# Apport des termes complexes à l'acquisition lexicale multilingue à partir de corpus comparables spécialisés : entre intuition et réalité

Emmanuel Morin

# Apport des termes complexes à l'acquisition lexicale multilingue à partir de corpus comparables spécialisés : entre intuition et réalité

Emmanuel Morin

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

**Résumé** : En acquisition lexicale multilingue à partir de corpus comparables, les méthodes traditionnelles pour identifier un terme et sa traduction s'appuient sur des contextes lexicaux composés de mots ou termes simples. Or en domaine spécialisé, les termes complexes sont tout autant précis, moins ambigus et reflètent mieux la spécificité du domaine que les termes simples. Leur prise en compte dans les contextes lexicaux devrait donc permettre d'améliorer la précision des méthodes d'alignement. Cette intuition ne s'avérant pas correcte, nous présentons dans cet article une série d'expériences visant à en préciser les causes.

## 1 Introduction

Depuis les travaux fondateurs de Rapp (1995); Tanaka & Iwasaki (1996); Fung & McKeown (1997), l'acquisition de lexiques bilingues à partir de corpus comparables a connu un essor important. Cet intérêt pour l'exploitation de corpus comparables est principalement lié aux difficultés de disposer de corpus parallèles lorsqu'il s'agit d'exploiter un matériau textuel ne faisant pas intervenir l'anglais. En outre, les lexiques bilingues obtenus à partir de corpus parallèles sont quelque peu biaisés. En effet, un corpus parallèle étant constitué d'un texte dans une langue source et de sa traduction dans une langue cible, le vocabulaire rencontré dans la partie traduite est fortement influencé par celui de la langue source en particulier dans les domaines spécialisés.

La méthode originelle pour l'acquisition de lexiques bilingues à partir de corpus comparables — qualifiée de « méthode par traduction directe » par Déjean & Gaussier (2002) — repose sur la recherche dans chaque partie monolingue d'« affinités du premier ordre » : « *First-order affinities describe what other*

*words are likely to be found in the immediate vicinity of a given word*[1] » (Grefenstette, 1994, p. 279). Ainsi à chaque unité lexicale à traduire sont associées les unités lexicales de son voisinage immédiat. Celles-ci forment alors un « vecteur de contexte ». L'unité lexicale à traduire est ensuite transférée vers la langue cible par traduction de l'ensemble des éléments de son vecteur de contexte. Les traductions potentielles de cette unité seront alors celles dont les vecteurs de contexte en langue cible sont les plus semblables au vecteur de contexte traduit.

Avec cette méthode, 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 chinois 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. En ce qui concerne, l'alignement de groupes nominaux relevant du domaine général, une première approche a été proposée pour des termes anglais/japonais par Shahzad *et al.* (1999) où l'évaluation est réduite à une liste de dix termes. Une deuxième approche plus aboutie, qui s'appuie sur l'algorithme EM (*Expectation-Maximization*), a été développée par Cao & Li (2002) pour l'extraction de groupes nominaux bilingues anglais/chinois. Ces derniers obtiennent une précision de 91 % sur les 3 premiers candidats en exploitant le web. Ces résultats, qui sont globalement inférieurs à ceux obtenus avec un corpus parallèle, permettent néanmoins d'envisager des applications concrètes notamment en recherche d'informations interlangues.

En ce qui concerne les domaines spécialisés, les résultats obtenus pour des couples de termes simples sont moins significatifs. Déjean & Gaussier (2002) obtiennent pour les 10 et 20 premiers candidats 44 % et 57 % de précision pour un corpus médical anglo-allemand de 100 000 mots et 35 % et 42 % pour un corpus anglo-allemand relatif aux sciences sociales de 8 millions de mots. Chiao & Zweigenbaum (2002) obtiennent, quant à eux, pour un corpus médical français-anglais de 1,2 million de mots : 61% et 94% de précision pour les 10 et 20 meilleurs candidats. En domaine spécialisé, le premier obstacle est bien sûr lié à la difficulté de disposer de corpus aussi volumineux que pour la langue générale. Ce qui ne permet pas de construire des vecteurs de contexte fortement nourris. Le second s'explique par l'appauvrissement du vecteur de contexte d'une unité à traduire lors de son transfert en langue cible si un trop grand nombre d'éléments ne peuvent être traduits.

