TALN 2004, Fès, 19–21 avril 2004

Extraction de terminologies bilingues à partir de corpus comparables

Emmanuel Morin, Samuel Dufour-Kowalski, Béatrice Daille
Université de Nantes - LINA - FRE CNRS 2729
2, rue de la Houssinière - BP 92208
44322 Nantes Cedex 3, France
{morin,dufour,daille}@lina.univ-nantes.fr

Résumé - Abstract

Cet article présente une méthode pour extraire, à partir de corpus comparables d'un domaine de spécialité, un lexique bilingue comportant des termes simples et complexes. Cette méthode extrait d'abord les termes complexes dans chaque langue, puis les aligne à l'aide de méthodes statistiques exploitant le contexte des termes. Après avoir rappelé les difficultés que pose l'alignement des termes complexes et précisé notre approche, nous présentons le processus d'extraction de terminologies bilingues adopté et les ressources utilisées pour nos expérimentations. Enfin, nous évaluons notre approche et démontrons son intérêt en particulier pour l'alignement de termes complexes non compositionnels.

This article presents a method of extracting bilingual lexica composed of simple and multi-word terms from comparable corpora of a technical domain. First, this method extracts the multi-word terms in each language, and then uses statistical methods to align them by exploiting the term contexts. After explaining the difficulties involved in aligning multi-word terms and specifying our approach, we show the adopted process for bilingual terminology extraction and the resources used in our experiments. Finally, we evaluate our approach and demonstrate its significance, particularly in relation to non-compositional multi-word term alignment.

Mots-clefs – Keywords

Terminologie bilingue, corpus comparable, termes complexes
Bilingual terminology, comparable corpora, multi-word terms

1 Introduction

Les recherches en extraction automatique de lexiques bilingues à partir de corpus se sont principalement concentrées sur l'exploitation de corpus parallèles (Veronis, 2000). À partir de textes alignés phrases à phrases, des techniques symboliques (Carl & Langlais, 2002), statistiques

(Gaussier & Langé, 1995) ou mixtes (Daille *et al.*, 1994) sont mises en œuvre pour aligner les différents éléments constitutifs de ces phrases. Les textes parallèles restent néanmoins des ressources rares, principalement pour des couples de langues ne faisant pas intervenir l'anglais. Les recherches récentes se sont donc intéressées à l'exploitation de corpus comparables. Déjean *et al* (2002) donnent la définition suivante d'un corpus comparable :

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

Les principaux travaux d'extraction de lexiques bilingues à partir de corpus comparables se focalisent sur l'extraction de termes simples bilingues. Ainsi, Fung (1998) extrait des couples de termes simples anglais/chinois avec une précision de 76 % sur les 20 premiers candidats proposés en exploitant deux ans du Wall Street Journal et du quotidien japonais Nikkei Financial News. Rapp (1999) porte cette précision à 89 % sur les 10 premiers candidats en exploitant des couples de termes simples anglais/allemand à partir d'un corpus journalistique de 85 millions de mots. Déjean & Gaussier (2002) obtiennent quant à eux une précision de 84 % sur les 10 premiers candidats pour des couples anglais/allemand à partir d'un corpus médical de 8 millions de mots. Pour ce qui est de l'extraction de groupes nominaux bilingues relevant du domaine général à partir de corpus comparables, elle a été réalisée avec succès par Cao & Li (2002) qui obtiennent une précision de 91 % sur les 3 premiers candidats en exploitant le web.

Si les résultats obtenus dans le domaine de l'extraction de lexiques bilingues à partir de corpus comparables sont encourageants, ils restent néanmoins limités soit à l'extraction de termes simples bilingues pour des textes génériques ou spécialisés, soit à l'extraction de groupes nominaux bilingues pour des textes relevant du domaine général. Nos travaux s'intéressent à l'extraction de termes complexes bilingues pour des domaines spécialisés dans des textes comparables.

