Compositionnalité et contextes issus de corpus comparables pour la traduction terminologique

Emmanuel Morin Béatrice Daille
Université de Nantes, LINA UMR CNRS 6241
2, rue de la Houssinière, BP 92208
F-44322 Nantes cedex 03
{emmanuel.morin,beatrice.daille}@univ-nantes.fr

RÉSUMÉ

Dans cet article, nous cherchons à mettre en correspondance de traduction des termes extraits de chaque partie monolingue d'un corpus comparable. Notre objectif concerne l'identification et la traduction de termes spécialisés. Pour ce faire, nous mettons en œuvre une approche compositionnelle dopée avec des informations contextuelles issues du corpus comparable. Notre évaluation montre que cette approche améliore significativement l'approche compositionnelle de base pour la traduction de termes complexes extraits de corpus comparables.

ABSTRACT

Compositionality and Context for Bilingual Lexicon Extraction from Comparable Corpora

In this article, we study the possibilities of improving the alignment of equivalent terms monolingually acquired from bilingual comparable corpora. Our overall objective is to identify and to translate highly specialised terminology. We applied a compositional approach enhanced with pre-processed context information. Our evaluation demonstrates that our alignment method outperforms the compositional approach for translationally equivalent term discovery from comparable corpora.

MOTS-CLÉS : Corpus comparable, compositionnalité, information contextuelle, lexique bilingue.

KEYWORDS: Comparable Corpora, compositionality, context information, bilingual lexicon.

1 Introduction

L'alignement lexical à partir de corpus comparables s'intéresse tout particulièrement aux domaines de spécialités, en particulier relevant de domaines scientifiques. Les domaines de spécialités sont caractérisés par des ressources textuelles réduites en comparaison à la langue générale et par une grande proportion de vocabulaire spécifique qui n'est pas présent dans les dictionnaires monolingues ou bilingues de langue générale. Un vocabulaire relevant d'un domaine de spécialité recense des termes simples, *i.e.* des mots simples, ou des termes complexes, *i.e.* des composés syntagmatiques, ces derniers étant particulièrement productifs (Sag *et al.*, 2002). Un terme est l'expression d'un concept dans un domaine de spécialité : par exemple, dans le domaine médical, *cancer* est un terme simple et *cancer du sein* un terme complexe.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 141–154,
Grenoble, 4 au 8 juin 2012.

Les corpus comparables qui rassemblent : « *des documents textuels dans des langues différentes qui ne sont pas des traductions les uns des autres* [1] » (Bowker et Pearson, 2002, p. 93) apparaissent comme une solution viable pour résoudre le manque de ressources linguistiques des domaines de spécialités : d'une part, leur statut de production monolingue garantit la qualité de leur vocabulaire spécialisé et, d'autre part, l'aspect multilingue du web garantit une disponibilité de ressources textuelles dans un grand nombre de langues et pour de nombreux domaines de spécialités. La comparabilité du corpus doit bien entendu être vérifiée à l'aide de caractéristiques communes partagées par les différentes langues lors de sa compilation (McEnery et Xiao, 2007). Pour les domaines de spécialités, le domaine et le sous-domaine sont des caractéristiques partagées obligatoires tout comme l'intention communicative et le genre textuel de manière à obtenir des traductions fiables (Bowker et Pearson, 2002). Cette comparabilité peut aussi être évaluée de manière quantitative en termes de degré de comparabilité et d'homogénéité du corpus comme dans Li et Gaussier (2010).

Pour construire des lexiques relevant de domaines de spécialités, les termes sont extraits de chaque partie monolingue du corpus comparable. Pour collecter les mêmes termes dans deux langues différentes, il est important d'utiliser un programme d'extraction terminologique adoptant la même méthode dans les deux langues. Les termes complexes constituant environ 80 % d'un lexique de domaine de spécialité, comme constaté par Nakagawa et Mori (2003) pour le japonais, il est essentiel de pouvoir les traduire.

Notre objectif est d'identifier pour un terme complexe dans une langue, sa bonne traduction au sein d'un ensemble de termes complexes candidats dans une autre langue. Une méthode triviale consiste à traduire chacun des éléments du terme complexe à l'aide d'un dictionnaire bilingue de la langue générale, à générer la combinatoire de toutes les traductions trouvées dans le dictionnaire, puis de ne conserver que les termes complexes apparaissant soit dans la liste des termes complexes candidats (Morin et Daille, 2010), soit dans le corpus comparable (Robitaille *et al.*, 2006), ou encore directement sur le web (Grefenstette, 1999). Cette méthode ne fonctionne que pour les termes complexes partageant une sémantique compositionnelle : Baldwin et Tanaka (2004) ont constaté que c'était le cas de 48,7 % des termes complexes de structure N N pour la paire de langue anglais/japonais.