Dans l'ensemble des travaux utilisant la méthode directe pour l'alignement de mots ou termes simples, Déjean & Gaussier (2002) sont les seuls à prendre en compte les termes complexes dans le processus d'alignement en s'appuyant sur les entrées multi-mots du dictionnaire bilingue avec des résultats contrastés. Or en domaine spécialisé, les termes complexes sont tout autant précis, moins ambigus et reflètent mieux la spécificité du domaine que les termes simples. Leur prise en compte dans les contextes lexicaux devrait donc permettre d'améliorer la précision des méthodes d'alignement. Les expériences que nous avons réalisées en

[1] « *Les affinités du premier ordre décrivent les mots qui sont susceptibles d'être trouvés dans le voisinage immédiat d'un mot donné.* »

ce sens semblent malheureusement indiquer que cette intuition n'est pas fondée. De manière à en préciser les causes, nous commençons par présenter en section 2 la plateforme de fouille terminologique multilingue mettant en œuvre la méthode directe et prenant en compte les termes complexes dans le calcul des contextes lexicaux. Nous présentons ensuite, en section 3, les résultats comparatifs obtenus en intégrant ou non les termes complexes dans le calcul des contextes lexicaux et une série d'expériences visant à préciser leur rôle dans le processus d'alignement. Enfin, la section 4 dresse le bilan de ce travail.

# 2 Plateforme de fouille terminologique multilingue

La plateforme de fouille terminologique multilingue que nous avons développée permet à partir d'un corpus comparable en deux langues d'associer une liste de termes simples et complexes à leurs traductions candidates. Celle-ci extrait en premier à l'aide d'une méthode linguistique les termes complexes et leurs variations dans chacune des langues, puis tente de les aligner à l'aide d'une méthode statistique[2].

## 2.1 Traitement linguistique

L'étape de traitements linguistiques vise à identifier les termes complexes dans chaque partie monolingue du corpus comparable. Dans un premier temps, les documents de chaque langue subissent un ensemble de pré-traitements : i) nettoyage des données indésirables (tableaux, caractères de contrôle... ), ii) segmentation du texte en occurrences de formes et de phrases et iii) assignation aux formes de leur étiquette grammaticale et de leur lemme. Ensuite, l'identification des termes complexes est réalisée à l'aide de l'extracteur de terminologie *ACABIT* (Daille, 2003). 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. À ce niveau, les candidats termes qui correspondent à des variantes morphologiques dérivationnelles synonymiques sont fusionnés par *ACABIT*. Par exemple, les candidats termes *produit de la forêt* et *produit forestier* n'en forment plus qu'un. 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. Au terme de cette première phase, les termes complexes identifiés par *ACABIT* et qui ne sont pas des hapax sont considérés comme une seule unité textuelle, dans le cas contraire ils sont décomposés en mots simples.

## 2.2 Traitement statistique

Le processus d'alignement lexical qui vise à fournir pour un terme à traduire une liste ordonnée de traductions candidates est réalisé en adoptant la « méthode

[2]Pour une présentation détaillée de la plateforme, nous renvoyons les lecteurs à l'article de Morin & Daille (2006).

par traduction directe » (Fung, 1998; Peters & Picchi, 1998; Rapp, 1999). Elle se résume aux traitements suivants :

**Identification des contextes lexicaux** Pour chaque langue du corpus comparable, le contexte de chaque unité lexicale[3] $i$ est extrait en repérant les unités qui apparaissent autour de $i$ dans une fenêtre contextuelle de $n$ mots[4]. 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* ou *Taux de vraisemblance*. À chaque unité $i$, nous associons ainsi un vecteur de contexte.

**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 à l'aide 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, un nombre plus ou moins grand 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.

**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* ou *Jaccard*.

**Obtention des traductions candidates** En fonction des précédentes valeurs de similarité, nous obtenons une liste ordonnée de traductions candidates pour l'unité visée.

## 2.3 Ressources associées

Dans le cadre de cette étude, les ressources associées à la plateforme sont composées d'un corpus comparable, d'un dictionnaire bilingue et d'un lexique de référence (pour évaluer la qualité des traductions obtenues).

