leur volume – *cf.* section 1) mais aussi de la couverture et de la pertinence des ressources linguistiques utilisées pour la traduction. À ce jour, il n'existe à notre connaissance aucune expérience et aucun jeu de ressources pouvant servir de référence.

Les résultats de l'approche directe s'évaluent sur le nombre de candidats correctement alignés, trouvés dans les x premiers candidats proposés (le Top_x). (Rapp, 1999) obtient par exemple 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. Précisons que nous nous intéressons dans cet article uniquement à l'alignement de termes simples (composés d'un seul mot), mais que d'autres recherches se sont portées sur l'alignement de termes complexes, notamment (Morin et Daille, 2004).

2.3. *Travaux reliés*

Dans le cadre de l'approche directe, le transfert du vecteur de contexte d'un mot à traduire de la langue source à la langue cible repose sur la traduction de chacun des éléments de son vecteur de contexte au moyen d'un dictionnaire bilingue. En fonction de l'adéquation du dictionnaire bilingue avec le corpus d'étude, plus ou moins d'éléments du vecteur de contexte seront traduits. Si aucun élément du vecteur de contexte ne peut être traduit, le vecteur transféré sera vide et il ne sera pas trouvé de traduction pour le mot visé. Dans ce type d'approche – où il y a toujours des éléments du vecteur de contexte qui ne peuvent être traduits – le mot à traduire perd naturellement de son potentiel de discrimination dans la langue cible. Pour limiter cet effet, l'adjonction de ressources supplémentaires permet d'améliorer significativement la qualité des lexiques extraits. Ainsi en associant un dictionnaire de langue générale à un dictionnaire spécialisé, (Chiao et Zweigenbaum, 2003) obtiennent une amélioration significative des performances d'alignement en faisant passer la précision de 61 à 94 % pour le Top_{20}. Dans le même esprit, (Déjean et Gaussier, 2002) s'appuient sur les propriétés hiérarchiques d'un thésaurus spécialisé pour améliorer le rang des tra-

ductions candidates. Avec cette ressource supplémentaire, ils font passer la précision de 57 à 63 % pour le Top_{20}.

Une autre technique visant à améliorer les résultats de l'approche directe consiste à filtrer les traductions candidates en s'appuyant sur une hypothèse de « symétrie distributionnelle » (Sadat *et al.*, 2003 ; Chiao *et al.*, 2004). Cette hypothèse est exprimée par (Chiao, 2004, p. 53) de la manière suivante : « *[...] si deux mots sont proches dans une direction de traduction ainsi que dans l'autre alors ils ont de plus fortes chances d'être traductions l'un de l'autre que s'ils ne sont proches que pour une seule direction de traduction.* »La mise en œuvre de cette hypothèse permet à (Chiao *et al.*, 2004) d'améliorer la précision de leurs résultats de 10 % pour le Top_{10}. (Sadat *et al.*, 2003) utilisent une technique similaire pour filtrer leurs résultats et y ajoutent une condition sur la catégorie syntaxique attendue de la traduction. Ce double filtrage permet un gain d'environ 12 %, toujours en précision, pour la recherche d'information interlangue.

D'autres méthodes ont été proposées pour pallier l'insuffisance des ressources bilingues utilisées lors de la traduction des vecteurs de contexte. (Déjean et Gaussier, 2002) proposent ainsi une approche dite par *similarité interlangue* qui repose sur une meilleure exploitation des ressources bilingues. Cette approche s'appuie sur l'hypothèse suivante : « *deux mots de l_1 et l_2 sont, avec une forte probabilité, traduction l'un de l'autre si leurs similarités avec les entrées des ressources bilingues disponibles sont proches* » (Déjean et Gaussier, 2002, p. 7). Le principe de cette approche consiste à identifier les vecteurs de contexte du dictionnaire bilingue (construit sur le corpus comparable, à partir de la liste des mots 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 ne perdent pas 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. Avec cette méthode, (Déjean et Gaussier, 2002) obtiennent pour des termes simples français/allemand une précision pour les 10 et 20 meilleurs candidats 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). Une approche similaire a été utilisée par (Morin et Daille, 2004) pour l'alignement de termes complexes.

En complément à cette approche, (Gaussier *et al.*, 2004) ont aussi étudié l'apport de l'analyse sémantique latente probabiliste (PLSA, *Probabilistic Latent Semantic Analysis*, (Hofmann, 1999)) à la problématique de l'alignement lexical bilingue. Les résultats obtenus avec la PLSA restent inférieurs à ceux obtenus avec l'approche directe mais proches de ceux obtenus avec l'approche par similarité interlangue.

