Apport d'un corpus comparable déséquilibré à l'extraction de lexiques bilingues

Emmanuel Morin
Université de Nantes, LINA - UMR CNRS 6241
2 rue de la Houssinière, BP 92208, 44322 Nantes Cedex 03
emmanuel.morin@univ-nantes.fr

Résumé. Les principaux travaux en extraction de lexiques bilingues à partir de corpus comparables reposent sur l'hypothèse implicite que ces corpus sont équilibrés. Cependant, les différentes méthodes computationnelles associées sont relativement insensibles à la taille de chaque partie du corpus. Dans ce contexte, nous étudions l'influence que peut avoir un corpus comparable déséquilibré sur la qualité des terminologies bilingues extraites à travers différentes expériences. Nos résultats montrent que sous certaines conditions l'utilisation d'un corpus comparable déséquilibré peut engendrer un gain significatif dans la qualité des lexiques extraits.

Abstract. The main work in bilingual lexicon extraction from comparable corpora is based on the implicit hypothesis that corpora are balanced. However, the different related approaches are relatively insensitive to sizes of each part of the comparable corpus. Within this context, we study the influence of unbalanced comparable corpora on the quality of bilingual terminology extraction through different experiments. Our results show the conditions under which the use of an unbalanced comparable corpus can induce a significant gain in the quality of extracted lexicons.

Mots-clés : Multilinguisme, corpus comparable, extraction de lexiques bilingues.

Keywords: Multilingualism, comparable corpus, bilingual lexicon extraction.

1 Contexte

L'extraction de lexiques bilingues à partir de corpus a initialement été entreprise en considérant des corpus parallèles (*i.e.* des textes en correspondance de traduction). Cependant, et en dépit des bons résultats obtenus avec ces corpus, ces derniers demeurent des ressources rares, notamment pour les domaines spécialisés et pour des couples de langues ne faisant pas intervenir l'anglais. Pour ces différentes raisons, les recherches en extraction de lexiques bilingues se sont penchées sur une autre sorte de corpus bilingue composé de documents partageant différentes caractéristiques telles que le domaine, le genre, la période... sans être en correspondance de traduction. Ces corpus, *bien connus maintenant sous le nom de corpus comparables*, ont été originellement qualifiés de *corpus non parallèles* (*non-parallel corpora* dans Fung (1995) et Rapp (1995)) ou encore de *corpus non alignés* (*non-aligned corpora* dans Tanaka & Iwasaki (1996)). Cette filiation entre les corpus parallèles et comparables a visiblement conduit à

travailler implicitement avec des corpus comparables équilibrés (*i.e.* des corpus comportant la même quantité de données en langues source et cible).

En extraction de lexiques bilingues à partir de corpus comparables, deux directions de recherche peuvent être observées. La première exploite des corpus comparables volumineux de langue générale pour extraire des couples de traductions. Ainsi, Fung & McKeown (1997) utilisent un corpus journalistique anglais-japonais composé de 49 millions d'octets issus du *Wall Street Journal* et de 127 millions d'octets issus du *Nikkei Financial News* (équivalent à environ 60 millions d'octets en anglais en raison du codage des caractères japonais), tandis que Rapp (1999) s'appuie sur un corpus journalistique allemand-anglais composé de 135 millions de mots extraits du *Frankfurter Allgemeine Zeitung* et de 163 millions de mots extraits du *Guardian*. Dans les deux cas, les corpus comparables exploités sont relativement bien équilibrés puisque le ratio de données entre les deux langues est de 1,2. La seconde direction s'intéresse quant à elle à l'exploitation de corpus comparables de domaines spécialisés qui sont par nature de taille plus réduite. Ainsi, Chiao & Zweigenbaum (2002b) construisent un corpus médical par consultation de CISMeF pour la partie française (591 594 mots) et de CliniWeb pour la partie anglaise (608 320 mots), Déjean & Gaussier (2002) constituent un corpus allemand-anglais composé d'environ 700 résumés d'articles scientifiques issus de MEDLINE (chaque partie contenant environ 100 000 mots), Morin *et al.* (2007) sélectionnent des documents français et japonais sur le web pour construire un corpus médical restreint à la thématique du diabète et de l'alimentation (693 666 mots pour la partie française et 807 287 mots pour la partie japonaise). Ici encore, les corpus comparables sont bien équilibrés avec un ratio de données compris entre 1 et 1,2. Lorsque l'on se place dans la perspective de travailler avec un corpus comparable équilibré, la quantité de données de la partie du corpus la plus importante doit être ramenée à la plus faible. Cet aspect est préjudiciable pour des corpus comparables spécialisés, notamment lorsqu'ils comportent des données en langue anglaise pour laquelle de nombreux documents sont souvent disponibles en raison de sa position dominante comme standard international pour les communications scientifiques.