Si les termes complexes bilingues rendent mieux compte de la terminologie d'un domaine, leur extraction pose un problème plus délicat que celle des termes simples, du fait que :

- Les termes simples et complexes ne se traduisent pas systématiquement par un terme de même longueur. Par exemple, le terme complexe *peuplement forestier* est traduit en anglais par le terme simple *crop*, et le terme *essence d'ombre* par *shade tolerant species*. Ce problème bien connu, décrit par Brown *et al.* (1993) sous le terme de *fertilité*, est rarement pris en compte dans l'extraction de lexiques bilingues. Une hypothèse de traduction *mot à mot* étant le plus souvent adoptée.

- La traduction d'un terme complexe n'est pas toujours obtenue par la traduction de ces composants (Melamed, 2001). Par exemple, le terme *plantation énergétique* est traduit en anglais par *fuel plantation*, où *fuel* n'est pas la traduction de *énergétique*.

- Un même terme peut se présenter sous différentes formes suite à des variations morphologique, syntaxique ou encore sémantique (dans le cas de la synonymie). Les variations des termes doivent donc être pris en compte dans le processus de traduction. Par exemple, les termes français *aménagement de la forêt* et *aménagement forestier* sont traduits par le même terme en anglais : *forest management*.

Notre approche prend en compte ces trois caractéristiques. Nous adoptons une approche mixte qui extrait en premier à l'aide d'une méthode linguistique les termes complexes dans chacune des langues puis tente de les aligner à l'aide d'une méthode statistique. L'alignement s'effectue

en comparant les contextes lexicaux des termes à traduire et ceux des termes identifiés en langue cible. Cette méthode permet d'éviter une traduction uniquement compositionnelle des termes complexes et rend possible des alignements de termes de longueur différente.

2 Processus d'extraction

Nous présentons dans cette section le processus d'extraction de terminologies bilingues mis en place qui se décompose en une phase de traitement linguistique et une phase de traitement statistique.

2.1 Traitement linguistique

L'objectif premier de cette phase est l'identification de l'ensemble des termes complexes présents dans le corpus. Pour ce faire, le corpus est nettoyé des données indésirables et segmenté en occurrences de phrases et de formes. Ce dernier est ensuite étiqueté et lemmatisé. Enfin, les candidats termes sont extraits du corpus par un outil d'acquisition terminologique fonctionnant sur le français et sur l'anglais : *ACABIT*[1] (Daille, 2003). Ce programme parcourt le corpus et relève les cooccurrences de candidats termes ou de leurs variations qui épousent des structures morphosyntaxiques prédéfinies. Les différentes occurrences qui réfèrent à un terme binaire ou à l'une de ses variantes sont regroupées sous un même candidat terme. Puis, une analyse morphologique est effectuée qui permet de regrouper sous un même candidat terme des variantes morphologiques dérivationnelles synonymiques comme *aménagement de la forêt* et *aménagement forestier*. Les variantes morphologiques et morphosyntaxiques induisant une distance sémantique ne sont pas regroupées sous le même candidat terme mais forment un groupement de candidats termes. Ainsi, les séquences comme *bois chauffé* et *bois non chauffé* forment un groupement de deux candidats termes liés par une relation d'antonymie. Dans la suite du traitement, nous considérons comme variantes de termes uniquement celles qui sont regroupées sous le même candidat terme. Cette opération, qui correspond à une normalisation terminologique, améliore le découpage du texte en unités de sens au même titre que la lemmatisation au niveau morphologique. De cette manière, nous ne traduisons pas un terme mais une classe d'équivalence de termes.

2.2 Traitement statistique

L'objectif de cette seconde phase, qui reprend et adapte la méthode proposée par Déjean & Gaussier (2002), a pour objectif de réaliser l'alignement des termes de la langue source avec ceux de la langue cible. Comme Déjean *et al.* (2002) pour les termes simples, nous supposons que la traduction des termes proches du terme à traduire donnera de meilleurs résultats que la traduction directe des termes de son contexte lexical (Fung, 1998; Rapp, 1999).

[1] http://www.sciences.univ-nantes.fr/info/perso/permanents/daille/

2.2.1 Calcul des vecteurs de contexte

Dans un premier temps, le contexte de chaque unité lexicale[2] i est extrait en repérant les unités qui occurrent avec i dans une *fenêtre contextuelle*[3] de n phrases autour de celle comportant i. Pour chaque unité i, nous créons ainsi un *vecteur de contexte* v_i. À chaque élément j de v_i, nous associons le nombre de fois où j et i occurrent ensemble occ_j^i.