(Pekar *et al.*, 2006) ont aussi apporté une contribution à l'approche directe en s'intéressant à l'alignement des mots de faibles fréquences. Dans cette approche, lorsque la probabilité d'un mot n apparaissant dans le contexte d'un mot k à traduire n'est

pas fiable ou ne peut être estimée, (Pekar *et al.*, 2006) proposent de calculer cette probabilité en s'appuyant sur la moyenne des probabilités des plus proches voisins de n apparaissant dans le contexte du mot k. Les différentes expériences associées à ce travail indiquent globalement une amélioration des performances, le rang moyen des traductions correctes étant amélioré de 2 points pour le Top_{10} et de 11 points pour le Top_{100} avec un corpus français/anglais.

Comme il est rappelé dans l'article de synthèse sur les corpus comparables de (Zweigenbaum et Habert, 2006, p. 36) : « *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).

3. Utilisation de points d'ancrage

De manière à améliorer le pouvoir de caractérisation des vecteurs de contexte dans le cas de petits corpus, nous proposons de nous appuyer sur des éléments de confiance, c'est-à-dire des points d'ancrage utilisés dans le processus d'alignement.

3.1. *Contexte*

Dans le cadre de cette étude, nous avons constitué un corpus trilingue français, anglais et japonais[4]. Les documents qui composent ce corpus ont été extraits du Web. Ils concernent le thème *alimentation et diabète* et appartiennent au registre *scientifique*[5]. Les documents ont été sélectionnés manuellement, à partir de requêtes sur des moteurs de recherche, mais aussi en suivant les liens internes des pages proposées. Ces documents ont été convertis de leur format source HTML ou PDF en texte brut. Nous avons ainsi collecté 257 000 mots pour la partie française, 235 000 pour la partie japonaise et 250 000 mots pour la partie anglaise.

4. Une description précise de la constitution du corpus français/japonais est donnée dans (Morin et Daille, 2006).

5. Documents écrits par des experts à destination d'autres experts (Pearson, 1998, p. 36).

Le dictionnaire français/japonais, nécessaire pour l'étape de traduction de l'approche directe, est composé de quatre dictionnaires disponibles librement sur Internet[6] ainsi que du *Dictionnaire scientifique français-japonais* (1989). Il contient 173 156 entrées, dont 114 461 sont des termes simples, avec une moyenne de 2,1 traductions par entrée. Nous avons utilisé le dictionnaire *JMDict* pour l'anglais/japonais[7] qui est disponible librement sous une licence *Creative-Commons* (Attribution-ShareAlike). Nous l'avons complété de traductions de termes techniques issus de différentes sources : dictionnaire du Ministère de l'Éducation japonais et du *National Institute of Informatics* (Tokyo)[8] ainsi que du *Dictionary of Technical Term* (Kotani et Kori, 1990). Sa version complétée contient 589 946 entrées avec une moyenne de 2,3 traductions par entrée et seulement 49 208 termes simples.

3.2. *Vocabulaire spécialisé comme point d'ancrage*

La construction des vecteurs de contexte dans le cadre des petits corpus est délicate : dans notre cas ils sont constitués de peu de documents. Même si les documents du corpus sont censés partager des traits communs, la prédominance d'un thème dans l'un d'eux peut faire varier grandement la fréquence d'un terme et donc son association dans les vecteurs de contexte. À l'inverse, les mesures d'association sont lissées dans le cas de grands corpus comparables, favorisant la significativité des vecteurs de contexte. Nous cherchons à renforcer cette significativité pour les vecteurs de contexte construits sur des petits corpus, en recherchant des *points d'ancrage*. Ces points d'ancrage doivent être des *éléments de confiance*, c'est-à-dire des éléments dont l'absence ou la présence dans un vecteur de contexte est particulièrement discriminante pour caractériser un terme. Notons que cette notion est proche de celle introduite initialement pour les corpus parallèles (Véronis, 2000), 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. Ces points d'ancrage sont des paires de traductions que nous cherchons à retrouver dans les vecteurs sources et cibles lors du calcul de similarité entre vecteurs.

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

1) ils doivent être faciles à identifier ;

2) ils doivent être pertinents, relativement aux thèmes des documents à aligner ;

3) ils doivent être peu polysémiques, pour ne pas être ambigus.