Dans cet article, nous proposons d'améliorer cette méthode fondée sur une traduction compositionnelle par l'utilisation de contextes extraits d'un corpus comparable. Ces informations contextuelles seront utilisées lorsqu'un ou plusieurs éléments du terme complexe à traduire n'apparaîtront pas dans le dictionnaire bilingue. Nous démontrons que l'utilisation du contexte permet de produire un nombre important de traductions correctes pour des termes complexes ne pouvant pas être traduit par la méthode compositionnelle.

Dans la suite de cet article, nous présentons en section 2 les problèmes de traductions rencontrés avec les termes complexes. La section 3 détaille la méthode fondée sur une traduction compositionnelle pour l'obtention de traductions de termes complexes. Notre nouvelle approche associant traduction compositionnelle et contextes issus de corpus comparables est introduite en section 4. La section 5 décrit les différentes ressources textuelles et dictionnairiques utilisées pour nos expériences. La section 6 évalue l'impact de notre approche mixte sur la qualité des lexiques bilingues ainsi obtenus. La section 7 examine quelques travaux similaires à l'approche proposée avant de conclure.

1. « *sets of texts in different languages, that are not translations of each other* ».

2 Traduction des termes complexes

Si les termes complexes sont moins polysémiques (Savary et Jacquemin, 2003) et plus représentatifs (Nomura et M., 1989; Nakagawa et Mori, 2003) d'un domaine de spécialité que les termes simples, le repérage de leurs traductions pose un certain nombre de difficultés comme la fertilité, la non compositionnalité ou encore la variation terminologique [2] :

Fertilité Elle correspond à un problème de différence de longueur entre le terme complexe de la langue source et celui de la langue cible (Brown *et al.*, 1993). Par exemple, le terme complexe français *dépistage du cancer du sein* (trois mots pleins) est traduit en anglais par le terme complexe *breast screening* (deux mots pleins).

Non compositionnalité Elle s'exprime lorsqu'un terme complexe de la langue cible n'est pas typiquement composé de la traduction des parties du terme de la langue source (Melamed, 2001). Par exemple, le terme complexe français *curage axillaire* est traduit en anglais par le terme *axillary dissection* où le mot anglais *dissection* n'est pas la traduction du mot français *curage*.

Variation terminologique Cela fait référence à un terme complexe qui apparaît dans des documents sous différentes formes reflétant des différences morphologiques, syntaxiques ou sémantiques. Par exemple, les termes complexes français *cancer du sein* et *cancer mammaire* sont traduits en anglais par le même terme complexe *breast cancer*. Les termes complexes source et cible peuvent apparaître dans différentes structures syntaxiques. Ansi, le terme complexe français *prolifération tumorale* de structure N Adj est traduit en anglais par le terme complexe *tumour proliferation* de structure N N où l'adjectif français *tumorale* est lié par dérivation morphologique à la traduction française du nom anglais *tumour*. La variation terminologique peut aussi impliquer une variation paradigmatique quand un élément du terme complexe est remplacé par un synonyme ou un hyperonyme tel que *tumour size* → *diameter tumour* en anglais et non en français *taille tumorale*. Ce dernier type de variante n'est généralement pas traité par les programmes d'extraction terminologique.

Il est assez difficile de concevoir un cadre général qui puisse répondre à l'ensemble de ces problèmes (Robitaille *et al.*, 2006). Par exemple, le problème de fertilité doit probablement être réglé en premier pour éviter des alignements incomplets entre les termes complexes des langues source et cible. En ce qui concerne le problème de variation terminologique, celui-ci pourrait être en partie résolu lors de l'extraction terminologie monolingue par le regroupement de toutes les variantes morphologiques et syntaxiques du terme complexe dans les langues source et cible. Un terme complexe est ainsi vu comme un ensemble de séquences de termes reflétant une forme de base ou une variante. Ce regroupement peut être interprété comme une normalisation terminologique de la même manière que la lemmatisation au niveau morphologique.