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 occ_j^i sur la base d'une mesure de récurrence contextuelle comme Information Mutuelle et Taux de vraisemblance (cf. équations 1 et 2 où N représente la taille du corpus). Après normalisation, à chaque élément j du vecteur de contexte de l'unité i nous associons le taux d'association $assoc_j^i$.

$$assoc_j^i = log \frac{a}{(a+b)(a+c)} \tag{1}$$

$$\begin{aligned} assoc_j^i = \quad & a\,log\,a + b\,log\,b + c\,log\,c + d\,log\,d + N\,log\,N \quad (a+b)\,log\,(a+b) \\ & (a+c)\,log\,(a+c) \quad (b+d)\,log\,(b+d) \quad (c+d)\,log\,(c+d) \end{aligned} \tag{2}$$

(avec $a = occ_j^i \quad b = occ_j \quad occ_j^i \quad c = occ_i \quad occ_j^i \quad d = N \quad occ_j \quad occ_i + occ_j^i$)

2.2.2 Transfert du vecteur de contexte de la langue source à la langue cible

Afin d'éviter les inconvénients de la traduction directe des vecteurs de contexte avec un dictionnaire, à savoir l'inadéquation des ressources bilingues aux corpus et l'impossibilité de traduire certains éléments des vecteurs de contexte, la traduction d'un vecteur de contexte est réalisée en s'appuyant sur les vecteurs de contexte qui lui sont proches. Le dictionnaire va ainsi permettre de « traduire » les vecteurs de contexte dans leur globalité et non élément par élément.

Ainsi, pour chaque unité k à traduire, nous identifions les vecteurs de contexte proches de v_k, en s'appuyant sur une mesure de distance vectorielle comme Cosinus et Jaccard (cf. équations 3 et 4). Par la suite, nous désignons par vecteur de similarité de l'unité k un vecteur dont les éléments sont les unités lexicales identifiant les vecteurs de contextes proches de v_k. À chaque élément l du vecteur de similarité de l'unité k, nous associons le taux de similarité $simil_{v_l}^{v_k}$ entre v_l et v_k. Le dictionnaire bilingue est alors utilisé pour assurer la traduction des éléments du vecteur de similarité de l'unité à traduire. Nous obtenons ainsi des vecteurs de contexte équivalents attestés en langue cible.

$$simil_{v_l}^{v_k} = \frac{\sum_t assoc_t^l\, assoc_t^k}{\sqrt{\sum_t {assoc_t^l}^2\, {assoc_t^k}^2}} \tag{3}$$

$$simil_{v_l}^{v_k} = \frac{\sum_t min(assoc_t^l, assoc_t^k)}{\sum_t {assoc_t^l}^2 + \sum_t {assoc_t^k}^2 \quad \sum_t assoc_t^l\, assoc_t^k} \tag{4}$$

La figure 1 est une illustration du processus de traduction sur un espace à deux dimensions. La position de l'unité à traduire est déterminée par son vecteur de contexte. L'espace vectoriel réel a un nombre de dimensions égal au nombre d'unités différentes du corpus.

[2]Nous employons ici le terme unité lexicale pour désigner un mot, un terme simple ou un terme complexe.

[3]Afin de diminuer le bruit et d'accélérer le calcul des contextes lexicaux, nous supprimons du corpus les mots fonctionnels.

