Comparabilité de corpus et fouille terminologique multilingue

Emmanuel Morin, Béatrice Daille

Comparabilité de corpus et fouille terminologique multilingue

Emmanuel Morin — **Béatrice Daille**

Université de Nantes, LINA - FRE CNRS 2729
2, rue de la Houssinière, BP 92208
F-44322 Nantes cedex 03
{emmanuel.morin,beatrice.daille}@univ-nantes.fr

RÉSUMÉ. *Les principaux travaux en fouille textuelle privilégient communément la taille du corpus sur sa qualité. Ainsi dans le cadre de l'alignement lexical à partir de corpus comparables, les meilleurs résultats sont obtenus pour des corpus de grande taille (plusieurs millions de mots). Pour les domaines de spécialité, et pour de nombreuses paires de langues, il n'est pas possible de disposer de corpus textuels aussi volumineux. Dans le cadre de ce travail, nous soutenons l'hypothèse que la qualité des données textuelles peut non seulement suppléer à leur quantité mais garantit aussi celle des ressources lexicales extraites. En particulier, nous montrons l'intérêt de prendre en compte le type du discours lors de la constitution du corpus comparable pour obtenir des listes terminologiques de qualité.*

ABSTRACT. *Current work in terminology mining from corpora tends to favour corpus size over corpus quality. Concerning multilingual lexical acquisition from comparable corpora, the best results are obtained on large corpora, i.e. several million words. But for many domains and language pairs, such huge corpora are simply not available. Our main hypothesis is that corpus quality is not only as important as corpus size, it is also what guarantees the quality of the acquired terminological resources. More precisely, we show how important it is to take discourse type into account when building the corpus.*

MOTS-CLÉS : *corpus comparable, type de discours, alignement lexical, termes complexes.*

KEYWORDS: *comparable corpora, type of discourse, lexical alignment, multi-word terms.*

TAL. Volume 47 – n° 1/2006, pages 113 à 136

1. Introduction

Benoît Habert (2000, p. 18) rappelle les deux positions concernant la création d'un corpus : « Gros, c'est beau »[1] ou « insécurité dans les grands ensembles ». La première, qui est la position communément adoptée en fouille textuelle, privilégie la quantité des données sur leur qualité. Les raisons qui pourraient être invoquées sont, d'une part, la nécessité de disposer de grandes masses de données pour mettre en œuvre les méthodes informatiques et, d'autre part, le manque de méthodes opérationnelles pour automatiser la construction d'un corpus *représentatif* d'un domaine, d'une activité ou encore d'une situation de communication et répondant donc à des critères langagiers précis.

Dans le cadre de l'alignement lexical à partir de corpus comparables, de bons résultats sont obtenus pour des corpus de grande taille – plusieurs millions de mots – lorsqu'il s'agit de proposer des traductions de mots simples : pour environ 80 % d'entre eux, les bonnes traductions apparaissent dans les vingt premiers candidats (Fung, 1998; Rapp, 1999; Chiao et Zweigenbaum, 2002). En utilisant le web, Cao et Li (2002) atteignent une précision de 91 % sur les trois premières traductions. En fouille terminologique multilingue, les documents relevant d'une thématique d'un domaine de spécialité ne sont pas disponibles en grande quantité. L'hypothèse de notre travail est que la qualité des données textuelles non seulement supplée à la quantité mais garantit celle des ressources lexicales extraites.

Un corpus comparable rassemble (Bowker et Pearson, 2002, p. 93) : « *des documents textuels dans des langues différentes qui ne sont pas des traductions les uns des autres*[2] ». Le terme *comparable* signifie que ces textes partagent des caractéristiques communes comme la période, le domaine, le thème, le support médiatique, l'auteur, le type de discours, etc. (Baayen, 1994). Cette comparabilité annoncée des corpus utilisés dans les travaux en fouille terminologique multilingue n'est guère précisée : elle se réduit souvent à un domaine générique comme le médical pour Chiao et Zweigenbaum (2002) ou la finance (Fung, 1998), ou à un support communicationnel comme les articles journalistiques (Fung, 1998). Pour l'acquisition terminologique, la prise en compte du type de discours, qui agit comme un sélectionneur sémantique sur les termes employés, nous paraît essentielle. Par exemple, en français, dans le domaine de la médecine, dans le sous-domaine des maladies liées à l'obésité, le terme *excès de poids* est utilisé uniquement dans les documents de discours vulgarisé, alors que son synonyne *excès pondéral* n'apparaîtra que dans le discours scientifique. Cette distinction de discours se retrouve par ailleurs sur des portails médicaux comme CISMeF[3] où les documents sont catégorisés selon qu'ils relèvent du discours scientifique ou vulgarisé. De manière à évaluer l'importance de la caractéristique du type de discours du corpus comparable pour la constitution de listes terminologiques bilingues, nous avons réalisé une série d'expérimentations sur deux corpus comparables français-

1. Il est fait ici référence à l'article de Marie-Paule Pery-Woodley (1995).
2. "*sets of texts in different languages, that are not translations of each other*".
3. Catalogue et Index des Sites Médicaux Francophones `http://www.cismef.org/`

japonais dans le domaine du médical dont les textes ont été moissonnés à partir du web et classés manuellement selon leur type de discours. Le premier corpus rassemble des documents uniquement de discours scientifique, le second corpus des documents quelconques, de discours scientifique mais aussi vulgarisé.

Nous utilisons une chaîne de fouille terminologique multilingue standard composée d'un extracteur de termes dans chaque langue et d'un module d'alignement. Les extracteurs de termes utilisés sont disponibles dans le domaine public et extraient des termes complexes qui sont plus précis que des termes simples. Le module d'alignement implémente les deux approches d'analyse contextuelle les plus connues : la « méthode directe » (Fung, 1998; Peters et Picchi, 1998; Rapp, 1999) et la « méthode par similarité interlangue » (Déjean et Gaussier, 2002; Morin et Daille, 2004). L'évaluation des traductions des termes complexes effectuée à l'aide d'une liste de référence montre que les résultats obtenus sont meilleurs lorsque le type de discours est une caractéristique du corpus comparable malgré une taille de corpus réduite de moitié. La quantité de données n'est donc pas suffisante pour obtenir des listes terminologiques de qualité et une véritable comparabilité du corpus impliquant notamment le domaine et le type de discours s'avère essentielle.

Dans la suite de cet article, nous présentons en section 2 la chaîne de fouille terminologique multilingue exploitée dans ce travail. La section 3 décrit et évalue les différentes ressources textuelles utilisées : les corpus comparables, le dictionnaire bilingue et les lexiques de référence. La section 4 présente les différentes expérimentations visant à vérifier l'hypothèse de cette étude. Enfin, la section 5 dresse le bilan de ce travail.

2. Chaîne de fouille terminologique multilingue

La chaîne de fouille terminologique multilingue permet à partir d'un corpus comparable en deux langues de proposer une liste de termes simples et complexes et leurs traductions candidates. L'architecture présentée en figure 1 est modulaire et est constituée d'un extracteur de termes dans chaque langue et d'un programme d'alignement lexical.