Nous émettons l'hypothèse que ces points d'ancrage sont des éléments discriminants dans la caractérisation des contextes des termes. La première propriété nous

6. `kanji.free.fr` ; `quebec-japon.com/lexique/index.php?a=index&d=25` ; `dico.fj.free.fr/index.php` ; `quebec-japon.com/lexique/index.php?a=index&d=3`
7. `www.csse.monash.edu.au/~jwb/j_jmdict.html`
8. `sciterm.nii.ac.jp/cgi-bin/reference.cgi`

permet de les utiliser dans le cadre d'un processus automatique. Les propriétés 2 et 3 assurent que ces points d'ancrage sont significatifs, c'est-à-dire aptes à caractériser efficacement les termes spécialisés que nous cherchons à aligner, sans introduire de nouvelles ambiguïtés. À partir des propriétés du corpus comparable à notre disposition, nous avons observé et extrait deux types de points d'ancrage pour appuyer notre hypothèse : les translittérations et les composés savants.

Nous appelons *translittération* le phénomène d'emprunt d'un mot d'une langue source à destination d'une langue cible qui ne partage pas nécessairement les mêmes phonèmes ni les mêmes symboles d'écriture. Le mot emprunté est adapté graphiquement dans la langue cible, sur la base de sa prononciation et non de son sens (Knight et Graehl, 1997). Dans (Prochasson *et al.*, 2008), nous avons montré la proéminence des translittérations dans le corpus spécialisé japonais que nous utilisons. Elles sont faciles à identifier, car écrites à l'aide d'un syllabaire principalement dédié aux mots d'emprunt en japonais (les *katakanas*) ; elles sont également représentatives d'un vocabulaire spécifique qui recouvre le vocabulaire spécialisé (Ito, 2007). Elles sont fréquemment employées dans le vocabulaire scientifique alors même qu'il existe la possibilité de construire un terme plus classique et équivalent en japonais. Les translittérations japonaises sont issues pour la plupart de l'anglais, mais peuvent dans de nombreux cas être alignées avec des termes français, en raison des relations de *cognats* fréquentes entre le français et l'anglais. Par exemple, le terme japonais インスリン / `i-n-su-ri-n` peut s'aligner en anglais avec *insulin* et en français avec *insuline*.

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 relèvent d'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.

3.3. *Modification de l'approche directe*

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 fonde en priorité sur les points d'ancrage, puis sur les éléments moins significatifs. Après avoir calculé l'association de façon classique, 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 2 à 4). Dans ces équations, PA est l'ensemble des points d'ancrage extraits ($|PA|_l$ le nombre de points d'ancrage trouvés 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.

$$assoc\_pondérée^l_j = assoc^l_j + \beta,\ si\ j\ \in\ PA \quad [2]$$

$$assoc\_pondérée^l_j = assoc^l_j - décalage_l,\ si\ j\ \notin\ PA \quad [3]$$

$$décalage_l = \frac{|PA|_l}{|\neg PA|_l} \times \beta \quad [4]$$

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 leur 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 plus faire 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, en fonction du type des points d'ancrage utilisés, de la confiance qui doit leur être accordée, mais aussi en fonction des mesures d'association utilisées, puisqu'elles ne retournent pas toutes les mêmes intervalles de valeurs.

3.4. *Détection des points d'ancrage*

De nombreux efforts ont été réalisés pour aligner automatiquement les translittérations. Dans ce travail, nous avons utilisé un outil réalisant la détection automatique de translittérations entre l'anglais et le japonais (Tsuji, 2001). Il génère un ensemble de correspondances potentielles pour une entrée donnée en katakana ou en anglais. Les résultats doivent alors être comparés avec un vocabulaire existant pour sélectionner les candidats les plus probables. Nous obtenons avec cet outil 589 paires de translittérations pour le couple anglais/japonais. En ce qui concerne la détection des translittérations entre le français et le japonais, nous avons utilisé le même outil, n'ayant pu disposer d'un outil efficace dédié. Avant traitement, les termes français à comparer sont normalisés pour faire disparaître les signes diacritiques spécifiques (mais ils sont réintégrés dans les couples alignés). Nous obtenons 526 paires de translittérations pour le couple français/japonais.

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[9]. Le processus d'ex-

9. www.medo.jp/a.htm

traction est trivial : en compilant une expression régulière par préfixe et par suffixe, il cherche les mots anglais correspondant 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*) est par exemple transformée en `-ie` en français (*psychologie*). Toutefois, 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*). Ainsi, 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 : 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 fournissent 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 résultent principalement de la nature des dictionnaires bilingues utilisés dans chaque paire de langues.