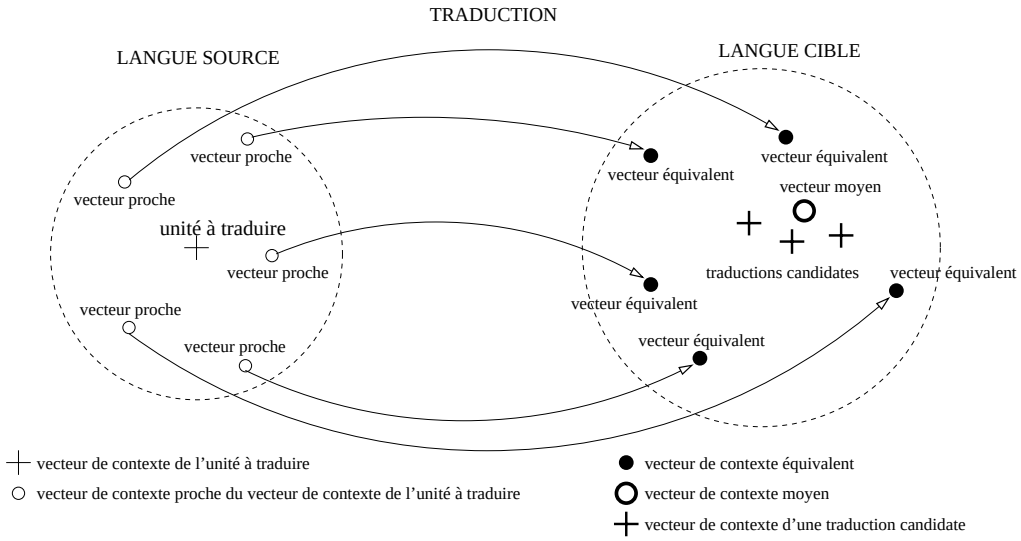

Figure 1: Processus de transfert des vecteurs de contexte de la langue source à la langue cible.

Suivant la longueur de l'unité lexicale à traduire, nous réalisons un traitement différent :

Traduction d'une unité de longueur 1 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*. 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.

Traduction d'une unité de longueur strictement supérieure à 1 Si l'unité est présente dans le dictionnaire, elle peut être traduite directement. Dans le cas contraire, nous avons adopté une approche reposant sur la traduction compositionnelle de ses composants. Chaque composant de l'unité est traduit à l'aide du dictionnaire bilingue et l'ensemble des combinaisons relevées par *ACABIT* est retenu. Nous réalisons alors l'union des vecteurs de contexte ainsi obtenus de la même manière que précédemment. Si la traduction compositionnelle échoue, l'unité n'est pas exploitée dans le processus de traduction.

2.2.3 Recherche des traductions en langue cible

Nous calculons en langue cible un vecteur de contexte moyen à partir des différents vecteurs de contexte équivalents issus de la traduction du vecteur de similarité en langue source. Ce vecteur moyen est obtenu par le barycentre des vecteurs de contexte équivalents pondéré par le coefficient de similarité $simil_{v_l}^{v_k}$ obtenu en langue source. Ainsi, nous favorisons les unités les plus proches de l'unité lexicale à traduire. Ce vecteur de contexte moyen est ensuite comparé à l'ensemble des vecteurs de contexte de la langue cible par la mesure du Cosinus. De cette manière, nous obtenons une liste ordonnée de traductions candidates.

3 Description des ressources

Nous présentons dans cette section les ressources utilisées pour nos expérimentations, à savoir le corpus comparable, le dictionnaire bilingue et le lexique de référence.

3.1 Corpus comparable

Dans le cadre de nos expériences, nous avons constitué un corpus comparable à partir de la revue *Unasylva* publié chaque trimestre en anglais, espagnol et français depuis 1947 par la FAO[4]. Cette revue internationale consacrée aux forêts et aux industries forestières couvre autant des aspects liés à la gestion et la conservation des plantations, des forêts et des animaux, que des aspects liés aux développements socio-économiques, au commerce international et à l'environnement. Afin d'obtenir un corpus comparable français/anglais, nous avons sélectionné les textes qui ne sont pas la traduction l'un de l'autre. Nous obtenons ainsi un corpus comparable composé de 2,6 millions de mots pour le français et de 2,3 millions pour l'anglais.

3.2 Dictionnaire bilingue

Un dictionnaire bilingue, nécessaire au processus d'alignement, a été construit à partir de ressources disponibles sur le web. Il est composé de 22 300 mots simples en français avec en moyenne 1,6 traductions par entrée. Il s'agit donc d'un dictionnaire de langue générale qui ne contient que peu de termes en rapport avec le domaine de la sylviculture.