La méthode par traduction directe associée à ces travaux n'impose pas de contraintes sur la taille de chaque partie du corpus comparable. Chaque partie est vue comme un ensemble clos sur lequel des statistiques de distributions lexicales basées sur le contexte des mots sont calculées. En fait, plus nombreux seront les contextes lexicaux, plus fiables seront les associations entre mots. En ce sens, un corpus comparable équilibré n'est pas requis pour l'extraction de lexiques bilingues.

Aucune attention, à notre connaissance, n'a été portée sur l'utilisation de corpus comparables déséquilibrés pour l'extraction de lexiques bilingues. Cet article vise précisément cet objectif en se concentrant sur les corpus comparables spécialisés.

2 Méthode par traduction directe

Les principaux travaux en extraction de lexiques bilingues à partir de corpus comparables reposent sur la simple observation qu'un mot et sa traduction ont tendance à apparaître dans les mêmes environnements lexicaux. La mise en œuvre de cette observation repose en premier lieu sur l'identification d'*affinités du premier ordre* dans les langues source et cible : « *First-order affinities describe what other words are likely to be found in the immediate vicinity of a given*

word[1] » (Grefenstette, 1994a, p. 279). Ces affinités peuvent être représentées sous la forme de vecteurs de contexte où chaque élément du vecteur est un mot qui apparaît dans la fenêtre du mot à traduire. Par exemple une fenêtre de sept mots (3 mots et avant et après le mot visé) permet d'identifier des dépendances syntaxiques. Afin d'identifier les mots les plus significatifs des vecteurs de contexte et de réduire l'influence des mots fréquents, les vecteurs de contexte sont normalisés au moyen d'une mesure d'association. Ensuite, la traduction d'un mot est obtenue en comparant son vecteur de contexte, dont les éléments ont été préalablement traduits à l'aide d'un dictionnaire bilingue, à l'ensemble des vecteurs de la langue cible.

Notre implémentation de la méthode par traduction directe se décompose de la manière suivante (Rapp, 1995; Fung & McKeown, 1997) :

Identification des contextes lexicaux

Pour chaque partie du corpus comparable, le contexte de chaque mot i est extrait en repérant les mots qui apparaissent autour de lui dans une fenêtre contextuelle de n mots[2]. Afin d'identifier les mots caractéristiques des contextes lexicaux et de supprimer l'effet induit par la fréquence des mots, nous normalisons l'association entre les mots sur la base d'une mesure de récurrence contextuelle comme l'*Information Mutuelle* - IM (Fano, 1961) ou le *Taux de vraisemblance* - TV (Dunning, 1993) (cf. les équations 1 et 2 et la table 1). Après normalisation, à chaque élément j du vecteur de contexte du mot i nous attachons le taux d'association $assoc(i, j)$.

Transfert d'un mot à traduire

Le transfert d'un mot 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 moyen d'un dictionnaire bilingue. Si le dictionnaire propose plusieurs traductions pour un élément, nous ajoutons au vecteur de contexte de k l'ensemble des traductions proposées[3] (lesquelles sont pondérées par la fréquence de la traduction en langue cible). Dans le cas où l'élément n'est pas présent dans le dictionnaire, il ne sera pas exploité dans le processus de traduction.