Le corpus comparable, noté [SYLV], a été constitué à partir de la revue *Unasylva*[5] publiée chaque trimestre en anglais, espagnol et français depuis 1947 par

[3]Nous employons ici le terme *unité lexicale* pour désigner la forme lemmatisée d'un mot, d'un terme simple ou d'un terme complexe. Par abus de langage, le terme *unité simple* fait référence à un mot ou un terme simple et *unité complexe* à un terme complexe.

[4]Les mots vides et les hapax ne sont pas exploités dans le processus d'alignement.

[5]`http://www.fao.org/forestry/site/unasylva/`

la FAO. 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.

Le dictionnaire bilingue, nécessaire au processus d'alignement, a été construit à partir de ressources disponibles sur le web. Afin de suppléer l'insuffisance du dictionnaire bilingue pour la traduction des mots composés, nous en avons élargi la couverture en utilisant une approche par traduction compositionnelle (Grefenstette, 1999). Ainsi, pour chaque mot composé de la langue source absent du dictionnaire et identifié par *ACABIT*[6], nous traduisons chacun de ses composants. Ensuite, nous combinons entre elles les traductions des différents composants pour obtenir des traductions candidates du mot composé en langue cible. Les traductions retenues sont celles qui font référence à un mot composé identifié en langue cible par *ACABIT*.

Le lexique de référence a été construit à partir de différentes ressources terminologiques : i) le glossaire bilingue de la terminologie de la sylviculture au Canada du service canadien des forêts[7], ii) le lexique multilingue du projet Eurosilvasur (plate-forme ressource forêt-bois-papier des régions de l'Europe du Sud)[8] et iii) le thesaurus multilingue AGROVOC de la FAO[9]. À partir de ces ressources, nous avons constitué un lexique de référence, noté [lexique TS], composé de 100 termes simples français (où chaque terme français est au moins présent cinq fois dans la partie française du corpus comparable) dont la traduction qui n'est pas présente dans notre dictionnaire bilingue est un terme simple. Ce lexique est essentiellement composé de termes peu fréquents dans la mesure où, d'une part, les différentes ressources utilisées pour le créer proposent des termes très spécifiques ou très génériques, d'autre part, le corpus utilisé couvre un grand nombre de domaines liés à la foresterie et ne constitue pas une ressource très spécialisée.

## 3 Expériences réalisées

Afin de vérifier l'apport des termes complexes au processus d'alignement, nous avons réalisé deux expériences en nous appuyant sur le [lexique TS] associé au corpus [SYLV] pour la méthode directe. D'un côté, les vecteurs de contexte sont limités à des unités simples, et de l'autre, les vecteurs de contexte sont composés d'unités simples et complexes.

[6] À ce niveau, nous exploitons aussi les variantes du terme si la forme canonique ne produit pas de traduction.

[7] `http://nfdp.ccfm.org/silviterm/silvi_f/silvitermintrof.htm`

[8] `http://www.eurosilvasur.net/francais/lexique.php`

[9] `http://www.fao.org/agrovoc/`

La table 1 présente les résultats de ces deux expériences où nous indiquons, pour chaque configuration, le nombre de traductions trouvées ($NB_{trad}$) ainsi que le nombre de traductions trouvées dans les dix et vingt meilleures positions ($TOP_{10}$ et $TOP_{20}$).

| | $NB_{trad}$ | $TOP_{10}$ | $TOP_{20}$ |
|---|---|---|---|
| Unités simples | 62 | 43 | 47 |
| Unités simples et complexes | 50 | 33 | 39 |

TAB. 1 – Apport des unités complexes à l'alignement

Les résultats obtenus en intégrant les termes complexes dans la construction des vecteurs de contexte sont globalement inférieurs à ceux décelés sans eux. Il est néanmoins délicat de comparer directement les résultats de ces deux expériences dans la mesure où nous ne travaillons plus dans le même espace vectoriel. En effet, l'intégration des termes complexes dans les vecteurs de contexte augmente le nombre de dimensions de l'espace vectoriel tout en diminuant leur représentativité numérique. Par exemple, le terme simple *débardage*, issu de la partie française du corpus [SYLV], passe d'une fréquence de 544 dans un espace vectoriel limité aux unités simples à 144 dans une espace intégrant des unités simples et complexes puisqu'il apparaît dans différents termes complexes comme *débardage mécanique*, *piste de débardage*, *technique de débardage*...