2.1. *Extraction terminologique*

L'extraction terminologique[4] est réalisée à l'aide de l'extracteur de terminologie *ACABIT* (Daille, 2003) disponible pour le français[5] et le japonais[6] (Takeuchi et al.,

4. Au préalable à l'extraction terminologique, chaque partie du corpus comparable est étiquetée, puis lemmatisée. Pour le français, il s'agit de Brill (Brill, 1994) associé à Flemm (Namer, 2000) et pour le japonais de ChaSen (Matsumoto et al., 1999).
5. `http://www.sciences.univ-nantes.fr/info/perso/permanents/daille/` et distribution Mandrake de LINUX.
6. `http://research.nii.ac.jp/~koichi/study/hotal/`

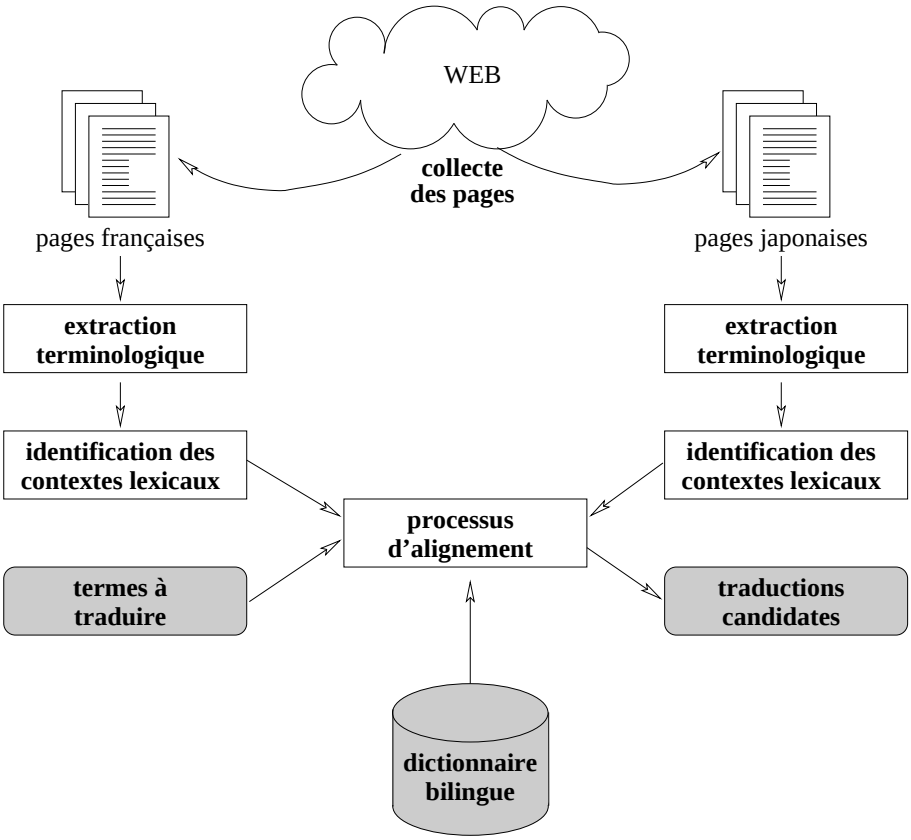


Figure 1. *Plateforme de fouille terminologique multilingue*

2004). Les unités lexicales extraites sont des termes complexes épousant des patrons morphosyntaxiques précis correspondant soit à la forme canonique du terme, soit à l'une de ses variations. Les patrons sont `N N`, `N Prep N` et `N Adj` pour le français et `N N`, `N Suff`, `Adj N` et `Pref N` pour le japonais. Les variantes traitées dans les deux langues relèvent de la morphologie et de la syntaxe. Nous détaillons ci-dessous un exemple d'un même terme extrait pour chaque langue.

En français, les occurrences suivantes du terme *sécrétion d'insuline* ont été relevées dans le corpus :

– **forme de base** : *sécrétion d'insuline* ;
– **variante flexionnelle** : *sécrétions d'insuline* ;
– **variante syntaxique (modification)** : *sécrétion pancréatique d'insuline* ;
– **variante syntaxique (coordination)** : *sécrétion de peptide et d'insuline*.

Les termes *sécrétion insulinique* et *hypersécrétion insulinique* qui ont aussi été identifiés forment avec *sécrétion d'insuline* un groupement de termes.

En japonais, les occurrences suivantes ont été relevées pour le même terme インスリン.分泌[7] (*sécrétion d'insuline*) :

– **forme de base** du patron `N N` : インスリン/N_1.分泌/N_2 ;

– **variante de composition** par agglutination de mots à la fin de la forme de base : インスリン/N_1.分泌/N_2.能力 /N_3 (*capacité de sécrétion d'insuline*).

À la différence du français, la version japonaise d'*ACABIT* n'effectue pas de regroupement de termes.

2.2. *Alignement terminologique*

Le processus d'alignement lexical peut être réalisé soit en adoptant la « méthode directe » (Fung, 1998; Peters et Picchi, 1998; Rapp, 1999) soit la méthode par « similarité interlangue » (Déjean et Gaussier, 2002; Morin et Daille, 2004). Nous rappelons brièvement ici le principe de chacune de ces méthodes, en mettant l'accent sur le processus de transfert des termes à traduire de la langue source à la langue cible qui en est la clef de voûte. Pour une présentation détaillée de chacune de ces méthodes, nous renvoyons les lecteurs à l'article de Morin et Daille (2004).

2.2.1. *Méthode directe*

Notre implémentation de la méthode directe, illustrée en figure 2.a, se décompose en quatre étapes :

Étape 1 : Identification des contextes lexicaux

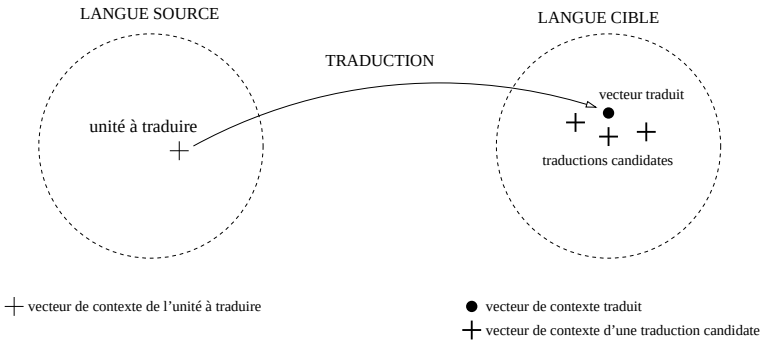
Pour chaque langue du corpus comparable, le contexte de chaque unité lexicale[8] i est extrait en repérant les unités qui apparaissent autour de i dans une fenêtre contextuelle de n mots. Afin d'identifier les unités lexicales caractéristiques des contextes lexicaux et de supprimer l'effet induit par la fréquence des unités lexicales, nous normalisons l'association entre les unités lexicales sur la base d'une mesure de récurrence contextuelle comme *Information Mutuelle* (Fano, 1961) ou *Taux de vraisemblance* (Dunning, 1993). À chaque unité i, nous associons ainsi un vecteur de contexte.

Étape 2 : Transfert d'une unité à traduire

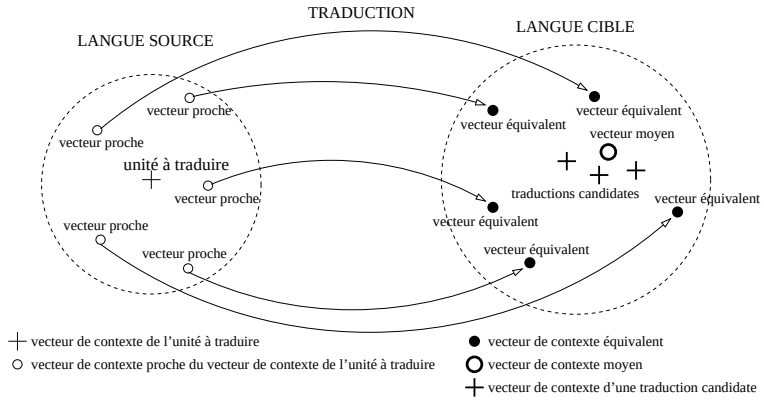
Le transfert d'une unité k à traduire de la langue source à la langue cible repose sur la traduction de chacun des éléments de son vecteur de contexte au

7. Pour tous les exemples de termes japonais, nous segmentons explicitement ses composants en utilisant le symbole « . ».

8. Nous employons ici le terme *unité lexicale* pour désigner la forme lemmatisée d'un mot, d'un terme simple ou d'un terme complexe. En ce qui concerne les mots fonctionnels, ils ne sont pas exploités dans le processus d'alignement. En japonais, les particules, auxiliaires des verbes, conjonctions et interjections sont reconnus comme mots fonctionnels.



(a) méthode directe



(b) méthode par similarité interlangue

Figure 2. *Processus de transfert d'une unité à traduire de la langue source à la langue cible*

moyen d'un dictionnaire bilingue. Dans le cas où le dictionnaire propose plusieurs traductions pour un élément, nous ajoutons au vecteur de contexte de l'unité k l'ensemble des traductions proposées (lesquelles sont pondérées par la fréquence de la traduction en langue cible). Dans le cas où la traduction échoue, l'élément ne sera pas exploité dans le processus de traduction.