Mettre en évidence un point d'ancrage uniquement dans la langue cible (respectivement source), en omettant sa traduction dans la langue source (resp. cible), ne contribuera qu'à déséquilibrer les scores d'association du vecteur cible (resp. source) et défavorisera l'alignement de ces vecteurs (les points d'ancrage ne seront pas transférés d'une langue à l'autre). La nature même des translittérations permet de les extraire directement du corpus japonais et de trouver leurs équivalents de traduction dans les corpus français et anglais. À l'inverse, les composés savants ne permettent pas de retrouver automatiquement leurs traductions. C'est la raison pour laquelle les translittérations sont extraites à partir du corpus comparable, alors que les composés savants sont extraits à partir des ressources bilingues. De plus, les paires de translittérations détectées sont ajoutées aux ressources bilingues, de manière à ce qu'elles puissent franchir l'étape de traduction et être des points d'ancrage convenables.

3.5. *Couverture des ressources linguistiques*

Le tableau 2 présente la couverture des ressources linguistiques sur les vecteurs de contexte des langues sources (anglais et français), c'est-à-dire le nombre de mots présents dans les vecteurs de contexte, dont au moins une traduction est disponible dans les ressources linguistiques, par rapport au nombre de mots total. Il présente également l'influence des points d'ancrage sur cette couverture. La colonne $mots_V$ indique le nombre de mots différents disponibles dans l'ensemble des vecteurs (c'est-à-dire l'ensemble du vocabulaire sur lequel nous cherchons à mesurer la couverture). La colonne *originale* indique la couverture dans le cas des ressources linguistiques

d'origine, non modifiées. La colonne *avec trans.* indique la couverture lorsque sont ajoutées les translittérations détectées automatiquement aux ressources, alors que la colonne *sans trans.* (respectivement *sans CS*) indique la couverture des ressources après retrait des translittérations détectées des ressources originales (resp. après retrait des composés savants). Il n'y a pas de colonne *avec CS* puisque les composés savants ont été extraits des ressources : elle est équivalente à la colonne *originale*. Le calcul est présenté en équation 5. Il correspond à l'intersection de l'ensemble des mots des vecteurs sources $mots_V$ et de la partie source des ressources bilingues res_{source}, divisée par le cardinal de $mots_V$.

Langue	$mots_V$	*originale*	*avec trans.*	*sans trans.*	*sans CS*
Français	8 709	57,1 %	57,2 %	57,1 %	47,7 %
Anglais	3 884	65,8 %	67,0 %	65,8 %	47,6 %

Tableau 2. *Couverture des ressources linguistiques originales et avec et sans translittérations et mots savants détectés, sur les vecteurs de contexte anglais et français*

$$C(res_{source}, mots_V) = \frac{|res_{source} \cap mots_V|}{|mots_V|} \quad [5]$$

Ces résultats permettent d'évaluer la qualité de la couverture dans le meilleur cas. En effet, les chiffres présentés indiquent combien de mots, présents dans les vecteurs de contexte sources, ont une ou plusieurs traductions en japonais données par les ressources linguistiques : la qualité de ces traductions n'est pas évaluée. En raison de phénomènes de polysémie, certaines de ces traductions ne sont peut-être pas pertinentes. Les résultats diffèrent entre le français et l'anglais, ce qui n'est pas surprenant au vu des différences entre les ressources utilisées pour chacune de ces langues. De plus, dans de nombreux cas, à un seul mot source sont associées plusieurs traductions dans les ressources, ils ne sont comptés qu'une fois.

Les translittérations ont un impact faible dans le cas du français : les ajouter ou les retirer modifie peu ou pas la couverture des ressources (car une traduction différente, n'impliquant pas les translittérations japonaises détectées pour le mot source existe déjà dans les ressources). Les translittérations détectées pour l'anglais semblent mieux couvrir le vocabulaire des vecteurs de contexte (1,2 point d'augmentation), mais les retirer ne modifie pas la couverture. Par ailleurs, les composés savants représentent une partie importante du vocabulaire de chaque langue : les retirer pénalise de plus de 18 points la couverture des ressources dans le cas de l'anglais, et de plus de 9 points dans le cas du français.

4. Expériences et résultats

Nous avons implémenté les modifications proposées en section 3 dans notre chaîne de traitements. Nous comparons les résultats obtenus avec les mesures d'association

pondérées, utilisant les points d'ancrage, avec les résultats obtenus dans le cadre de l'approche directe standard.

4.1. *Protocole*

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

– (*a*) approche directe standard ;

– (*b*) en utilisant les translittérations détectées automatiquement ;

– (*c*) en utilisant les composés savants extraits automatiquement.