## 3.1 Hypothèses vérifiées

Afin de mieux comprendre la nature des résultats obtenus en intégrant les termes complexes au contexte lexical et aussi d'affiner notre intuition initiale, nous avons cherché à vérifier un ensemble d'hypothèses que nous restituons ici.

**Hypothèse 1** « *Les termes complexes ne sont pas suffisamment présents dans les vecteurs de contexte.* »

Pour vérifier cette première hypothèse, nous avons mesuré la proportion d'unités simples par rapport aux unités complexes dans les vecteurs de contexte en nous limitant aux 100 premières entrées des vecteurs de contexte du [lexique TS]. Cette proportion est d'environ 50 % en langue source comme en langue cible. En outre, le nombre d'unités pivots de notre corpus (c'est-à-dire celles qui sont traduites par le dictionnaire) est de 7 352 unités simples pivots et de 6 769 unités complexes pivots sur un nombre total de 55 013 unités simples et complexes. Il apparaît donc que cette hypothèse est non fondée puisque les unités complexes sont aussi nombreuses que les unités simples dans les vecteurs de contexte.

**Hypothèse 2** « *Les termes complexes ne sont pas "porteurs de sens".* »

Le fait que les termes complexes soient suffisamment présents dans les vecteurs de contexte ne permet pas de juger de leur potentiel de discrimination par

rapport aux unités simples. Afin de vérifier cette hypothèse, nous avons manuellement inspecté les vecteurs de contexte des termes du [lexique TS] dans le cas d'une construction limitée à des unités simples, puis d'une construction intégrant des unités simples et complexes. D'une manière générale, les vecteurs de contexte construits à partir d'unités simples et complexes sont plus précis que ceux construits uniquement à partir d'unités simples. Ceci est essentiellement dû à une meilleure représentation du contexte lexical des unités. Par exemple, le vecteur de contexte du terme *débardage* comporte le terme complexe *tracteur à chenille* qui est plus discriminant que les mots pleins le composant (cf. colonnes 1 et 2 de la table 2). Les termes complexes sont bien des éléments « porteurs de sens ». Ils semblent donc pertinents dans la construction des vecteurs.

| *Unités simples* | *Unités simples et complexes* | *Unités complexes* |
|---|---|---|
| tracteur | groupement | tracteur à chenille |
| câble | abattage | coût de le abattage |
| distance | chargement | traction animal |
| abattage | transport | utilisation de le traction |
| groupement | tracteur à chenille | équipement mécanique |
| arche | coût de le abattage | progrès ultérieur |
| montée | distance | progrès réalisé |
| chenille | manuel sur le méthode | homme au brélage |
| coût | tronçonnage | homme dirigeant |
| piste | tracteur | énergie économique |

TAB. 2 – Premières entrées ordonnées du vecteur de contexte de *débardage*

Afin de conforter cette affirmation, nous avons réalisé une nouvelle expérience où nous limitons la description des vecteurs de contexte aux seules unités complexes. D'une manière générale, les éléments des vecteurs de contexte ainsi construits ont un « rapport sémantique » plus ou moins proche et ne semblent pas correspondre à une distribution aléatoire (cf. colonne 3 de la table 2). À noter que certains termes complexes sont spécifiques à la problématique du débardage comme *homme au brélage* ou encore partiellement identifiés comme *homme dirigeant*. Les résultats de cette expérience, présentés en table 3, indiquent que seulement 12 traductions correctes ont été identifiées à partir du [lexique TS] parmi les 100 premières traductions candidates. Les résultats de traduction obtenus ici — qui doivent être interprétés avec prudence dans la mesure où les unités complexes sont en moyenne moins fréquentes que les unités simples — sont bien en dessous de ceux décelés précédemment (cf. table 1). Dans cette expérience, nous avons observé que peu de termes complexes des vecteurs de contexte de la langue source sont transférés en langue cible. Nous devons donc déterminer l'importance de ce phénomène dans le processus de traduction.