En fonction de l'adéquation du dictionnaire bilingue avec le corpus d'étude, plus ou moins d'éléments du vecteur de contexte seront traduits. L'unité à traduire sera d'autant plus discriminante en langue cible que le nombre d'éléments traduits de son vecteur de contexte sera important.

Étape 3 : Identification des vecteurs proches de l'unité à traduire

Le vecteur de contexte ainsi traduit est ensuite comparé à l'ensemble des vecteurs de contexte de la langue cible en s'appuyant sur une mesure de distance vectorielle comme Cosinus (Salton et Lesk, 1968) ou Jaccard (Tanimoto, 1958).

Étape 4 : Obtention des traductions candidates

En fonction des précédentes valeurs de similarité, nous obtenons une liste ordonnée de traductions candidates.

2.2.2. *Méthode par similarité interlangue*

Dans le cadre de la méthode par similarité interlangue, le transfert d'une unité à traduire ne repose plus sur la traduction directe des différents éléments de son vecteur de contexte, mais sur la traduction des unités de la langue source qui lui sont « proches ». Notre implémentation de cette méthode, illustrée en figure 2.b, reprend les étapes 1 et 4 de la méthode directe et adapte les étapes 2 et 3 de la manière suivante :

Étape 2 : Transfert d'une unité à traduire

Le transfert d'une unité k à traduire repose, dans un premier temps, sur l'identification des unités lexicales dont les vecteurs de contexte lui sont similaires en exploitant une mesure de distance vectorielle comme Cosinus ou Jaccard. Le dictionnaire bilingue est ensuite utilisé pour assurer la traduction directe des unités similaires à k. Nous obtenons ainsi des vecteurs de contexte équivalents attestés en langue cible. Dans le cas où le dictionnaire bilingue propose plusieurs traductions pour une unité, comme il est essentiel de prendre en compte l'ensemble des vecteurs de contexte issus des différentes traductions, nous en réalisons l'union[9]. Le vecteur de contexte résultant est alors composé de l'ensemble des éléments des différents vecteurs de contexte originaux. Si plusieurs vecteurs ont un élément commun, alors le taux d'association de cet élément sera le plus grand des taux d'association des différents éléments. De cette manière, nous prenons en compte l'ensemble des traductions possibles et pas seulement la plus courante.

Ici encore, si la traduction échoue, l'unité lexicale ne sera pas exploitée dans le processus de traduction. Néanmoins, dans le cadre de cette méthode, la traduction n'altère par le vecteur de contexte puisque ce dernier est transféré directement de la langue source à la langue cible. Si une unité ne peut être traduite, elle ne sera pas prise en compte dans la recherche des vecteurs proches de l'unité à traduire. Cela ne pénalise pas, *a priori*, la recherche des unités similaires en langue cible.

Étape 3 : Identification des vecteurs proches de l'unité à traduire

En s'appuyant sur les vecteurs de contexte équivalents précédemment obtenus, nous calculons en langue cible un vecteur de contexte moyen. Celui-ci est obtenu par le barycentre des vecteurs de contexte équivalents pondéré par le coefficient de similarité issu de la langue source. Ce vecteur de contexte moyen est ensuite comparé à l'ensemble des vecteurs de contexte de la langue cible en exploitant une mesure de similarité comme Cosinus ou Jaccard.

9. Cette technique correspond à l'approche couramment adoptée lorsque les traductions ne sont pas ordonnées dans le dictionnaire bilingue.

3. Description des ressources

Nous décrivons dans cette section les différentes ressources textuelles utilisées dans cette étude : les corpus comparables, le dictionnaire bilingue et les lexiques de référence.

3.1. *Corpus comparable*

Pour démontrer l'importance du degré de comparabilité du corpus dans la tâche de fouille terminologique, nous devons disposer d'un corpus de langue de spécialité riche en terme de variation de discours ou de genres. Nous avons décidé de construire un corpus à partir du web dans le domaine du médical, restreint à la thématique « hygiène et santé » et à la sous-thématique des « régimes alimentaires ». Cette sous-thématique proche du corps valorise différemment selon les cultures la représentation du corps, la longévité, la force vitale, etc. Une première recherche de textes en français sur le web sur les « régimes alimentaires des êtres humains vivant en France » révèle une prépondérance des sites de type « esthétisant » dédiés à la réduction de la masse corporelle et dont le vocabulaire ne relève pas d'un domaine de spécialité : conseils de minceur, silhouette, esthétique, recettes de cuisine... Nous avons donc décidé de nous restreindre aux « maladies liées aux régimes alimentaires » et plus particulièrement au « diabète ». Le diabète, maladie des pays fortement industrialisés, fait l'objet d'une importante production langagière en français, mais aussi en japonais, comme l'atteste la présence de nombreux sites sur le web et l'existence de terminologies monolingue et multilingue (cf. section 3.3).

L'élaboration du corpus comparable s'est déroulée en trois étapes : le moissonnage du web, la sélection et la classification manuelle des textes, et leur normalisation. Les deux premières étapes ont été réalisées pour le français et le japonais par des locuteurs natifs et linguistes.

3.1.1. *Moissonnage du web*

Pour le moissonnage de documents sur le web, nous avons effectué :

1) une recherche sur l'ensemble du web à l'aide des moteurs `www.google.fr` pour le français et `www.google.co.jp` pour le japonais. En ce qui concerne le français, nous avons limité notre recherche aux pages francophones du domaine `.fr` en incluant les sites européens, mais en excluant les sites internationaux susceptibles de contenir des traductions en français de documents anglais comme celui de l'OMS. Nous n'avons identifié aucun site international hébergeant des documents français traduits en japonais, ou inversement, qui nous permettrait d'exploiter des corpus parallèles ;

2) une recherche interne sur des portails, notamment `http://www.diabsurf.com` pour le français, en utilisant le cas échéant le moteur de recherche propre au site.

La circonscription de la recherche à la thématique choisie s'effectue par l'intermédiaire de l'utilisation de mots-clés qui doivent être à la fois précis mais aussi permettre d'obtenir un large spectre de documents en terme de discours ou de genres. Deux stratégies de recherche sont possibles : la première en largueur qui examine la plupart des documents renvoyés par une seule requête, la seconde en profondeur qui n'examine que les premiers documents et en explore les liens internes. Pour le français, nous avons mené une recherche en profondeur sur les vingt premiers résultats en utilisant les combinaisons des mots-clés *alimentation*, *diabète* et *obésité*. Pour garantir une bonne couverture de la thématique, nous avons élargi la recherche en utilisant des synonymes[10] comme *conduite alimentaire* et en nous appuyant sur des termes relevés dans les pages visitées comme *hyperglycémie* ou *excès de poids*. Pour le japonais, nous avons mené une recherche en largeur sur une partie des 180 000 documents renvoyés par la combinaison des mots-clés 糖尿.病 (*diabète*) et 食事.療法 (*régime alimentaire*).

3.1.2. *Classification en fonction du type du discours*

Toutes les pages retenues ont été visualisées et leur pertinence à la thématique a été vérifiée. Très rapidement, deux types de discours, scientifique ou vulgarisé, correspondant à deux paramètres situationnels (Biber, 1993) ont émergé : les textes sont essentiellement écrits par des spécialistes (médecin, diététicien, infirmier, etc.) mais sont destinés, soit à un public restreint composé lui-même de spécialistes, soit à un public plus large incluant les malades, les populations à risque, etc. Les textes de non-spécialistes à non-spécialistes sont pratiquement inexistants. Nous avons décidé de classer les documents selon le type de discours scientifique ou vulgarisé, la seule contrainte étant d'atteindre la taille minimale de 200 000 mots pour chaque type de discours et pour chaque langue. Pour la classification, nous nous sommes appuyés non seulement sur le genre du document du web (Karlgren et Cutting, 1994; Beauvisage, 2001), mais aussi sur des critères internes reflétant le contexte sociolinguistique de production comme, par exemple, le niveau de style, la personnalisation, la technicité (Krivine et al., 2006). Pour les deux langues, le genre dominant du web est le rapport mais des différences dans les sous-genres apparaissent : pour le japonais, les rapports émanent presque exclusivement d'institutions privées alors que, pour le français, ils émanent de sites gouvernementaux, d'instituts universitaires ou d'institutions publiques (hôpitaux). La prépondérance du rapport d'institution privée pour le japonais s'explique par le statut privé et concurrentiel de l'hôpital. Notons une bonne présence du matériel journalistique et de pages interactives pour le français, ce qui est moins le cas pour le japonais. Les deux types de discours se déclinent dans tous les genres du web : par exemple, l'article de recherche relève du rapport scientifique, les conseils nutritionnels pour l'enfant diabétique du rapport vulgarisé. Quelques documents pour lesquels il est difficile de statuer sur le type de discours, comme ceux émanant de sectes ou de personnes dont le statut de spécialiste n'est pas certain, n'ont pas été conservés.