L'expérience *a* est une expérience témoin, utilisée comme étalon pour être comparée avec les expériences *b* et *c*.

4.1.1. *Paramètres*

Les expériences sont réalisées avec les mêmes paramètres :

– le *Taux de vraisemblance* (eq. 1) comme mesure d'association ;

– le *Cosinus* comme mesure de similarité (eq. 6) ;

$$cos(v_l, v_k) = \frac{\sum_j assoc_j^l \times assoc_j^k}{\sqrt{\sum_j {assoc_j^l}^2} \times \sqrt{\sum_j {assoc_j^k}^2}} \quad [6]$$

– une taille de fenêtre de 25 mots avant et 25 mots après la tête du vecteur pour la construction des vecteurs de contexte (les vecteurs sont constitués après retrait des mots fonctionnels du corpus) ;

– une fréquence minimale de 3 occurrences pour qu'un mot soit pris en compte dans la constitution des vecteurs de contexte.

Ces paramètres sont ceux qui donnent les meilleurs résultats pour l'expérience témoin (a), que nous comparons avec les expériences *b* et *c*.

La mesure de l'équation 6 correspond au cosinus de l'angle formé par les deux vecteurs v_l et v_k, obtenus après calcul des mesures d'association[10]. Dans le cas des expériences *b* et *c*, nous faisons varier le paramètre β entre 1 et 20 (par pas de 1).

4.1.2. *Liste d'évaluation*

Pour évaluer la qualité de l'extraction, nous avons construit une liste de traductions connues. Nous avons sélectionné les mots français et anglais les plus fréquents dans le corpus ($N_{occ} > 50$) dont la traduction en japonais est connue. Parmi ces traductions, nous avons sélectionné celles apparaissant fréquemment dans le corpus japonais

10. Nous renvoyons le lecteur à l'article de (Gaussier *et al.*, 2004) qui explique le principe de l'alignement d'un point de vue géométrique.

($N_{occ} > 50$) pour construire une liste de 98 traductions français/japonais et 99 traductions anglais/japonais. Ce protocole pour constituer une liste de termes utilisée pour l'évaluation est semblable à celui présenté dans (Chiao et Zweigenbaum, 2002). Ils travaillent avec une liste de 95 termes pour un corpus médical anglais/français d'environ 600 000 mots pour chaque langue.

4.1.3. *Vecteurs de contexte*

Les vecteurs de contexte français et anglais sont construits à partir des corpus étiquetés et lemmatisés en utilisant l'étiqueteur de *Brill* (Brill, 1992) et le lemmatiseur *FLEMM* (Namer, 2000, pour la partie française). Les vecteurs de contexte japonais sont construits à partir du corpus segmenté par l'analyseur morphosyntaxique *Chasen* (Matsumoto *et al.*, 1999)[11]. Ils sont constitués après retrait des mots fonctionnels du corpus. Ils contiennent uniquement des termes simples. Nous obtenons 3 884 vecteurs de contexte anglais, 8 709 vecteurs français et 4 559 vecteurs japonais. Ces nombres correspondent au nombre de mots différents enregistrés dans les vecteurs.

4.2. *Résultats*

Le tableau 3 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), avec $\beta = 8$.

	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 %]

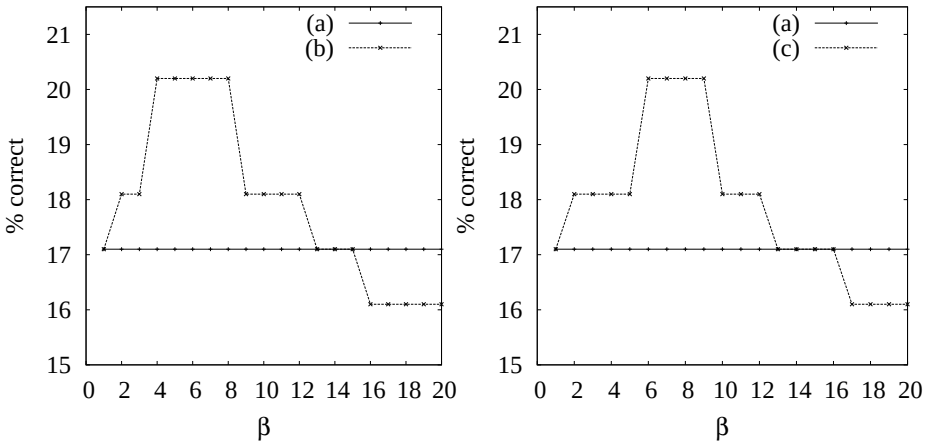
Tableau 3. *Résultats de l'alignement anglais/japonais et français/japonais ($\beta = 8$) ;* a *: expérience témoin ;* b *: utilisation des translittérations ;* c *: utilisation des composés savants*

Les résultats pour l'expérience de contrôle (exp. *a*) sont comparables aux résultats obtenus par (Chiao et Zweigenbaum, 2002) discutés en section 2.2.4. Dans le cas de l'anglais, le gain obtenu en s'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 au traitement anglais/japonais, sans oublier que les translittérations entre le français et le japonais sont plus rares.