3.3 Lexique de référence

Dans cette étude, l'évaluation du processus d'extraction de terminologies bilingues a été réalisée automatiquement à l'aide d'un lexique de référence. Celui-ci a été construit à partir de trois ressources terminologiques :

1. Le glossaire bilingue de la terminologie de la sylviculture au Canada du service canadien des forêts[5]. Celui-ci couvre les domaines usuels de la pratique de la sylviculture au Canada. Il est composé de 700 termes spécialisés dont 70 % sont des termes complexes.

2. Le lexique multilingue du projet Eurosilvasur (plate-forme ressource forêt-bois-papier des régions de l'Europe du Sud)[6]. Celui-ci couvre un ensemble de domaines liés à l'exploitation de la forêt : économie et exploitation forestière, transformation du bois, sylviculture, etc. Il est composé de 2 800 termes dont 66 % sont des termes complexes.

3. Le thesaurus multilingue AGROVOC de la FAO[7]. Ce thesaurus, destiné à l'indexation des données entrant dans les systèmes d'informations agricoles, couvre les domaine de

[4] http://www.fao.org/forestry/foris/webview/forestry2/index.jsp?siteId=2342
[5] http://nfdp.ccfm.org/silviterm/silvi_f/silvitermintrof.htm
[6] http://www.eurosilvasur.net/francais/lexique.php
[7] http://www.fao.org/agrovoc/

l'agriculture, de la pêche, de la sylviculture, de la nutrition, de l'innocuité des produits alimentaires, ainsi que divers sujets connexes comme l'environnement. Il comporte 15 000 descripteurs pour le français dont 47 % sont des termes complexes.

Ces trois ressources terminologiques sont complémentaires dans la mesure où elles proposent des termes plus ou moins spécialisés. Ainsi, le glossaire bilingue est plus spécialisé que le lexique multilingue, qui lui-même est plus spécialisé que le thesaurus multilingue.

À partir de ces ressources, nous avons sélectionné automatiquement 300 termes français, où chaque terme est au moins présent cinq fois dans le corpus comparable, pour constituer notre lexique de référence. Ces termes se répartissent en trois sous-listes :

- [liste 1] 100 termes simples français dont la traduction, qui n'est pas présente dans notre dictionnaire bilingue, est un terme simple.
- [liste 2] 100 termes complexes français dont la traduction peut être un terme simple ou complexe. Ces termes ne peuvent pas être traduits directement ou de manière compositionnelle à partir de notre dictionnaire bilingue.
- [liste 3] 100 termes complexes français dont la traduction est un terme complexe. Ces termes ne peuvent pas être traduits directement mais le sont de manière compositionnelle par leurs composants à partir de notre dictionnaire bilingue.

Ce lexique de référence est essentiellement composé de termes peu fréquents (cf. table 1). Deux raisons majeures expliquent ce phénomène. D'une part, les différentes ressources utilisées pour créer le lexique de référence proposent des termes très spécifiques ou très génériques. D'autre part, le corpus utilisé couvre un grand nombre de domaines liés à la foresterie, il ne constitue pas une ressource très spécialisée.

	$NB_{occ.} < 50$	$NB_{occ.} \leq 100$	$NB_{occ.} \leq 1\,000$	$NB_{occ.} > 1\,000$
[liste 1]	50	21	18	11
[liste 2]	54	21	25	0
[liste 3]	51	18	29	2

Table 1: Fréquence dans le corpus comparable des termes français du lexique de référence

4 Évaluation

Nous présentons dans cette section les résultats de l'évaluation du processus d'extraction de terminologies bilingues pour les ressources précédemment présentées.

4.1 Estimation des paramètres

Comme nous l'avons vu en section 2, la chaîne de traitement utilisée pour l'extraction de termes bilingues fait intervenir de nombreux paramètres. En fonction de ces derniers, les résultats varient fortement. Les résultats les plus intéressants ont été obtenus avec les paramètres suivants :

1. Taille de la fenêtre contextuelle : 3 phrases.
2. Réduction des éléments contenus dans les vecteurs de contexte aux termes simples.
3. Mesure d'association pour le calcul des vecteurs de contexte : Information Mutuelle ou Taux de vraisemblance.
4. Taille des vecteurs de contexte : 30 ou 20 éléments.
5. Mesure de similarité pour le calcul des vecteurs de similarité : Cosinus ou Jaccard.
6. Taille des vecteurs de similarité : 30 ou 20 éléments.