10. Présents dans le dictionnaire des synonymes `http://elsap1.unicaen.fr/cgi-bin/cherches.cgi`

3.1.3. *Normalisation*

Pour chacun des documents sélectionnés, nous avons enregistré son `url`, la date d'aspiration et l'avons converti au format UNICODE. Ce nouveau standard de codage de caractères est loin d'être la norme sur le web. Pour le japonais, 70 % des textes collectés sont encodés sous Shift-JIS (encodage par défaut sous Windows), 10 % sous EUC-JP (encodage UNIX/LINUX), 5 % sous ISO-2022-JP (encodage standard des industries japonaises), UTF-8 ne représentant que 1 %. Nous n'avons pas pu normaliser 14 % des documents présentant un encodage inconnu. Les formats de documents rencontrés sont principalement du `html` ou du `pdf`.

Le tableau 1 présente les principales caractéristiques des documents récoltés, à savoir leur nombre ainsi que le nombre de mots pour chaque langue et pour chaque type de discours[11]. En français comme en japonais, les documents scientifiques sont moins nombreux que les documents vulgarisés mais plus volumineux en terme de nombre de mots. L'élaboration du corpus a nécessité deux mois/homme par langue.

	Français		**Japonais**	
	Nb. doc.	**Nb. mots**	**Nb. doc.**	**Nb. mots**
Discours scientifique	65	425 781	119	234 857
Discours vulgarisé	183	267 885	419	572 430
Total	248	693 666	538	807 287

Tableau 1. *Caractéristiques générales des documents français-japonais récoltés à partir du web.*

À partir des documents précédemment récoltés, nous avons constitué deux corpus comparables : [corpus scientifique] avec les documents de discours scientifique et [corpus mixte] avec tous les documents, de discours scientifique ou vulgarisé.

3.2. *Dictionnaire bilingue*

Le dictionnaire français-japonais, utilisé dans cette étude, a été construit à partir de quatre ressources disponibles sur le web (notées [dico 1][12], [dico 2][13], [dico 3][14] et [dico 4][15]) et d'un dictionnaire électronique de termes techniques et scientifiques (Hak, 1989) (noté [dico 5]). En dehors du [dico 4] spécialisé dans le domaine médical,

11. Pour le japonais, le nombre de mots correspond au nombre d'occurrences des formes lexicales reconnues par Chasen (Matsumoto et al., 1999).
12. `http://kanji.free.fr/`
13. `http://quebec-japon.com/lexique/index.php?a=index&d=25`
14. `http://dico.fj.free.fr/index.php`
15. `http://quebec-japon.com/lexique/index.php?a=index&d=3`

les autres ressources sont des dictionnaires généralistes ou techniques. Le tableau 2 présente les principales caractéristiques des différents dictionnaires utilisés[16]. La fusion de ces différentes ressources permet de constituer une dictionnaire bilingue composé de 173 156 entrées distinctes pour le français avec en moyenne 2,1 traductions par entrée.

	Nature	**Nb. mots**	**Nb. mots simples**	**Nb. mots composés**	**Nb. traductions par entrée**
[dico 1]	généraliste	9 939	7 414	2 525	1,4
[dico 2]	généraliste	45 042	41 046	3 996	1,6
[dico 3]	généraliste	63 772	45 624	18 148	3,8
[dico 4]	médicale	2 329	1 136	1193	2,2
[dico 5]	technique	65 154	31 269	33 885	1,3
Total		**173 156**	**114 461**	**58 695**	**2,1**

Tableau 2. *Caractéristiques des différents dictionnaires français-japonais*

L'adéquation du dictionnaire bilingue au corpus comparable est primordiale dans le processus d'alignement puisqu'il assure le transfert direct ou indirect du vecteur de contexte de la langue source vers la langue cible. Afin de juger de l'apport de ce dictionnaire dans le processus d'alignement, nous avons compté le nombre de mots simples et composés (respectivement nb. MS et nb. MC) de l'espace vectoriel[17] qui peuvent être traduits à l'aide du dictionnaire (respectivement nb. MST et nb. MCT) et dont les traductions sont présentes dans le corpus comparable. Le tableau 3 présente les résultats obtenus pour chaque langue et pour chaque corpus. Pour les deux corpus, le taux de traduction des mots composés est très faible (autour de 1 %) en comparaison avec celui des mots simples (environ 30 % pour le français et 20 % pour le japonais) qui reste lui aussi faible.

3.2.1. *Élargissement de la couverture du dictionnaire bilingue*

Le faible taux de mots composés traduits directement à partir de notre dictionnaire bilingue constitue un handicap majeur dans le processus d'alignement, et plus particulièrement lors de l'étape de transfert d'une unité à traduire de la langue source à la langue cible. Afin de suppléer l'insuffisance du dictionnaire bilingue pour la traduction des mots composés, nous proposons d'en élargir la couverture en utilisant une approche par traduction compositionnelle (Janssen, 1996; Melamed, 1997; Grefenstette, 1999).

16. Pour disposer d'une ressource adaptée au corpus comparable, les entrées du dictionnaire bilingue ont été lemmatisées.

17. Nous rappelons que l'espace vectoriel est constitué d'unités lexicales qui ne sont pas des hapax et qui font référence à des mots simples ou aux termes complexes reconnus par *ACABIT*.

	Français		Japonais	
	Nb. MST/ Nb. MS	**Nb. MCT/ Nb. MC**	**Nb. MST/ Nb. MS**	**Nb. MCT/ Nb. MC**
[corpus scientifique]	2 300/7 443	80/7 225	2 614/12 941	78/8 655
[corpus mixte]	3 085/9 888	131/10 110	4 293/9 604	95/11 847

Tableau 3. *Éléments lexicaux des corpus comparables traduits à partir du dictionnaire bilingue*

Ainsi, pour chaque mot composé de la langue source absente du dictionnaire et identifié par *ACABIT*, nous traduisons chacun de ses composants. Par exemple, pour le terme français *fatigue chronique*, nous obtenons pour *fatigue* les 4 traductions suivantes :疲れ, 疲労, 倦怠 et 飽き et pour *chronique* les 2 traductions suivantes : 記事番組 et 慢性. Ensuite, nous combinons entre-elles les traductions des différents composants pour obtenir des traductions candidates du mot composé en langue cible. Pour le même exemple, nous obtenons les 8 traductions candidates[18] présentées en tableau 4. Les traductions retenues sont celles qui font référence à un mot composé existant en langue cible (c'est-à-dire identifié par la version japonaise d'*ACABIT*). Dans notre exemple, le seul mot composé identifié en japonais est 慢性.疲労.

chronique	*fatigue*
記事番組	疲れ
慢性	疲れ
記事番組	疲労
慢性	疲労
記事番組	倦怠
慢性	倦怠
記事番組	飽き
慢性	飽き

Tableau 4. *Exemple de traduction compositionnelle*

Cette approche présente aussi des limites (Baldwin et Tanaka, 2004; Brown et al., 1993; Morin et Daille, 2004) notamment en ce qui concerne la *fertilité*. Par exemple, le terme simple français *hypertension* est traduit en japonais par le terme complexe 高.血圧 (où le caractère 高 (*taka*) signifie *haut* et le terme 血圧 (*ketsuatsu*) désigne la *pression sanguine*). Elle diffère aussi de celle utilisée par Robitaille et al. (2006) pour la traduction compositionnelle de termes complexes anglo-