11. `http://chasen.aist-nara.ac.jp/`

4.3. *Discussion*

Les résultats que nous avons présentés dans le tableau 3 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 2.1 et 2.2 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, ne dépendant pas du paramètre β).



2.1 : expérience *b* (translittérations) 2.2 : expérience *c* (composés savants)

Figure 2. *Influence du paramètre β, comparé à l'expérience témoin. Alignement anglais/japonais*

Ces figures montrent que dans le cas des points d'ancrage sélectionnés (figures 2.1 et 2.2), 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 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.

4.4. *Influence des points d'ancrage*

La figure 3 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 3 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).

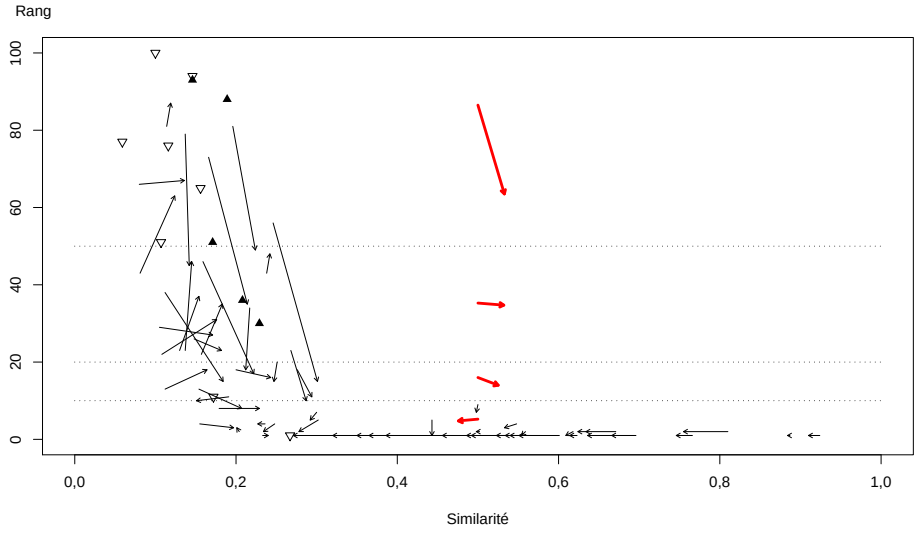


Figure 3. *Rangs et scores des traductions correctes pour l'alignement français/japonais, avec et sans utilisation des points d'ancrage (composés savants – $\beta = 8$)*

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 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 représentent 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 semblent correspondre à des traductions instables, très sensibles aux différents paramètres utilisés (taille de la fenêtre, mesure d'association et de similarité...). Les flèches permettent de mieux comprendre l'influence des points d'ancrage. Elles indiquent, 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 : il existe une tendance au réarrangement des candidats à la traduction vers des positions plus avantageuses, quel que soit leur rang initial, malgré une amélioration des $Top\ 1$ et 10 peu importante.

Afin d'évaluer la significativité de cette amélioration, nous avons effectué un *t-test apparié* (comparaison des paires de rangs des traductions correctes sans/avec points d'ancrage (Harris, 1998)). Nous proposons comme hypothèse nulle que l'utilisation des points d'ancrage ne permet pas d'amélioration des rangs des traductions correctes. Les résultats du test ($t = 1,8694$; $p = 0,0333$) nous permettent de rejeter l'hypothèse nulle avec un intervalle de confiance à 95 %[12]. Ces tests statistiques nous permettent également de rejeter l'hypothèse nulle dans le cas de l'alignement anglais/japonais (en utilisant les translittérations et les composés savants), mais pas dans le cas de l'alignement français/japonais en utilisant les translittérations, probablement en raison de la mauvaise qualité de la détection des translittérations entre le français et le japonais.