3 Approche compositionnelle

La compositionnalité est définie comme la propriété où « *le sens du tout est fonction du sens des parties* [3] » (Keenan et Faltz, 1985, p. 24-25) : une *poêle à frire* est en effet une *poêle* destinée à *frire*.

2. Les exemples de termes français et anglais sont extraits d'un corpus comparable médical décrit en section 5.
3. « *the meaning of the whole is a function of the meaning of the parts* ».

La mise en œuvre du principe de traduction compositionnelle à partir de corpus comparables repose sur les étapes suivantes (Grefenstette, 1999; Tanaka, 2002; Robitaille *et al.*, 2006) :

Traduction du terme complexe de la langue source Pour un terme complexe de la langue source à traduire, chaque mot plein composant le terme complexe est traduit à l'aide d'un dictionnaire bilingue. Généralement, lors de cette phase de projection, la catégorie grammaticale des composants du terme complexe n'est pas utilisée. Par exemple pour le terme complexe français *examen clinique*, nous avons six traductions en anglais pour *examen* (*consideration*/N, *examen*/N, *examination*/N, *inspection*/N, *review*/N et *test*/N) et deux traductions pour *clinique* (*clinic*/N et *clinical*/Adj).

Génération des traductions candidates Toutes les combinaisons sont générées sans tenir compte de l'ordre des mots avec un total de $O(n! \prod_{i=1}^{p} t_i)$ combinaisons possibles (où t_i est le nombre de traductions du mot plein i et n le nombre total de mots pleins). 24 combinaisons sont obtenues avec le précédent exemple.

Sélection des traductions candidates À partir de l'ensemble des traductions candidates, les traductions les plus probables sont ordonnées en fonction de leur fréquence d'apparition dans la langue cible. Pour l'exemple précédent, les traductions candidates sont les termes complexes de la langue cible identifiés par le système d'extraction de terminologie.

Le principe de traduction compositionnelle est restrictif. Pour palier cette difficulté, Robitaille et al. (2006) proposent d'utiliser une méthode de repli : s'il n'y a pas suffisamment de données dans le dictionnaire bilingue pour traduire un terme de longueur n (avec $n > 2$ mots pleins) alors ce terme sera décomposé en toutes les combinaisons de termes de longueur inférieure ou égale à n. Cette approche permet de pouvoir traduire directement une sous partie du terme complexe s'il est présent dans le dictionnaire bilingue. Par exemple, pour le terme complexe français *technique du ganglion sentinelle* quatre combinaisons seraient générées : (i) [technique du ganglion sentinelle], (ii) [technique du ganglion] [sentinelle], (iii) [technique] [ganglion sentinelle] et (iv) [technique] [ganglion] [sentinelle]. Morin et Daille (2010) ont proposé quant à eux une méthode compositionnelle étendue qui comble le fossé entre termes complexes de différentes structures syntaxiques en exploitant des liens morphologiques. En s'appuyant sur une liste de règles morphologiques de codage/décodage associée à un système d'extraction de terminologie, leur méthode est plus efficace que la méthode compositionnelle de base. Pour 859 termes complexes français de structure N Adjr (où Adjr est un adjectif relationnel), ils retrouvent dans un dictionnaire bilingue français/japonais 30 termes complexes et traduisent 8 termes complexes avec une précision de 62 % avec la méthode compositionnelle de base et 128 termes complexes avec une précision de 88 % avec leur approche compositionnelle étendue.

L'approche compositionnelle est aussi appelée « approche par sacs de mots équivalents [4] » par Vintar (2010) lorsque le dictionnaire bilingue est construit à partir d'un corpus parallèle et qu'il contient tous les mots qui apparaissent dans le corpus avec leurs équivalences de traduction accompagnées d'un score de probabilité. Le nombre de traductions généré peut être réduit en utilisant des structures syntaxiques de traductions entre les termes des langues source et cible. Par exemple, Tanaka et Baldwin (2003) utilisent les structures suivantes pour filtrer les candidats de traduction : un terme complexe japonais de structure N_1 N_2 est traduit en anglais par un terme complexe de structure N_1 N_2 (dans 33,2 % des cas), par Adj_1 N_2 (28,4 %) ou encore par N_2 of (the) N_1 (4,4 %).

4. « *bag-of-equivalents approach* ».