18. Entre le français et le japonais les constituants sont inversés.

japonais. Dans Robitaille et al. (2006), les termes de longueur n (avec $n > 2$) sont préalablement décomposés en toutes les combinaisons de termes de longueur inférieure ou égale à n éléments. Cette approche permet de pouvoir éventuellement traduire directement une sous-partie du terme complexe s'il est présent dans le dictionnaire bilingue. Par exemple, pour le terme *syndrome de fatigue chronique*, Robitaille et al. (2006) génèrent les quatre combinaisons suivantes i) [*syndrome de fatigue chronique*], ii) [*syndrome de fatigue*] [*chronique*], iii) [*syndrome*] [*fatigue chronique*] et iv) [*syndrome*] [*fatigue*] [*chronique*]. Pour la troisième combinaison, si le dictionnaire dispose de la traduction de *fatigue chronique* sa traduction sera alors directement utilisée. Dans ce travail, nous nous limitons à la dernière combinaison[19]. La première combinaison, quant à elle, sera obtenue directement si elle est présente dans le dictionnaire.

Avec cette approche compositionnelle, nous définissons un nouveau dictionnaire comportant 111 traductions de mots composés pour le [corpus scientifique] et 201 traductions pour le [corpus mixte]. En combinant ces entrées avec celles obtenues directement par le dictionnaire, le taux de traduction des mots composés passe de 1 à environ 3 %, ce qui reste globalement très faible.

3.2.2. *Analyse de la couverture du dictionnaire*

Afin de mieux comprendre la faible quantité de mots simples et composés traduits, nous les avons analysés. Les tableaux 5 et 6 présentent respectivement les différentes catégories de mots simples et composés à traduire[20].

En ce qui concerne les mots simples, nous constatons que sur 100 mots simples seulement 48 sont traduits à l'aide du dictionnaire. Parmi les 52 mots non traduits, 14 sont des mots mal orthographiés ou segmentés, des acronymes et entités nommées, ou des mots anglais. Les 34 mots bien formés mais non traduits sont principalement des verbes (*aboutir*, *obliger*, *justifier*, *suggérer*...) et dans une moindre mesure des adjectifs relationnels du domaine médical (*diabétogène*, *ischémique*, *organique*...) ou encore des noms spécialisés (*cataracte*, *glucotoxicité*, *mitochondrie*...).

Catégorie	**Exemple**	**Taux**
Mots communs corrects	*obliger*	34
Mots mal orthographiés ou segmentés	*traitementdu*	8
Acronymes et entités nommées	*VLDL*	4
Mots anglais	*drug*	2

Tableau 5. *Catégories de mots simples non traduits*

19. Environ 90 % des candidats termes fournis par *ACABIT*, après regroupement, ne sont composés que de deux mots pleins.
20. Dans les deux cas, l'analyse a été réalisée sur un échantillon de 100 mots choisis aléatoirement dans la partie française du corpus comparable.

Au niveau des mots composés, il est intéressant de noter que 32 d'entre eux ne peuvent pas être traduits dans la mesure où l'un des composants est absent du dictionnaire (ce qui est cohérent avec les résultats obtenus pour la traduction des mots simples présentés dans le tableau 3). Le niveau de bruit induit par *ACABIT* [21] est de l'ordre de 17 mots composés (mots composés mal orthographiés ou segmentés, anglais, incomplets ou incohérents). Il n'y a donc que 51 mots composés qui peuvent prétendre à une traduction compositionnelle.

Catégorie	**Exemple**	**Taux**
Mots composés où tous les éléments sont traduits	*facteur de croissance*	51
Mots composés dont un élément n'est pas traduit	*mortalité périnatale*	32
Mots composés mal orthographiés ou segmentés	*diabtique de type 2*	9
Mots composés anglais	*young adult*	4
Mots composés incomplets	*ministère chargé*	2
Mots composés incohérents	*laitier lait*	2

Tableau 6. *Catégories de mots composés à traduire*

Quoique Chiao et Zweigenbaum (2003) aient démontré l'intérêt d'utiliser un dictionnaire de langue générale en alignement de termes simples spécialisés pour un corpus français-anglais, le dictionnaire français-japonais que nous avons construit présente beaucoup de limites pour la traduction des termes complexes.

3.3. *Lexiques de référence*

Pour évaluer la qualité de la chaîne de fouille terminologique multilingue, nous avons construit deux listes de référence bilingues. La première liste sera utilisée pour nous positionner par rapport aux travaux existants en extraction de terminologies bilingues pour des termes simples et la seconde pour évaluer l'aide à la constitution de terminologies précises en se concentrant sur les termes complexes :

– [lexique_TS] rassemble 100 termes simples (TS) français dont la traduction japonaise est un terme simple. Ces termes n'appartiennent pas au dictionnaire bilingue.

– [lexique_TC] rassemble 60 termes simples ou complexes (TC) avec les caractéristiques suivantes : un terme simple français est traduit par un terme complexe japonais, ou inversement ; un terme complexe français est traduit par un terme complexe japonais. Là encore, ces termes ne peuvent pas être traduits directement ou à l'aide d'un processus de traduction compositionnelle à partir du dictionnaire bilingue.

21. À ce niveau, il s'agit d'une accumulation d'erreurs de la chaîne de traitement dont *ACABIT* est le reflet.

Les termes complexes du [lexique_TC] doivent répondre à trois contraintes supplémentaires :

1) ils attestent d'au moins deux occurrences dans le [corpus scientifique] ;

2) ils ont été proposés par les programmes d'extraction terminologique français et japonais ;

3) soit le terme français et sa traduction japonaise appartiennent au méta-thesaurus de l'UMLS[22], soit le terme français a été recensé par le *Grand dictionnaire terminologique*[23] dans le domaine de la médecine.

Ces contraintes ne nous ont pas permis d'atteindre la taille de 100 termes simples ou complexes pour le [lexique_TC]. L'intersection entre les candidats termes complexes français extraits, d'une fréquence de deux occurrences au moins et les référentiels terminologiques est de 177 termes. Sur ces 177 termes, seuls 60 d'entre eux ont une traduction en japonais qui correspond à un candidat terme complexe identifié dans le [corpus scientifique].

En extraction de terminologie bilingue à partir de corpus comparables spécialisés, les lexiques de référence exploités sont souvent composés d'une centaine de mots (180 termes simples dans les travaux de Déjean et Gaussier (2002), 97 termes simples pour Chiao et Zweigenbaum (2002) et trois lexiques de 100 termes simples ou composés pour Morin et Daille (2004)).

4. Expérimentations

L'objectif des expérimentations présentées ci-dessous est de vérifier la validité de notre hypothèse sur l'apport du type de discours en extraction de terminologies bilingues à partir de corpus comparables. Comme nous l'avons déjà indiqué, les principaux travaux en extraction de terminologies bilingues à partir de corpus comparables se focalisent principalement sur l'alignement de termes simples. Afin de pouvoir comparer notre travail à ces approches, nous proposons dans un premier temps de réaliser une première expérience avec le [lexique_TS] composé de termes simples. Puis, nous présenterons les résultats obtenus pour le [lexique_TC] qui permettent d'évaluer l'hypothèse de cette étude.

4.1. *Évaluation de l'alignement pour le* [*lexique_TS*]

Le tableau 7 présente les résultats de la première expérience[24] avec le [lexique_TS]. Pour chacun des corpus comparables et chacune des méthodes (MD pour méthode

22. http://www.nlm.nih.gov/research/umls
23. http://www.granddictionnaire.com/
24. Les résultats obtenus peuvent varier en fonction des paramètres utilisés. Les résultats présentés en tableau 7 sont ceux qui donnent le meilleur TOP_{20}. Il peut donc s'agir de résultats obtenus pour des combinaisons différentes de paramètres.

directe et MSI pour méthode par similarité interlangue), nous indiquons le nombre de traductions trouvées (NB_{trad}), la position moyenne (MOY_{pos}) et son écart type (ECT_{pos}) dans la liste ordonnée des 100 premières traductions proposées, puis le nombre de traductions trouvées dans la liste des 10 et 20 premières traductions proposées (TOP_{10} et TOP_{20}).