**Hypothèse 3** « *Les termes complexes des vecteurs de contexte de la langue source ne sont pas correctement transférés en langue cible.* »

| | $NB_{trad}$ | $TOP_{10}$ | $TOP_{20}$ |
|---|---|---|---|
| Unités complexes | 12 | 6 | 8 |

TAB. 3 – Apport des unités complexes à l'alignement en se limitant aux unités complexes

Pour évaluer la quantité d'unités complexes des vecteurs de contexte de la langue source transférée en langue cible, nous avons considéré les 100 premières entrées des vecteurs de contexte associées au [lexique TS] (dans cette expérience les vecteurs de contexte sont décrits par des unités simples et complexes). Nous constatons ainsi que la proportion de termes complexes dans les vecteurs de contexte traduits passe de 50 % à environ 20 %. Plus de la moitié des termes complexes ne sont pas conservés pendant la phase de transfert. Comparativement, 92 % des unités simples des vecteurs de contexte de la langue source sont conservées en langue cible. En ce qui concerne la qualité des termes complexes traduits, nous ne l'avons pas jugé particulièrement mauvaise, quoique limitée le plus souvent aux traductions compositionnelles ajoutées au dictionnaire. Dans l'ensemble, il ne semble pas y avoir de décalage sémantique fortement marqué. Les difficultés de traduction des termes complexes semblent essentiellement liées à l'insuffisance de dictionnaire. Par exemple, le terme *tracteur à chenille* du vecteur de contexte du terme *débardage* ne peut être traduit directement à partir de notre dictionnaire bilingue, son transfert en langue cible repose sur la traduction compositionnelle de ses composants. Dans ce cas, puisque notre dictionnaire bilingue propose *tractor* pour *tracteur* et *caterpillar* pour *chenille*, la seule traduction obtenue est *caterpillar tractor*. Or cette traduction ne correspond pas à celles usitées le plus souvent en langue cible, à savoir *crawler* et *crawler tractor*. Nous sommes confrontés ici à deux problèmes importants, d'une part, un terme ne se traduit pas systématiquement par un terme de même longueur (le terme complexe *tracteur à chenille* est traduit par le mot simple *crawler*) , et d'autre part, la traduction d'un terme complexe ne s'obtient pas simplement par la traduction de ses composants (*tracteur à chenille* est aussi traduit par *crawler tractor*, où *crawler* n'est pas la traduction de *chenille*).

Il semble donc que les difficultés de transfert des termes complexes soient une explication à leur manque d'apport dans l'alignement. Néanmoins, nous devons préciser si ces 20 % de termes complexes transférés de la langue source à la langue cible participent ou non à l'identification des traductions candidates.

**Hypothèse 4** « *Le nombre de termes communs entre le vecteur original et le vecteur traduit est trop faible.* »

Pour vérifier cette dernière hypothèse, nous avons compté le nombre de termes communs obtenus lors de la comparaison du vecteur traduit avec les vecteurs originaux de la langue cible. À partir des 100 premières entrées des vecteurs de contexte du [lexique TS], nous avons trouvé en moyenne 14,7 termes communs en

ne considérant que la meilleure traduction candidate (celle qui a le plus de termes communs). Si nous considérons toutes les traductions candidates, nous obtenons une moyenne de 8,1 termes communs sur 100. Sachant que nous avons environ 20 % de termes complexes dans les vecteurs traduits, nous ne pouvons espérer trouver plus de deux ou trois termes complexes communs entre les vecteurs source et cible. Ceci semble être une explication cohérente à l'inefficacité des termes complexes dans les vecteurs de contexte. Pour compléter cette analyse, nous devrions aussi vérifier que la traduction effective des termes complexes engendre bien un gain dans la précision des résultats. Si cette condition, plus délicate à vérifier, n'était pas avérée, cela engendrerait plutôt une remise en cause du processus de sélection des termes complexes que du processus d'alignement lequel est intrinsèquement lié à la capacité du dictionnaire à assurer un pont entre les langues cible et source.