4 Approche compositionnelle enrichie par des informations contextuelles

L'approche compositionnelle de base qui propose des traductions pour des termes complexes est facile à mettre en œuvre, mais elle échoue lorsque :
– les termes complexes ne partagent de propriété compositionnelle, *i.e.* 50 % des situations (Baldwin et Tanaka, 2004) ;
– l'un des composants du terme complexe ne peut être traduit directement par un dictionnaire bilingue ;
– les combinaisons proposées de termes candidats ne sont pas présentes dans la liste des termes complexes extraits de la langue cible ou plus généralement dans la langue cible du corpus comparable.

Pour palier cette difficulté, une première solution serait de trouver des synonymes dans la langue source. Pour les mots de basse fréquence, Pekar et al. (2006) prédisent des valeurs de cooccurrences absentes en s'appuyant sur des mots similaires dans la même langue. Pour les traductions jugées plus difficiles, Sharoff et al. (2009) identifient des mots similaires dans la langue source pour produire une similarité plus fiable. Dans notre cas, nous avons déjà réalisé un regroupement de synonymes en exploitant un ensemble de variantes du terme au lieu d'un terme unique (voir la section 5).

L'approche proposée pour identifier les traductions d'un terme complexe à partir d'un corpus comparable s'appuie sur l'exploitation du contexte des composants du terme à traduire lorsque l'approche compositionnelle de base échoue. Nous référons ici aux deux parties monolingues du corpus comparables comme le corpus source et cible. Ce modèle se décompose en quatre étapes :

Calcul du contexte des termes complexes Pour un terme complexe à traduire défini par $C_{s1}C_{s2}\cdots C_{sk}$ (où k est le nombre de mots pleins du terme), nous recherchons chaque composant C_{si} dans le dictionnaire bilingue. Ici, chaque composant non traduit par le dictionnaire est remplacé par des informations de cooccurrence. Plus précisément, nous calculons les mots qui cooccurrent avec C_{si} dans une fenêtre de w mots autour de C_{si} dans le corpus source. L'information mutuelle comme le rapport de vraisemblance sont deux bonnes mesures pour déterminer la relation de cooccurrence entre deux mots. Ces informations de cooccurrence, normalisées avec l'une des précédentes mesures, sont représentées sous la forme d'un vecteur de contexte (V_{si}). À titre d'exemple, considérons le terme complexe français *antécédent familial* ($C_{s1}C_{s2}$). Si le premier composant *antécédent* (C_{s1}) n'est pas présent dans le dictionnaire bilingue alors ce composant est remplacé par son vecteur de contexte (V_{s1}) (voir la figure 1).

Transfert des termes complexes Pour chaque vecteur de contexte V_{si}, ses éléments sont projetés dans la langue cible en utilisant le dictionnaire bilingue et le vecteur de contexte transféré devient V'_{si}. Si le dictionnaire bilingue propose plusieurs traductions pour un élément, elles sont toutes utilisées, mais chaque traduction est pondérée en fonction de la fréquence de l'élément dans la langue cible. Si un élément n'est pas trouvé dans le dictionnaire bilingue, il est alors écarté. En revanche, quand le composant C_{si} est présent dans le dictionnaire, nous calculons les informations de cooccurrence de chaque traduction dans le corpus cible et les sauvegardons dans un vecteur de contexte, V'_{si}. Par exemple, si nous avons trouvé deux traductions anglaises pour le composant *familial* (C_{s2}) telles que *fami-*

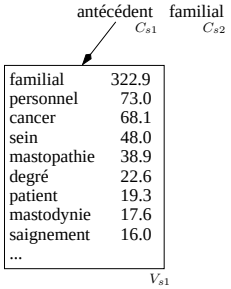

FIGURE 1 – Calcul du contexte d'un terme complexe

lial et *family*, nous retenons alors deux vecteurs de contexte $V_{s2_1}^{'}$ et $V_{s2_2}^{'}$ dans le corpus cible (voir la figure 2).