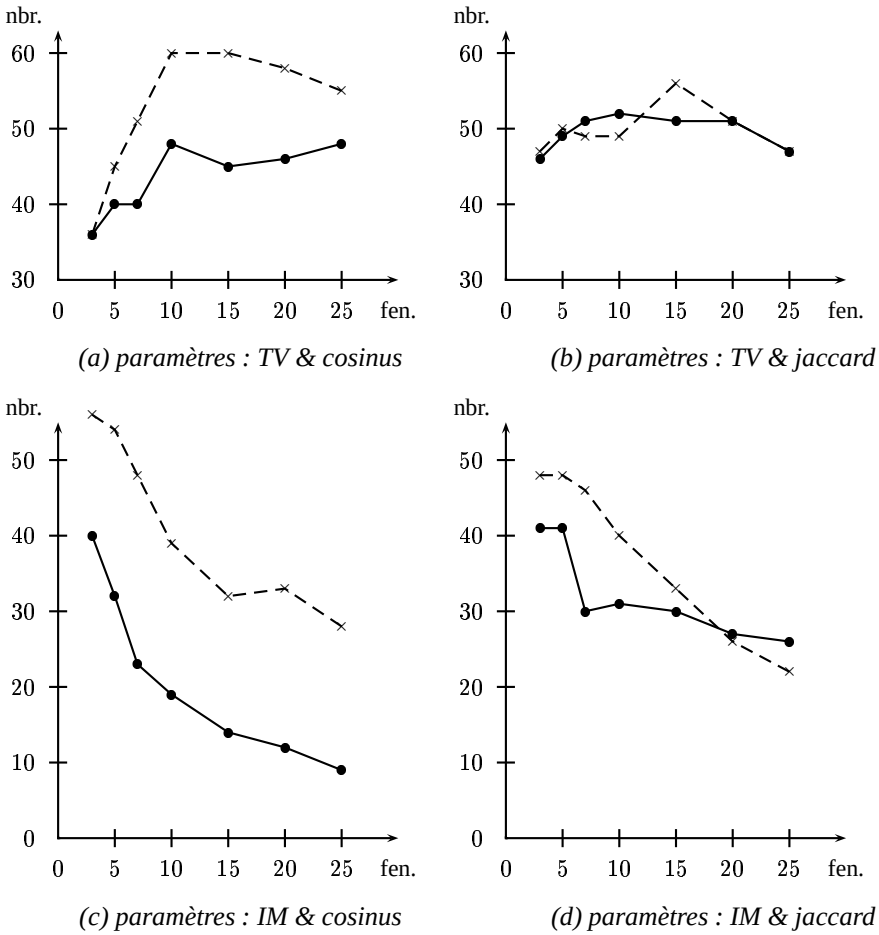
	NB_{trad}	MOY_{pos}	ECT_{pos}	TOP_{10}	TOP_{20}
[corpus scientifique] MD	64	11,6	20,2	49	52
[corpus scientifique] MSI	52	26,3	24,7	21	26
[corpus mixte] MD	76	11,5	16,3	51	60
[corpus mixte] MSI	55	23,3	23,1	19	29

Tableau 7. *Évaluation du processus d'alignement pour le* [*lexique_TS*]

Nous pouvons constater que les résultats obtenus pour la méthode directe sont bien meilleurs que ceux obtenus avec la méthode par similarité interlangue. En outre, le [corpus mixte] semble plus pertinent que le [corpus scientifique] dans cette expérience. En effet, les termes du [lexique_TS] étant limités à des termes simples, ils ne sont pas plus caractéristiques du discours scientifique que vulgarisé. Comme nous pouvons le constater sur la figure 3, ce comportement est presque toujours vérifié quels que soient les paramètres utilisés.

Comme nous l'avons mentionné en introduction à cette section, cette première expérience vise à nous positionner par rapport aux travaux existants sur les termes simples. À titre de comparaison, Déjean et Gaussier (2002) obtiennent 44 % et 57 % en méthode directe et 43 % et 51 % en méthode par similarité interlangue pour les 10 et 20 premiers candidats pour un corpus médical anglo-allemand de 100 000 mots (pour un lexique de référence composé de 1 800 mots qui comporte des mots de faible fréquence ainsi que des hapax). En utilisant un corpus anglo-allemand relatif aux sciences sociales de 8 millions de mots, Déjean et Gaussier (2002) obtiennent 35 % et 42 % pour la méthode directe et 79 % et 84 % pour la méthode par similarité interlangue pour les 10 et 20 premiers candidats (pour un lexique de référence de 180 termes simples dont la fréquence des termes est très souvent supérieure à 100 voire à 1000). Chiao et Zweigenbaum (2002) obtiennent, quant à eux, pour un corpus médical français-anglais de 1,2 million de mots : 61 % et 94 % de précision pour la méthode directe pour les 10 et 20 meilleurs candidats (pour un lexique de référence de 97 termes simples où la fréquence n'est pas précisée). Dans notre cas, nous obtenons 51 % et 60 % de précision pour les 10 et 20 premiers candidats pour un corpus français/japonais de 1,6 million de mots (pour un lexique de référence composé de 100 termes simples peu fréquents). Ces résultats, pour la méthode directe, se situent donc en dessous de ceux de Chiao et Zweigenbaum (2002) et au-dessus de ceux de Déjean et Gaussier (2002). Ils sont donc cohérents par rapport aux travaux existants et pour un couple de langues à grande distance linguistique.

Dans le cadre de cette expérience, nous travaillons avec des termes de faible fréquence. La fréquence des termes à traduire est un critère important de l'approche vectorielle. En effet, plus la fréquence du terme à traduire est importante, plus le vecteur de contexte associé sera discriminant. Le tableau 8 confirme bien cette hypothèse dans la mesure où les termes les plus fréquents sont les mieux traduits. Dans ce tableau



(a) paramètres : TV & cosinus *(b) paramètres : TV & jaccard*

(c) paramètres : IM & cosinus *(d) paramètres : IM & jaccard*

Figure 3. *Évolution du nombre de traductions trouvées en TOP_{20} en fonction de la taille de la fenêtre contextuelle pour différentes combinaisons de paramètres dans le cadre de la méthode directe pour le* [*lexique_TS*] ([*corpus scientifique*] —— *et* [*corpus mixte*] - - -*, les points indiqués sont les seules valeurs calculées)*

nous indiquons pour différents seuils de fréquence, le nombre de traductions trouvées par rapport au nombre de traductions présentes dans le lexique de référence.

	[2,10]	**[11,50]**	**[51,100]**	**[101,...]**
Français	3/17	12/29	17/23	28/31
Japonais	4/26	32/41	14/20	10/13

Tableau 8. *Fréquence des termes traduits du* [*lexique_TS*] *pour le* [*corpus mixte*] *appartenant au* TOP_{20} *pour la méthode directe (le résultat présenté correspond à la troisième ligne du tableau 7 où* $TOP_{20} = 60$*)*

4.2. *Évaluation de l'alignement pour le* [*lexique_TC*]

Les résultats de la deuxième expérience sont présentés dans le tableau 9[25]. L'analyse des résultats indique que seulement une faible quantité des termes du [lexique_TC] est retrouvée. Ce résultat était attendu dans la mesure où nous travaillons avec des corpus de petite taille. De la même manière que précédemment, la méthode directe donne de meilleurs résultats que la méthode par similarité interlangue. En revanche, les termes du [lexique_TC] sont mieux identifiés à partir du [corpus scientifique] que du [corpus mixte]. En outre, si nous comptons le nombre de termes communs qui sont correctement traduits entre le [corpus scientifique] et le [corpus mixte], nous retrouvons la plus grande partie des termes bien traduits avec le [corpus mixte] dans ceux obtenus avec le [corpus scientifique] [26]. Une analyse plus systématique de différentes combinaisons de paramètres (cf. figure 4) semble indiquer que ce comportement est souvent vérifié. Dans ce cas, il semble que l'exploitation d'un corpus générique (corpus scientifique et vulgarisé) n'apporte pas d'informations supplémentaires, mais vient ajouter du bruit par rapport aux résultats obtenus avec le seul corpus scientifique. Ce résultat semble confirmer l'intérêt de la prise en compte du type du discours dans la constitution du corpus qui va être utilisé pour proposer des terminologies bilingues spécialisées.