Identification des vecteurs proches du mot à traduire

Le vecteur de contexte v_k ainsi traduit est ensuite comparé à l'ensemble des vecteurs de la langue cible en s'appuyant sur une mesure de distance vectorielle comme *Cosinus* (Salton & Lesk, 1968) ou *Jaccard pondéré* (Grefenstette, 1994b) (cf. les équations 3 et 4).

Obtention des traductions candidates

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

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

TAB. 1 – Table de contingence

$$IM(i, j) = log\frac{a}{(a + b)(a + c)} \quad (1)$$

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

[2] En ce qui concerne les mots agrammaticaux, ils ne sont pas exploités dans le processus d'alignement.

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

$$\begin{aligned} TV(i,j) = \; & a\log(a) + b\log(b) + c\log(c) + d\log(d) \\ & +(a+b+c+d)\log(a+b+c+d) - (a+b)\log(a+b) \\ & -(a+c)\log(a+c) - (b+d)\log(b+d) - (c+d)\log(c+d) \end{aligned} \quad (2)$$

$$Cosinus_{v_l}^{v_k} = \frac{\sum_t assoc_t^l \, assoc_t^k}{\sqrt{\sum_t assoc_t^{l\,2}}\sqrt{\sum_t assoc_t^{k\,2}}} \quad (3)$$

$$Jaccard\,pondéré_{v_l}^{v_k} = \frac{\sum_t min(assoc_t^l, assoc_t^k)}{\sum_t max(assoc_t^l, assoc_t^k)} \quad (4)$$

En fonction de l'adéquation du dictionnaire bilingue avec le corpus d'étude, un nombre plus ou moins important d'éléments du vecteur de contexte sera traduit. Le mot à traduire sera d'autant plus informatif en langue cible que le nombre d'éléments traduits de son vecteur de contexte sera important. Les difficultés de traduction des éléments des vecteurs de contexte peuvent être partiellement contournées : i) en combinant un dictionnaire de langue générale avec un dictionnaire spécialisé ou un thesaurus multilingue (Chiao & Zweigenbaum, 2003; Déjean *et al.*, 2002) ou ii) en utilisant la méthode par similarité interlangue (Déjean & Gaussier, 2002; Daille & Morin, 2005).

3 Expériences et résultats

3.1 Ressources linguistiques

Dans le cadre de cette étude, nous avons construit un corpus comparable spécialisé français-anglais à partir de documents extraits du portail Elsevier[4]. L'ensemble des documents collectés relève du domaine médical restreint à la thématique du « cancer du sein ». Nous avons utilisé l'interface de recherche du portail pour sélectionner les publications scientifiques comportant dans le titre ou les mots clés le terme *cancer du sein* en français et *breast cancer* en anglais pour la période de 2001 à 2008. Nous avons ainsi automatiquement collecté 130 documents (environ 530 000 mots) pour le français et 1 640 documents (environ 7,4 millions de mots) pour l'anglais. Les documents anglais ont été divisés en 14 parties contenant chacune environ 530 000 mots. L'ensemble des documents ont été nettoyés et normalisés à travers les traitements suivants : segmentation en occurrences de formes, étiquetage morpho-syntaxique et lemmatisation. Enfin, les mots agrammaticaux ont été supprimés et les mots apparaissant moins de deux fois dans la partie française et dans chaque partie anglaise écartés. Le tableau 2 présente les principales caractéristiques des différentes parties après ces étapes à savoir le nombre de documents (# documents) et le nombre de mots distincts (# mots distincts). Dans la suite de cet article, nous désignons par [corpus i] le corpus comparable construit en s'appuyant sur la partie française et la partie anglaise i ($i \in [1, 14]$).