### 3.2 Discussion

Ces différentes expériences, qui permettent de mieux comprendre le faible apport des termes complexes à l'alignement, induisent une difficulté supplémentaire lorsque l'on s'intéresse à l'alignement de termes complexes. Dans ce cas, les unités lexicales associées aux vecteurs de contexte sont tout autant des unités simples que complexes. Pour « contourner cet obstacle » et limiter la construction des vecteurs de contexte à des unités simples, deux approches sont possibles. Prenons par exemple la phrase suivante : « *La forêt boréale perdrait ainsi 37 pour cent de sa superficie [...]* » et le terme complexe *forêt boréale* à traduire. Une première « *approche amont* » consiste à s'appuyer sur l'identification en corpus du terme complexe pour construire son vecteur de contexte, qui serait alors limité aux unités *perdre* et *superficie*, puis à associer aux autres unités des contextes composés d'unités simples (p. ex. pour *superficie* : *forêt*, *boréal* et *perdre*). Une deuxième « *approche aval* » consiste pour chaque unité simple à lui associer uniquement les unités simples de son contexte (p. ex. pour *forêt* : *boréal*, *perdre*, *superficie* ; pour *boréal* : *forêt*, *perdre*, *superficie* ; etc.), puis à construire le vecteur de contexte d'un terme complexe à traduire comme étant la conjonction des vecteurs de contexte des unités simples le composant (ce qui correspond à la technique utilisée par Déjean & Gaussier (2002)). Dans les deux cas, le vecteur de contexte associé à *forêt boréale* est globalement le même (à quelques occurrences liées au décalage de la fenêtre contextuelle). De cette manière, il est possible d'utiliser la méthode directe pour l'alignement de termes complexes.

## 4 Conclusion

Dans cet article, nous avons cherché à rendre compte de l'intérêt mais aussi de la difficulté d'intégration des unités complexes dans le calcul des contextes lexicaux. D'une part, la prise en compte des termes complexes induit bien une meilleure représentation du contexte lexical. D'autre part, l'augmentation du nombre de dimensions de l'espace vectoriel associée aux difficultés de transfert

des unités complexes de la langue source à la langue cible engendrent un appauvrissement du vecteur de contexte traduit en langue cible. Par conséquent, il devient plus délicat d'identifier en langue cible les vecteurs de contexte similaires au vecteur de contexte traduit.

Ces résultats ne doivent pas conduire les travaux en fouille terminologique multilingue à partir de corpus comparables à limiter systématiquement le calcul des contextes lexicaux aux unités simples. Cette dualité, entre une meilleure représentation du contexte lexical par rapport à un appauvrissement du vecteur de contexte traduit, doit ouvrir de nouvelles perspectives de recherche en vue de trouver le meilleur compromis. Lequel est non seulement intéressant pour l'alignement de termes simples, mais plus encore, indispensable à l'alignement de termes complexes.

## Références


Cao Y. & Li H. (2002). Base Noun Phrase Translation Using Web Data and the EM Algorithm. In *Proceedings of COLING'02*, p. 127–133, Tapei, Taiwan.

Chiao Y.-C. & Zweigenbaum P. (2002). Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of COLING'02*, p. 1208–1212, Tapei, Taiwan.

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

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*, p. 1–22.

Fung P. (1998). A Statistical View on Bilingual Lexicon Extraction : From Parallel Corpora to Non-parallel Corpora. In *Proceedings of AMTA'98*, p. 1–16, Langhorne, PA, USA.

Fung P. & McKeown K. (1997). Finding Terminology Translations from Non-parallel Corpora. In *Proceedings of VLC'97*, p. 192–202, Hong Kong, China.

Grefenstette G. (1994). Corpus-Derived First, Second and Third-Order Word Affinities. In *Proceedings of EURALEX'94*, p. 279–290, Amsterdam, Pays-Bas.

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

Morin E. & Daille B. (2006). Comparabilité de corpus et fouille terminologique multilingue. *TAL*, **47**(2), 113–136.

Peters C. & Picchi E. (1998). Cross-language information retrieval : A system for comparable corpus querying. In G. Grefenstette, Ed., *Cross-language information retrieval*, p. 81–90. Kluwer Academic Publishers.

Rapp R. (1995). Identify Word Translations in Non-Parallel Texts. In *Proceedings of ACL'95*, p. 320–322, Boston, MA, USA.

Rapp R. (1999). Automatic Identification of Word Translations from Unrelated English and German Corpora. In *Proceedings of ACL'99*, p. 519–526, College Park, MD, USA.

Shahzad I., Ohtake K., Masuyama S. & Yamamoto K. (1999). Identifying Translations of Compound Nouns Using Non-aligned Corpora. In *Proceedings of MAL'99*, p. 108–113, Beijing, China.

Tanaka K. & Iwasaki H. (1996). Extraction of Lexical Translations from Non-Aligned Corpora. In *Proceedings of ACL'96*, p. 580–585, Copenhagen, Denmark.