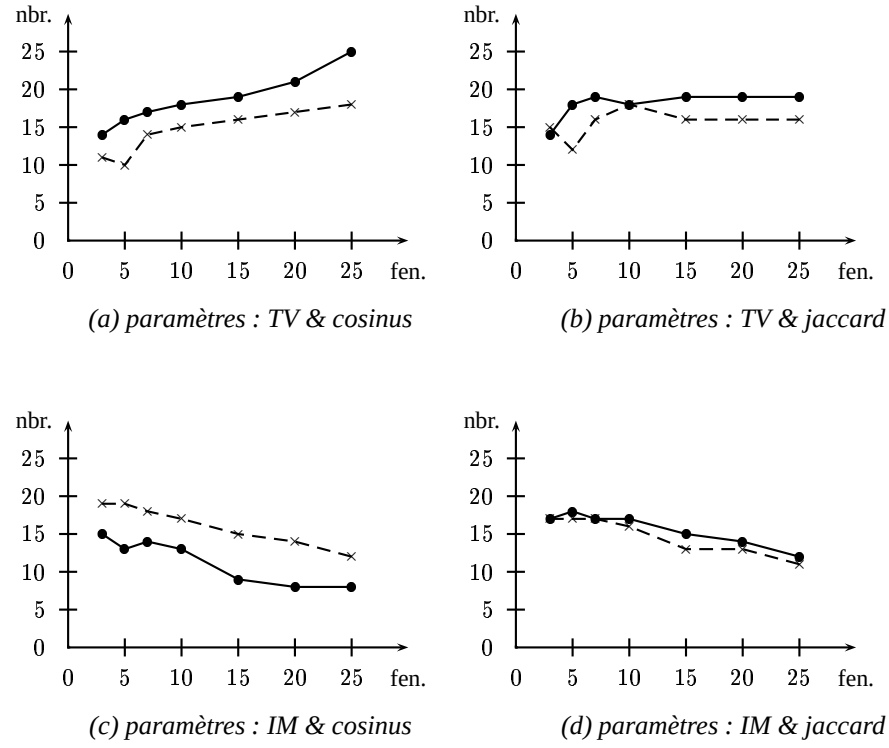
Ici encore, les résultats du tableau 10 confirment bien que la fréquence des termes à traduire est un critère important de l'approche vectorielle. Les termes les plus fréquents tels que *diabète* (#occ. 899 - 糖尿.病), *facteur de risque* (#occ. 267 - 危険.因子), *hyperglycémie* (#occ. 127 - 高.血糖), *tissu adipeux* (#occ. 62 - 脂肪.組織) sont les mieux traduits. Il est aussi intéressant de noter que des termes de faible fréquence comme *édulcorant* (#occ. 13 - 甘味.料) ou *choix alimentaire* (#occ. 11 - 食品.選択),

25. Comme pour la première expérience, les résultats obtenus peuvent varier en fonction des paramètres utilisés. Les résultats présentés en tableau 9 sont ceux qui donnent le meilleur TOP_{20}.

26. Pour la méthode directe nous obtenons pour le $TOP_{10}18 \cap 17 = 15$, pour le $TOP_{20}25 \cap 20 = 18$ et pour $NB_{trad}32 \cap 32 = 28$, et pour la méthode par similarité interlangue $TOP_{10}10 \cap 4 = 4$, pour le $TOP_{20}18 \cap 8 = 7$ et pour $NB_{trad}24 \cap 24 = 22$.

	NB_{trad}	MOY_{pos}	ECT_{pos}	TOP_{10}	TOP_{20}
[corpus scientifique] MD	32	16,1	21,9	18	25
[corpus scientifique] MSI	24	21,3	28,9	10	18
[corpus mixte] MD	32	23,9	27,6	17	20
[corpus mixte] MSI	24	38,0	28,5	4	8

Tableau 9. *Évaluation du processus d'alignement du* [*lexique_TC*]



(a) paramètres : TV & cosinus

(b) paramètres : TV & jaccard

(c) paramètres : IM & cosinus

(d) paramètres : IM & jaccard

Figure 4. *Évolution du nombre de traductions trouvées en TOP_{20} en fonction de la taille de la fenêtre contextuelle pour différentes combinaisons de paramètres dans le cadre de la méthode directe pour le* [*lexique_TC*] *(*[*corpus scientifique*] *—— et* [*corpus mixte*] *- - -, les points indiqués sont les seules valeurs calculées)*

ou de très faibles fréquences comme *obésité massive* (#occ. 6 - 高度.肥満) sont aussi identifiés par cette approche.

	[2,10]	**[11,50]**	**[51,100]**	**[101,...]**
Français	1/11	11/25	6/14	7/10
Japonais	5/21	13/25	5/9	2/5

Tableau 10. *Fréquence des termes traduits du* [*lexique_TC*] *pour le* [*corpus scientifique*] *appartenant au* TOP_{20} *pour la méthode directe (le résultat présenté correspond à la première ligne du tableau 9 où* $TOP_{20} = 25$*)*

4.3. *Discussion*

Notre objectif était de démontrer l'importance de la prise en compte du type du discours, garant d'une véritable comparativité des corpus, pour la fouille terminologique multilingue. Si les résultats obtenus confirment cette hypothèse, en particulier, dans le cadre de l'acquisition de termes complexes, la méthode employée montre aussi des limites et soulève plusieurs problèmes.

4.3.1. *Termes complexes en français et en japonais*

Malgré le fait que la méthodologie d'extraction terminologique soit la même dans les deux langues considérées, les termes complexes en français et en japonais ne sont pas de même nature : en français, ce sont des syntagmes nominaux alors qu'en japonais ce sont des composés morphologiquement construits. Un syntagme nominal en japonais n'est donc pas considéré comme un terme : ainsi, la version japonaise d'*ACABIT* ne propose pas 一.型.糖尿.病 (*diabète de type 1*) comme candidat terme mais uniquement 糖尿.病 (*diabète*), à l'inverse de la version française d'*ACABIT*. La moitié de la centaine de termes complexes français présents dans l'UMLS et incluant la racine *diabète* comme *diabète de type 1*, *diabète sucré*, *diabète avec insulinorésistance*, *régime diabétique*, qui sont extraits par la version française d'*ACABIT*, sont présents dans le partie japonaise du [corpus scientifique] mais malheureusement pas extraits par la version japonaise d'*ACABIT*. Ce problème a été mis en exergue lors de la construction de la liste du [lexique_TC] (cf. section 3.3). Il conviendra donc d'améliorer le parallélisme entre les deux extractions monolingues et d'étendre la couverture des termes complexes japonais aux syntagmes nominaux.

4.3.2. *Type de discours et multilinguisme*

Le type de discours est un élément important pour s'assurer de la comparabilité des textes. La terminologie relevant d'un discours scientifique est différente de celle d'un discours vulgarisé et nous l'avons constaté pour le français. Néanmoins, cette distinction semble beaucoup moins pertinente pour le japonais : les traductions d'une

cinquantaine de termes complexes extraits par *ACABIT* dans la partie française du [corpus scientifique] ont été identifiées par *ACABIT* dans la partie japonaise du [corpus mixte]. Cette atténuation du caractère distinctif du type du discours a sans doute une explication culturelle. Elle semble confirmer en tout cas la hiérarchie de catégorisation des textes proposée par Rastier (2001) où la langue domine le discours.

5. Conclusion

Cette étude, qui s'inscrit dans le cadre général de la fouille terminologique multilingue par des méthodes mixtes, visait à vérifier l'hypothèse que la qualité des données textuelles peut non seulement suppléer à leur quantité mais garantit aussi celle des ressources lexicales extraites. En particulier, nous souhaitions montrer l'intérêt d'exploiter de véritable corpus comparables pour obtenir des ressources terminologiques de qualité composées de termes complexes.