Le dictionnaire français-anglais nécessaire à l'étape de traduction de la méthode directe a été construit à partir de différentes ressources disponibles sur le web. Il comporte, après normalisation, 22 300 mots pour le français avec en moyenne 1,6 traductions par entrée. Il s'agit d'un

[4] www.elsevier.com

	# documents	# mots distincts
Français		
Partie 1	130	7 376
Anglais		
Partie 1	118	8 214
Partie 2	114	7 788
Partie 3	101	8 370
Partie 4	114	7 992
Partie 5	119	7 958
Partie 6	117	8 230
Partie 7	109	8 035
Partie 8	116	8 008
Partie 9	129	8 334
Partie 10	114	7 978
Partie 11	126	8 373
Partie 12	137	8 065
Partie 13	123	7 847
Partie 14	103	8 457

TAB. 2 – Caractéristiques des corpus collectés

dictionnaire de langue générale qui ne contient que peu de termes en rapport avec le domaine médical.

En extraction de lexiques bilingues à partir de corpus comparables spécialisés, le lexique de référence nécessaire pour évaluer les performances de la méthode d'alignement est souvent composé d'une centaine de mots (180 dans Déjean & Gaussier (2002), 95 dans Chiao & Zweigenbaum (2002a), ou encore 100 mots simples dans Daille & Morin (2005)). Afin de construire notre lexique de référence, nous avons sélectionné 400 couples de mots simples français-anglais à partir du meta-thesaurus UMLS[5] et du *Grand dictionnaire terminologique*[6]. Nous n'avons ensuite retenu que les couples pour lesquels le mot français apparaît au moins cinq fois dans la partie française et sa traduction au moins cinq fois dans chaque partie anglaise i. Au terme de ce processus de sélection, nous disposons d'une liste de référence composée de 122 couples de termes simples français-anglais.

3.2 Expériences préliminaires

Afin d'évaluer l'influence d'un corpus comparable déséquilibré sur l'extraction de lexiques bilingues différentes expériences sont réalisées. Dans un premier temps, nous utilisons chaque [corpus i] équilibré indépendamment des autres. Puis, nous faisons varier la taille de la partie anglaise du corpus comparable de 530 000 à 7,4 millions de mots suivant un pas de 530 000 mots. Dans ces différentes expériences, ainsi que dans toutes celles qui vont suivre, la taille n de la fenêtre contextuelle est fixée à 7 mots (3 mots et avant et après le mot visé) et les paramètres indiqués (mesures d'association et de similarité) sont ceux qui donnent en moyenne

[5] www.nlm.nih.gov/research/umls
[6] www.granddictionnaire.com

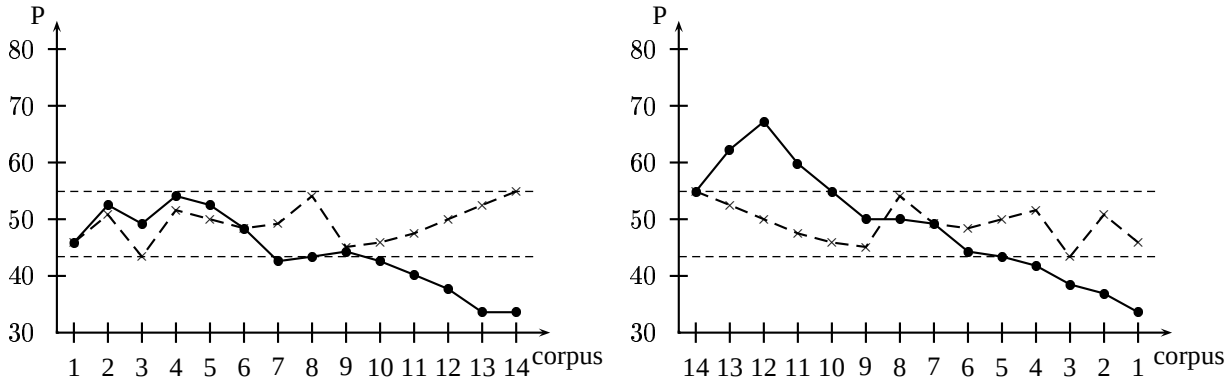

FIG. 1 – Précision des traductions obtenues pour la direction français-anglais (paramètres : IM et cosinus)