4.2 Analyse des résultats

La table 2 présente les résultats bruts de nos expériences. Pour chacune de trois listes de 100 termes, nous indiquons le nombre de traductions trouvées (NB_{trad}), la position moyenne et son écart type pour les traductions trouvées dans la liste ordonnée de traductions proposées (AVG_{pos}, $STDDEV_{pos}$).

L'analyse de ces résultats montre que les termes complexes de la [liste 3] (qui se traduisent de manière compositionnelle) sont relativement bien repérés et apparaissent le plus souvent dans les 20 meilleurs candidats. En revanche les termes des [listes 1 et 2] sont moyennement repérés, mais plus encore, n'apparaissent que rarement dans les 20 premiers candidats.

En dehors de l'analyse numérique de ces résultats, il convient de préciser que les traductions proposées pour un terme donné se situent le plus souvent dans le même champ sémantique. Ce qui démontre l'intérêt de notre approche. Par exemple, les 20 premières traductions proposées pour le terme *gaz à effet de serre* de fréquence 33 sont *carbon*, *carbon cycle*, *atmosphere*, *greenhouse gas*, *greenhouse*, *global carbon*, *atmospheric carbon*, *emission*, *sink*, *carbon dioxide*, *fossil fuel*, *fossil*, *carbon pool*, *mitigate*, *global warming*, *climate change*, *atmospheric*, *dioxide*, *sequestration*, *quantity of carbon*.

	NB_{trad}	AVG_{pos}	$STDDEV_{pos}$
[liste 1]	60	35,1	38,5
[liste 2]	65	37,5	45,9
[liste 3]	90	3,4	15,2

Table 2: Évaluation du processus d'extraction de terminologies bilingues

Comme nous l'avons indiqué précédemment, nos résultats varient fortement en fonction des paramètres choisis pour nos expérimentations. Il est malheureusement assez difficile de statuer précisément sur le rôle de chaque paramètre en raison des temps de calculs importants et du manque de traçabilité de la chaîne de traitement. Cependant une analyse manuelle des traductions candidates pour différentes configurations a permis de mettre en évidence que :

- Des termes correctement traduits pour une configuration de paramètres donnée ne le sont pas forcément dans une autre, et inversement, des termes non traduits dans la première

configuration peuvent être traduits correctement dans la deuxième. De ce fait, il est difficile de déterminer quelle est la configuration à adopter. Ce phénomène est principalement visible pour les [listes 1 et 2].

- Plus précisément, pour un terme donné, les premières traductions candidates pour différentes configurations de paramètres sont souvent différentes. Par exemple, pour le terme *pâte à papier (paper pulp)*, les 50 meilleures traductions candidates d'une vingtaine de configurations différentes n'ont qu'un trentaine d'éléments en commun.

- Par extension, la bonne traduction d'un terme donné apparaît souvent à des positions très différentes. Cette position variant en fonction des configurations.

À partir de ce constat et afin de capturer davantage de bonnes traductions, nous avons choisi de prendre en compte les résultats donnés par différentes configurations de paramètres et non par une seule. Pour cela, le processus de traduction est exécuté sur un ensemble de configurations différentes et nous ne retenons que les x meilleures traductions candidates pour chacune d'entre elles. Ces différents résultats sont alors fusionnés en sommant les scores de traductions associés à chaque candidat pour les différentes configurations. Ainsi, nous ne retenons que les candidats les mieux « classés » sur l'ensemble des configurations. Cela permet d'améliorer nettement la position des traductions correctes parmi l'ensemble des traductions candidates. Il convient cependant de choisir un x qui n'est pas trop grand, par exemple 20, afin d'améliorer le comportement et de supprimer du bruit inutile.

Nous avons ainsi fusionné différentes configurations, en particulier en faisant varier les tailles des vecteurs de contexte et de similarité et les mesures d'association et de similarité[8].