Nous avons ainsi collecté à partir du web des documents relevant exclusivement de la sous-thématique du diabète et de l'alimentation et les avons classés suivant leur appartenance au discours scientifique ou vulgarisé. À partir de ces documents, nous avons créé deux corpus comparables, l'un composé exclusivement de documents scientifiques, et l'autre de documents scientifiques ou vulgarisés. Nous avons ensuite exploité une chaîne de fouille terminologique multilingue, composée d'un extracteur de termes et d'un module d'alignement, pour évaluer l'impact de la caractéristique du type de discours du corpus comparable sur la constitution de listes terminologiques bilingues.

La première expérience réalisée avec une liste de référence composée de termes simples, nous a permis de nous positionner convenablement par rapport aux travaux existants en obtenant une précision de 51 % et 60 % pour les 10 et 20 premiers candidats pour la méthode directe. À ce niveau, l'apport de la classification suivant le type de discours n'est pas significative puisque les résultats sont meilleurs avec le corpus composé de documents scientifiques et vulgarisés par comparaison au corpus composé exclusivement de documents scientifiques. Ce résultat était attendu dans la mesure où les termes simples utilisés dans cette évaluation ne sont pas plus caractéristiques du discours scientifique que vulgarisé. Par exemple, le terme *excès* peut tout aussi bien faire référence au discours scientifique dans *excès pondéral* qu'au discours vulgarisé dans *excès de poids*.

La seconde expérience mettait en jeux une liste de référence composée de termes complexes caractéristiques du discours scientifique. Ici, les meilleurs résultats sont obtenus avec le corpus scientifique (précision de 30 % et 42 % pour les 10 et 20 premiers candidats pour la méthode directe). Plus précisément, il semble que les documents vulgarisés – qui utilisent par ailleurs les termes complexes du discours scientifique – induisent du bruit dans le corpus comparable. Dans ce cas, la prise en compte du type de discours agit bien comme un sélectionneur sémantique adéquat.

Une telle étude nous semble ouvrir de nouvelles perspectives à la fouille terminologique multilingue par la prise en compte du type de discours comme caractéristique pertinente lors de la création d'un corpus comparable au même titre que la période, ou la thématique...

Remerciements

Ce travail a été mené dans le cadre du projet DECO, programme CNRS-TCAN 2004-2006, en partenariat avec le NII, Kyo Kageura et Koichi Takeushi, et l'INALCO, Monique Slodzian et Natalia Grabar. Nous remercions Estelle Dubreil, Lorraine Goeuriot et Masaru Tomimitsu pour la constitution des corpus français et japonais.

6. Bibliographie

Baayen R. H., « Derivational Productivity and Text Typology », *Journal of Quantitative Linguistics*, vol. 1, n° 1, p. 16-34, 1994.

Baldwin T., Tanaka T., « Translation by Machine of Complex Nominals : Getting it Right », *Proceedings of the ACL 2004 Workshop on Multiword Expressions : Integrating Processing*, Barcelona, Spain, p. 24-31, July, 2004.

Beauvisage T., « Morphosyntaxe et genres textuels. Exploiter des données morphosyntaxiques pour l'étude statistique des genres textuels : application au roman policier. », *Traitement Automatique des Langues (TAL)*, vol. 42, n° 2, p. 579-608, 2001.

Biber D., « Representativeness in Corpus Design », *Literary and Linguistic Computing*, vol. 8, n° 4, p. 243-257, 1993.

Bowker L., Pearson J., *Working with Specialized Language : A Practical Guide to Using Corpora*, London/New York : Routeledge, 2002.

Brill E., « Some Advances in Transformation-Based Part of Speech Tagging », *Proceedings of the 12th National Conference on Artifi cal Intelligence (AAAI'94)*, Seattle, Washington, USA, p. 722-727, 1994.

Brown P., Della Pietra S., Della Pietra V., Mercer R., « The Mathematics of Statistical Machine Translation : Parameter Estimation », *Computational Linguistics*, vol. 19, n° 2, p. 263-311, 1993.

Cao Y., Li H., « Base Noun Phrase Translation Using Web Data and the EM Algorithm », *Proceeding of the 19th International Conference on Computational Linguistics (COLING'02)*, Tapei, Taiwan, p. 127-133, 2002.

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 identifi cation 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, The Netherlands, p. 397-402, 2003.

Daille B., « Terminology Mining », *in* M. Pazienza (ed.), *Information Extraction in the Web Era*, Springer, p. 29-44, 2003.

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

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

Fano R. M., *Transmission of Information : A statistical Theory of Communications*, MIT Press, Cambridge, MA, 1961.

Fung P., « A Statistical View on Bilingual Lexicon Extraction : From Parallel Corpora to Non-parallel Corpora », *in* D. Farwell, L. Gerber, E. Hovy (eds), *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA'98)*, Springer, Langhorne, PA, USA, p. 1-16, 1998.

Grefenstette G., « The Word Wide Web as a Ressource for Example-Bases Machine Translation Tasks », *ASLIB'99 Translating and the Computer 21*, London, UK, November, 1999.

Habert B., « Des corpus représentatifs : de quoi, pour quoi, comment ? », *in* M. Bilger (ed.), *Linguistique sur corpus. Études et réflexions*, n° 31 in *Cahiers de l'université de Perpignan*, Presses Universitaires de Perpignan, Perpignan, p. 11-58, 2000.

Hak, « Dictionnaire des termes techniques et scientifiques français-japonais », Hakusuisha, 1989. 4e édition.

Janssen T. M. V., « Compositionality », *in* J. van Benthem, A. ter Meulen (eds), *Handbook of Logic and Language*, Elsevier, Amsterdam, p. 417-473, 1996.

Karlgren J., Cutting D., « Recognizing Text Genres with Simple Metrics Using Discriminant Analysis », *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, vol. II, Kyoto, Japan, p. 1071-1075, 1994.

Krivine S., Tomimitsu M., Grabar N., Slodzian M., « Relever des critères pour la distinction automatique entre les documents médicaux scientifiques et vulgarisés en russe et en japonais », *Actes de la 13e conférence sur le Traitement Automatique des Langues Naturelles (TALN'06)*, Leuven, Belgique, p. 522-531, April, 2006.

Matsumoto Y., Kitauchi A., Yamashita T., Hirano Y., Japanese Morphological Analysis System ChaSen 2.0 Users Manual, Technical report, Nara Institute of Science and Technology (NAIST), 1999.

Melamed I. D., « A Word-to-Word Model of Translational Equivalence », *in* P. R. Cohen, W. Wahlster (eds), *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, Somerset, New Jersey, p. 490-497, 1997.

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

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.

Péry-Woodley M.-P., « Quels corpus pour quels traitements automatiques ? », *Traitement Automatique des Langues (TAL)*, vol. 36, n° 1-2, p. 213-232, 1995.

Peters C., Picchi E., « Cross-Language Information Retrieval : A System for Comparable Corpus Querying », *in* G. Grefenstette (ed.), *Cross-Language Information Retrieval*, Kluwer Academic Publishers, chapter 7, p. 81-90, 1998.

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, Maryland, USA, p. 519-526, 1999.

Rastier F., « Éléments de théorie des genres », *Revue électronique Texto !*, 2001. http ://www.revue-texto.net.

Robitaille X., Sasaki X., Tonoike M., Sato S., Utsuro S., « Compiling French-Japanese Terminologies from the Web », *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*, Trento, Italy, p. 225-232, April, 2006.

Salton G., Lesk M. E., « Computer Evaluation of Indexing and Text Processing », *Journal of the Association for Computational Machinery*, vol. 15, n° 1, p. 8-36, 1968.

Takeuchi K., Kageura K., Daille B., Romary L., « Construction of Grammar Based Term Extraction Model for Japanese », *in* S. Ananadiou, P. Zweigenbaum (eds), *Proceeding of the COLING 2004, 3rd International Workshop on Computational Terminology (COMPUTERM'04)*, Geneva, Switzerland, p. 91-94, August, 2004.

Tanimoto T. T., An elementary mathematical theory of classification, Technical report, IBM Research, 1958.