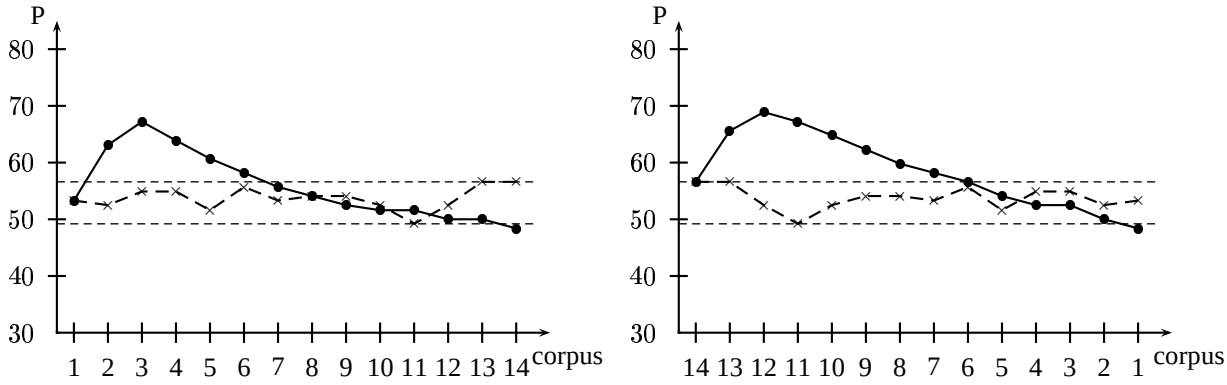

FIG. 2 – Précision des traductions obtenues pour la direction anglais-français (paramètres : TV et jaccard pondéré)

les meilleurs résultats sur les corpus comparables équilibrés.

Les figures[7] 1 et 2 présentent les résultats de ces expériences évaluées au niveau de la précision du top 20 (c'est-à-dire que la bonne traduction est identifiée dans la liste ordonnée des vingt premières traductions candidates) pour les termes de la liste de référence suivant les directions français-anglais et anglais-français et pour deux configurations distinctes du corpus comparable déséquilibré. Par exemple, la valeur 3 sur l'axe des abscisses du graphique gauche de la figure 1 indique pour la direction français-anglais la précision obtenue pour un corpus comparable composé i) seulement du [corpus 3] (43,4 % - $\times$) et ii) des [corpus 1, 2 et 3] (49,2 % - $\bullet$).

Comme remarque préliminaire, nous pouvons constater que les résultats varient sensiblement suivant les corpus comparables utilisés individuellement. Ce phénomène est plus visible quand la langue source est le français (variation entre 43,4 et 54,9 %) que l'anglais (variation entre 49,2 et 56,6 %). En outre, pour n'importe quel corpus comparable équilibré les résultats obtenus avec l'anglais comme langue source sont toujours supérieurs à ceux obtenus avec le français. Globalement, nos résultats sont conformes à ceux relevés dans la littérature avec la méthode par traduction directe pour des corpus comparables ayant une taille et un domaine semblables. À titre de comparaison, Chiao & Zweigenbaum (2002a) affichent une précision autour de 60 % pour le top 20 en exploitant un corpus comparable français-anglais de 1,2 millions de mots et une liste de référence composée de 95 mots (chaque terme français ayant une fréquence

[7] Dans toutes les figures présentées dans cet article, les courbes en pointillés désignent un corpus comparable équilibré et les courbes pleines un corpus comparable déséquilibré.