5. Conclusion

Nous avons proposé dans cet article d'introduire la notion de point d'ancrage dans le processus d'alignement lexical à partir de corpus comparables. Notre hypothèse repose sur l'identification et la confiance en un vocabulaire spécialisé pour améliorer les résultats de l'approche directe. Cette hypothèse a été confirmée par l'expérience : nous avons montré que les résultats sont améliorés pour les Top_1 et Top_{10}, mais aussi que le classement des candidats à la traduction est globalement amélioré en utilisant des points d'ancrage.

Cette étude ouvre la voie à de nouvelles perspectives pour améliorer les méthodes d'alignement lexical à partir de corpus comparables spécialisés de taille réduite. D'un côté, la qualité des points d'ancrage que nous avons extraits peut être améliorée en utilisant des outils plus performants. De l'autre, il est probable qu'il existe d'autres points d'ancrage pertinents, dont l'extraction est permise par des travaux transversaux en traitement automatique des langues naturelles (typiquement, l'extraction des cognats (Kraif, 1999)). Ces points d'ancrage doivent être choisis avec soin pour chaque couple de langue envisagé.

Enfin, cette étude invite à repenser la façon dont sont caractérisés et comparés les contextes des termes. En effet, l'approche directe s'appuie sur la comparaison des as-

12. Le *test de Wilcoxon* retourne une *p-value* de $0,032$.

sociations entre les têtes et les éléments des vecteurs de contexte, d'une langue vers une autre. Nous l'avons modifiée et avons obtenu des résultats prometteurs. En parallèle à la recherche et à l'exploitation de nouveaux points d'ancrage, nous souhaitons réfléchir à de nouvelles mesures discriminantes, plus adaptées que ne le sont les mesures d'association pour l'extraction à partir de corpus comparables spécialisés de taille réduite.

Remerciements

Les auteurs tiennent à remercier Kyo Kageura (Université de Tokyo, Japon) et Akiko Aizawa (*National Institute of Informatics*, Tokyo, Japon) pour leur contribution à l'étude des translittérations japonaises et leurs conseils avisés. Nous remercions également Chantal Enguehard et Annie Tartier pour leurs relectures attentives et précises. Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-08-CORD-009.

6. Bibliographie

Brill E., « A Simple Rule-Based Part of Speech Tagger », *Proceeding of the 3rd Conference on Applied Natural Language Processing (ANLP'92)*, Trento, Italie, p. 152-155, 1992.

Chiao Y.-C., 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, 2004.

Chiao Y.-C., Sta J.-D., Zweigenbaum P., « A Novel Approach to Improve Word Translations Extraction from Non-Parallel, Comparable Corpora », *Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP'04)*, Hainan, Chine, 2004.

Chiao Y.-C., Zweigenbaum P., « Looking for candidate translational equivalents in specialized, comparable corpora », *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, Tapei, Taiwan, p. 1208-1212, 2002.

Chiao Y.-C., Zweigenbaum P., « The Effect of a General Lexicon in Corpus-Based Identification of French-English Medical Word Translations », *in* R. Baud, M. Fieschi, P. Le Beux, P. Ruch (eds), *The New Navigators : from Professionals to Patients, Actes Medical Informatics Europe*, vol. 95 of *Studies in Health Technology and Informatics*, IOS Press, Amsterdam, p. 397-402, 2003.

Claveau V., « Inférence de règles de réécriture pour la traduction de termes biomédicaux », *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN'07)*, Toulouse, France, p. 111-120, 2007.

Dunning T., « Accurate Methods for the Statistics of Surprise and Coincidence », *Computational Linguistics*, vol. 19, n° 1, p. 61-74, 1993.

Déjean H., Gaussier E., « Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables », *in* J. Véronis (ed.), *Lexicometrica, Alignement lexical dans les corpus multilingues*, p. 1-22, 2002.

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

Fung P., « Compiling Bilingual Lexicon Entries from a Non-Parallel English-Chinese Corpus », *in* D. Yarovsky, K. Church (eds), *Proceedings of the 3rd Workshop on Very Large Corpora (VLC'95)*, Somerset, NJ, États-Unis d'Amérique, p. 173-183, 1995.

Fung P., « A Statistical View on Bilingual Lexicon Extraction : From Parallel Corpora to Non-parallel Corpora. », *in* D. Farwell, L. Gerber, E. H. Hovy (eds), *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA'98)*, Langhorne, Pennsylvanie, États-Unis d'Amérique, p. 1-17, 1998.

Gaussier E., Renders J.-M., Matveeva I., Goutte C., Déjean H., « A Geometric View on Bilingual Lexicon Extraction from Comparable Corpora », *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04)*, Barcelone, Espagne, p. 526-533, July, 2004.