	NB_{trad}	AVG_{pos}	$STDDEV_{pos}$	**Top 10**	**Top 20**
[liste 1]	60	22,7	26,1	41	51
[liste 2]	65	21,3	25,0	45	55
[liste 3]	90	2,1	10,3	87	88

Table 3: Évaluation du processus d'extraction par combinaison de paramètres

Les résultats obtenus par cette méthode, présentés en table 3, améliorent globalement le processus d'extraction de termes bilingues. Les résultats obtenus pour la [liste 3] restent très satisfaisants. En revanche ceux de la [liste 1], quoique meilleurs, restent inférieurs aux résultats obtenus par Déjean & Gaussier (2002) : pour les 10 et 20 meilleurs candidats 43 % et 51 % pour un corpus médical de 100 000 mots et 79 % et 84 % pour un corpus de 8 millions de mots. Cette différence de résultats peut s'expliquer par notre évaluation automatique du processus de traduction qui est sans doute plus contraignante qu'une évaluation humaine. Par exemple, notre lexique de référence indique que la traduction du terme *piste de débardage* est *haulage road*. Ce terme n'est jamais présent dans la liste des traductions candidates. Par contre, le terme *skid trail* est présent dans les 20 premiers candidats. Même si cette traduction nous semble acceptable, elle n'est pas comptabilisée comme valide. Nous avons aussi délibérément ôté de nos listes de termes simples, ceux étant déjà présents dans le dictionnaire.

[8]Pour des raisons de temps de calcul, nous ne faisons pas varier la taille de la fenêtre contextuelle. Nous limitons aussi les unités contenues dans les vecteurs de contexte à des termes simples, puisque nos expériences indiquent que cette configuration est la plus intéressante.

En ce qui concerne, les résultats de la [liste 2], ils sont d'une qualité légèrement supérieure à ceux de [liste 1]. Cette méthode semble donc intéressante, en particulier pour les termes dont la traduction est non compositionnelle.

5 Conclusion

Nous avons proposé et évalué une méthode mixte pour l'extraction de termes complexes bilingues prenant en compte les problèmes de fertilité, de non compositionnalité et de variation. Cette méthode extrait d'abord les termes complexes dans chaque langue, puis les aligne à l'aide de méthodes statistiques exploitant le contexte des termes. Cette approche mixte donne des résultats satisfaisants comparables à ceux obtenus pour les mots simples. Elle permet d'obtenir la traduction de termes non compositionnelle. Les prochains travaux porteront sur l'examen des groupes de termes identifiés lors de la phase d'extraction terminologique, sur le délicat problème de réglage des paramètres et sur l'extension de la méthode à d'autres langues.

Références

BROWN P., DELLA PIETRA S., DELLA PIETRA V., MERCER R. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, Vol. 19(2), 263–311.

CAO Y., LI H. (2002). Base Noun Phrase Translation Using Web Data and the EM Algorithm. In *Actes, COLING'02*, p. 127–133.

CARL M., LANGLAIS P. (2002). An intelligent Terminology Database as a pre-processor for Statistical Machine Translation. In *Actes, COMPUTERM'02*, p. 15–21.

DAILLE B. (2003). Terminology Mining. In M. PAZIENZA, Ed., *Information Extraction in the Web Era*, p. 29–44. Springer.

DAILLE B., GAUSSIER E., LANGÉ J. (1994). Towards Automatic Extraction of Monolingual and Bilingual Terminology. In *Actes, COLING'94*, p. 515–521.

DÉJEAN H., GAUSSIER E. (2002). Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica, Alignement lexical dans les corpus multilingues.*

DÉJEAN H., SADAT F., GAUSSIER E. (2002). An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Actes, COLING'02*, p. 218–224.

FUNG P. (1998). A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-parallel Corpora. In D. FARWELL, L. GERBER & E. HOVY, Eds., *Actes, AMTA'98*, p. 1–16: Springer.

GAUSSIER E., LANGÉ J.-M. (1995). Modèles statistiques pour l'extraction de lexiques bilingues. *Traitement automatique des langues*, Vol. 36(1-2), p. 133–155.

MELAMED I. D. (2001). *Empirical Methods for Exploiting Parallel Texts.* MIT Press.

RAPP R. (1999). Automatic Identification of Word Translations from Unrelated English and German Corpora. In *Actes, ACL'99*, p. 519–526.

VERONIS J. (2000). *Parallel Text Processing.* Kluwer Academic Publishers.