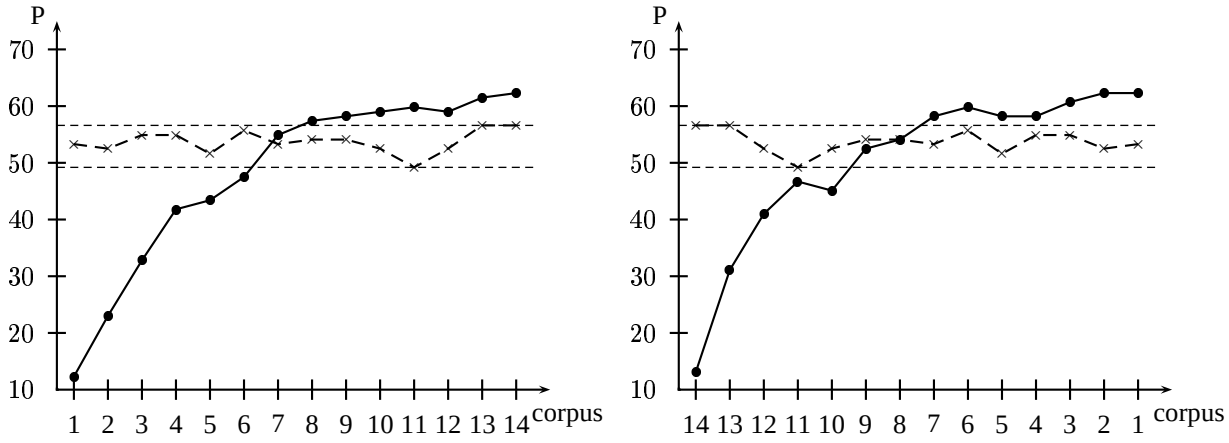

FIG. 3 – Précision des traductions obtenues pour la direction anglais-français (paramètres pour les courbes en pointillés : IM en langue source, IM3 en langue cible et jaccard pondéré)

supérieure à 100 occurrences dans ce travail).

Quand le français est la langue source et que nous faisons varier la taille du corpus comparable, il est difficile de discerner une tendance commune entre les deux expériences de la figure 1. Le seul phénomène vraiment visible est une chute importante de la précision quand le corpus devient fortement déséquilibré.

Maintenant quand l'anglais est la langue source, nous pouvons observer la même tendance entre les deux graphiques de la figure 2. Ainsi lorsque le corpus comparable est faiblement déséquilibré (ratio entre 2:1 et 6:1), la précision obtenue est toujours supérieure à celle de n'importe quel corpus comparable équilibré. La précision pouvant être améliorée jusqu'à 11 points lorsque le ratio est de 3:1 par comparaison avec la meilleure précision obtenue avec les corpus comparables équilibrés. D'un autre côté, nous observons la même chute de précision que précédemment, bien que moins importante, lorsque le corpus comparable devient fortement déséquilibré (ratio supérieur à 8:1).

3.3 Combinaison de mesures

Dans les expériences précédentes, les résultats obtenus lorsque le corpus comparable est fortement déséquilibré sont relativement surprenants. En fait, nous appliquons la même stratégie (*i.e.* les mêmes mesures d'association) pour les langues source et cible quelle que soit la taille de la partie anglaise du corpus comparable. Or lorsque la taille de la partie anglaise augmente, l'association entre les mots des vecteurs de contexte est plus fiable. Ainsi, une mesure d'association inadaptée pour une partie anglaise de taille réduite peut se révéler pertinente pour une taille plus grande.

Nous avons exploré cette hypothèse en utilisant une mesure d'association distincte entre les langues source et cible. Les résultats les plus intéressants ont été obtenus en associant l'IM3 (information mutuelle au cube[8]) (Daille, 1995) pour la partie française et l'IM pour la partie anglaise. La figure 3 présente les résultats obtenus pour cette configuration pour la direction anglais-français. Comme nous pouvons le constater, les résultats que nous obtenons s'amé-

[8]L'IM3 est similaire à la mesure d'IM (cf. équation 1), mais avec le numérateur élevé à la puissance 3. Alors que l'IM a tendance à surdimensionner les mots rares, l'IM3 réduit cette tendance.