Harris M. B., *Basic Statistics for Behavioral Science Research*, 2nd edn, Allyn & Bacon, 1998.

Hofmann T., « Probabilistic Latent Semantic Analysis. », *in* K. Laskey, H. Prade (eds), *In Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI'99)*, Stockholm, Suède, p. 289-296, 1999.

Ito M., « Loan words and foreign words in 90 magazines and 70 magazines », *The 51st Annual Meeting of the Mathematical Linguistic Society of Japan*, 2007.

Knight K., Graehl J., « Machine Transliteration », *in* P. R. Cohen, W. Wahlster (eds), *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)*, Madrid, Espagne, p. 128-135, 1997.

Kotani T., Kori A., *Dictionary of Technical Terms*, Kenkyusha, 1990.

Kraif O., « Identification des cognats et alignement bi-textuel : une étude empirique », *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN'99)*, Cargese, France, p. 205-214, 1999.

Lovis C., Baud R., Michel P. A., Scherrer J. R., Rassinoux A. M., « Building Medical Dictionaries for Patient Encoding Systems : A methodology », *Lecture Notes in Computer Science*, vol. 1211, p. 373-380, 1997.

Manning C. D., Schütze H., *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, MA, États-Unis d'Amérique, 1999.

Matsumoto Y., Kitauchi A., T. T. Y., Hirano Y., *Japanese Morphological Analysis System ChaSen 2.0 User Manual*. 1999.

Morin E., Daille B., « Extraction terminologique bilingue à partir de corpus comparables d'un domaine spécialisé », *Traitement Automatique des Langues (TAL)*, vol. 45, n° 3, p. 103-122, 2004.

Morin E., Daille B., « Comparabilité de corpus et fouille terminologique multilingue », *Traitement Automatique des Langues (TAL)*, vol. 47, n° 1, p. 113-136, 2006.

Namer F., « FLEMM : Un analyseur flexionnel du français à base de règles », *Traitement Automatique des Langues (TAL)*, vol. 41, n° 2, p. 523-547, 2000.

Namer F., « Morphosémantique pour l'appariement de termes dans le vocabulaire médical : approche multilingue », *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN'05)*, Dourdan, France, p. 63-72, 2005.

Namer F., Zweigenbaum P., « Acquiring meaning for French medical terminology : contribution of morphosemantics », *in* M. Fieschi, E. Coiera, Y.-C. J. Li (eds), *Studies in Health Technology and Informatics*, vol. 107, Amsterdam, Pays-Bas, p. 535-539, 2004.

Pearson J., *Terms in Context*, John Benjamins publishing company, 1998.

Pekar V., Mitkov R., Blagoev D., Mulloni A., « Finding translations for low-frequency words in comparable corpora », *Machine Translation*, vol. 20, n° 4, p. 247-266, 2006.

Prochasson E., Kageura K., Morin E., Aizawa A., « Looking for Transliterations in a trilingual English, French and Japanese Specialised Comparable Corpus », *Proceedings of the 1st Workshop on Building and Using Comparable Corpora, Language Resources and Evaluation Conference (LREC'08)*, Marrakech, Maroc, p. 83-86, 2008.

Rapp R., « Identifying word translations in non-parallel texts », *Proceedings of the 33rd annual meeting on Association for Computational Linguistics (ACL'95)*, Cambridge, MA, États-Unis d'Amérique, p. 320-322, 1995.

Rapp R., « Automatic Identification of Word Translations from Unrelated English and German Corpora », *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, College Park, MD, États-Unis d'Amérique, p. 519-526, 1999.

Sadat F., Yoshikawa M., Uemura S., « Learning Bilingual Translations from Comparable Corpora to Cross-Language Information Retrieval : Hybrid Statistics-based and Linguistics-based Approach », *Proceedings of the Sixth Internation Workshop on Information Retrieval with Asian Languages (IRAL'03)*, Sappro, Japon, p. 57-64, 2003.

Tsuji K., « Automatic Extraction of Translational Japanese-KATAKANA and English Word Pairs from Bilingual Corpora », *Proceedings of the 19th International Conference on Computer Processing of Oriental Langages (ICCPOL'01)*, Taichung, Taiwan, p. 245-250, 2001.

Véronis J. (ed.), *Parallel Text Processing*, Kluwer Academic Publishers, 2000.

Zweigenbaum P., Habert B., « Faire se rencontrer les parallèles : regards croisés sur l'acquisition lexicale monolingue et multilingue », *Revue de sociolinguistique en ligne GLOTTOPOL*, vol. 8, p. 22-44, 2006.