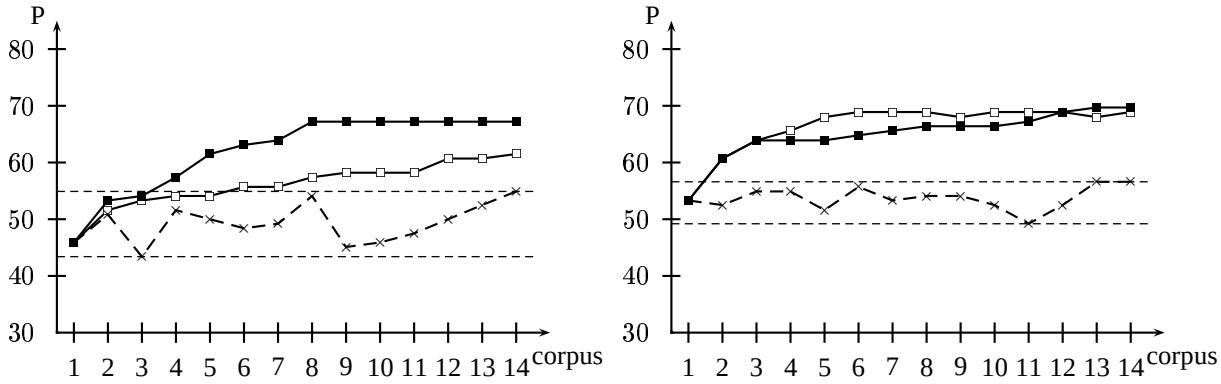

FIG. 4 – Précision des traductions obtenues par combinaison de résultats pour les directions français-anglais à gauche et anglais-français à droite (moyennes arithmétique □ et harmonique ■)

liorent en fonction de la taille de la partie anglaise du corpus comparable. En outre, lorsque le ratio entre la partie anglaise et la partie française devient supérieur à 8:1, la précision obtenue est supérieure à n'importe quel corpus comparable équilibré. La précision est améliorée au mieux de 6 points par comparaison à la meilleure précision obtenue avec les corpus comparables équilibrés.

3.4 Combinaison de résultats

Comme cela est rappelé dans l'article de synthèse sur les corpus comparables de Zweigenbaum & Habert (2006, p. 36) : « *Lorsque l'on dispose de plusieurs méthodes pour résoudre un problème, il est souvent plus productif de chercher à les combiner* ». Puisque nous disposons de plusieurs corpus comparables équilibrés qui indiquent des variations dans les résultats, nous pouvons appliquer ce propos à notre travail en combinant les résultats obtenus pour ces différents corpus. Cette technique correspond finalement à une autre manière d'exploiter un corpus comparable déséquilibré. Nous avons donc exploré cette voie à travers différentes combinaisons de résultats. Les résultats les plus intéressants ont été obtenus en utilisant pour un mot à traduire i) la moyenne arithmétique des scores de similarité des traductions candidates et ii) la moyenne harmonique des positions des traductions candidates relevés pour les différents corpus comparables équilibrés.

La figure 4 présente les résultats obtenus avec ces mesures pour les deux directions de traduction. Nous pouvons constater que la combinaison des résultats obtenus pour chaque corpus équilibré permet d'améliorer significativement la qualité des traductions identifiées pour les deux directions de traduction. Lorsque le français est la langue source et que le ratio de données entre les deux langues devient supérieur à 3:1, la précision obtenue est supérieure au meilleur résultat relevé avec les corpus équilibrés avec la moyenne harmonique. En outre, à partir du ratio 8:1 et pour la même mesure, la précision atteint son maximum (67,2 %), soit un gain de 11 points par rapport au meilleur résultat obtenu avec les corpus équilibrés. En ce qui concerne la seconde direction de traduction, les résultats sont toujours supérieurs à ceux relevés avec les corpus équilibrés utilisés individuellement. Ici, la précision atteint au mieux 69,7 %, soit un gain de 13 points par rapport au meilleur résultat obtenu avec les corpus équilibrés.

4 Conclusion

Les travaux fondateurs en extraction de lexiques bilingues à partir de corpus comparables ont été entrepris pour suppléer aux difficultés rencontrées avec les corpus parallèles, notamment en ce qui concerne leur disponibilité. Partant de ce constat, nous avons émis l'hypothèse que la filiation entre ces deux sortes de corpus bilingues avait visiblement conduit à travailler avec des corpus équilibrés. En domaines spécialisés, pour lesquels les données disponibles sont moins volumineuses, le respect de l'équilibre du corpus implique de ramener la taille de la partie du corpus la plus importante à la plus faible. Cette stratégie ne permet pas de tirer profit de la plus grande masse de données disponible, notamment lorsque le corpus comporte des données en langue anglaise pour laquelle de nombreux documents spécialisés sont accessibles.

Nous avons donc cherché dans cet article à étudier l'influence que pouvait avoir un corpus comparable spécialisé déséquilibré sur la qualité des lexiques bilingues extraits. Les différentes expériences que nous avons réalisées montrent sous quelles conditions un corpus comparable déséquilibré améliore la précision des traductions. Ainsi, lorsque la langue source est celle qui dispose de la plus grande quantité de données, un corpus faiblement déséquilibré (ratio entre 2:1 et 6:1) permet d'obtenir un gain significatif qui peut atteindre 11 points ; et un corpus fortement déséquilibré (ratio supérieur à 8:1) un gain de 6 points. En outre, une autre stratégie d'exploitation d'un corpus déséquilibré consiste à combiner les résultats obtenus pour différents corpus équilibrés. Dans ce cas, le gain peut atteindre 13 points.

D'autres expériences seront nécessaires pour confirmer ces premiers résultats. Néanmoins, nous espérons que ce travail ouvrira la voie à l'utilisation de corpus comparables spécialisés déséquilibrés pour l'extraction de lexiques bilingues.

Remerciements

Ce travail qui s'inscrit dans le cadre du projet METRICC (`www.metricc.com`) a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-08-CORD-009.

Références

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

CHIAO Y.-C. & ZWEIGENBAUM P. (2002b). Looking for French-English Translations in Comparable Medical Corpora. *Journal of the American Society for Information Science*, **8**, 150–154.

CHIAO Y.-C. & ZWEIGENBAUM P. (2003). The Effect of a General Lexicon in Corpus-Based Identification of French-English Medical Word Translations. In R. BAUD, M. FIESCHI, P. LE BEUX & P. RUCH, Eds., *The New Navigators: from Professionals to Patients, Actes Medical Informatics Europe*, volume 95 of *Studies in Health Technology and Informatics*, p. 397–402, Amsterdam: IOS Press.

DAILLE B. (1995). *Combined approach for terminology extraction: lexical statistics and linguistic filtering*. Rapport interne 5, Unit for Computer Research on the English (UCREL), Lancaster University.

DAILLE B. & MORIN E. (2005). French-English Terminology Extraction from Comparable Corpora. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCLNP'05)*, p. 707–718, Jeju Island, Korea.

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.

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

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

FANO R. M. (1961). *Transmission of Information: A Statistical Theory of Communications*. Cambridge, MA, USA: MIT Press.

FUNG P. (1995). Compiling Bilingual Lexicon Entries from a non-Parallel English-Chinese Corpus. In *Proceedings of the 3rd Annual Workshop on Very Large Corpora (VLC'95)*, p. 173–183, Cambridge, MA, USA.

FUNG P. & MCKEOWN K. (1997). Finding Terminology Translations from Non-parallel Corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora (VLC'97)*, p. 192–202, Hong Kong.

GREFENSTETTE G. (1994a). Corpus-Derived First, Second and Third-Order Word Affinities. In *Proceedings of the 6th Congress of the European Association for Lexicography (EURALEX'94)*, p. 279–290, Amsterdam, The Netherlands.

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

MORIN E., DAILLE B., TAKEUCHI K. & KAGEURA K. (2007). Bilingual Terminology Mining – Using Brain, not brawn comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, p. 664–671, Prague, Czech Republic.

RAPP R. (1995). Identify Word Translations in Non-Parallel Texts. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (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 the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, p. 519–526, College Park, MD, USA.

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

TANAKA K. & IWASAKI H. (1996). Extraction of Lexical Translations from Non-Aligned Corpora. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, volume 2, p. 580–585, Copenhagen, Denmark.

ZWEIGENBAUM P. & HABERT B. (2006). Faire se rencontrer les parallèles : regards croisés sur l'acquisition lexicale monolingue et multilingue. *Revue de sociolinguistique en ligne GLOTTOPOL*, **8**, 22–44.