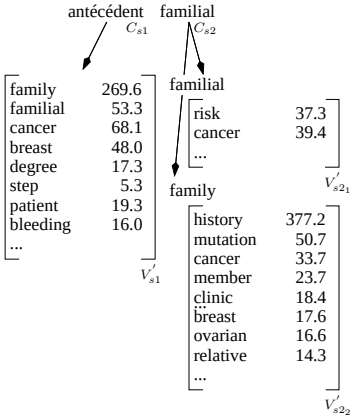

FIGURE 2 – Projection dans la langue cible

Génération des traductions candidates Chaque terme complexe de la langue cible, pour lequel chaque composant C_{ti} est décrit par son vecteur de contexte V_{ti}, est ensuite comparé au vecteur de contexte transféré à travers une mesure de distance vectorielle comme le Cosinus ou le Jaccard pondéré. Pour un terme complexe en langue cible composé de deux vecteurs de contexte V_{t1} et V_{t2} et un terme complexe transféré composé de deux vecteurs de contexte $V_{s1}^{'}$ et $V_{s2}^{'}$, deux paires de scores de similarité correspondantes aux différentes combinaisons possibles seraient calculées : $sim(V_{t1}, V_{s1}^{'})$ avec $sim(V_{t2}, V_{s2}^{'})$ et $sim(V_{t1}, V_{s2}^{'})$ avec $sim(V_{t2}, V_{s1}^{'})$. Le score terminal pour chaque paire est quant à lui défini comme la moyenne géométrique de chaque score de similarité : $\sqrt{sim(V_{t1}, V_{s1}^{'}).sim(V_{t2}, V_{s2}^{'})}$ et

$\sqrt{sim(V_{t1}, V_{s2}^{'}).sim(V_{t2}, V_{s1}^{'})}$ (voir la figure 3).

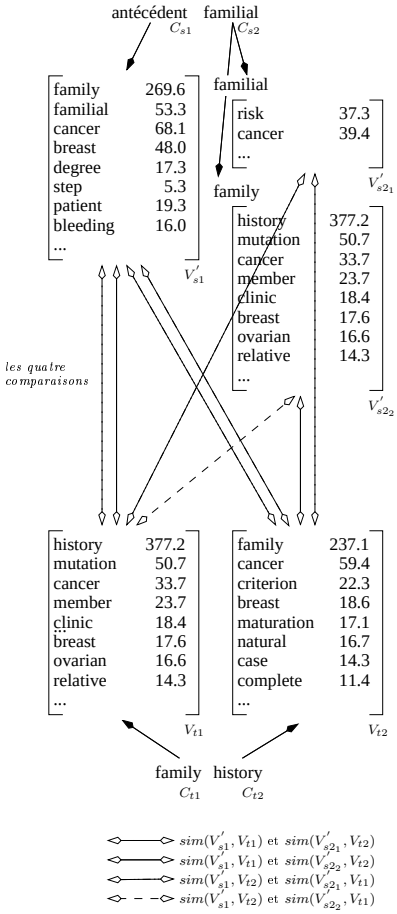

FIGURE 3 – Comparaison entre un terme complexe à traduire et un terme complexe de la langue cible (une paire de flèches correspond à une comparaison entre les composants du terme source et du terme cible - par exemple la paire de flèches pleines correspond à la comparaison entre $V_{s1}^{'}$ et V_{t1} et entre $V_{s2_2}^{'}$ et V_{t2})

Sélection des traductions candidates Les traductions candidates sont finalement ordonnées en fonction du score d'association (voir la figure 4).

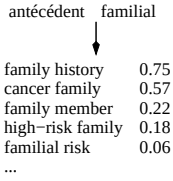

FIGURE 4 – Liste ordonnée des traductions candidates

5 Ressources

Dans cette section, nous décrivons les différentes ressources utilisées pour nos expériences, à savoir le corpus comparable, le dictionnaire bilingue et la liste de référence.

5.1 Corpus comparable

Le corpus comparable spécialisé français/anglais utilisé dans cette étude relève du domaine médical et plus précisément du sous-domaine du cancer du sein. Les documents composant ce corpus ont été sélectionnés automatiquement à partir du site d'Elsevier [5] en sélectionnant les articles scientifiques publiés sur la période 2001 et 2008 pour lesquels le titre ou les mots-clés des articles contiennent les termes « *cancer du sein* » en français et « *breast cancer* » en anglais. La compilation de ce corpus comparable remplit les exigences d'un corpus comparable de spécialité en termes de domaine, sous-domaine, de paramètres de communication (experts-à-experts) et de genre textuel qui sont des caractéristiques communes à travers les langues. Nous avons ainsi automatiquement collecté 130 documents pour le français et 118 pour l'anglais ce qui représente environ 530 000 mots par langue. L'ensemble des documents ont été nettoyés et normalisés à travers les traitements suivants : segmentation en occurrences de formes, étiquetage morpho-syntaxique [6], lemmatisation [7] et extraction terminologique [8]. Enfin, les mots outils et les hapax ont été supprimés dans les parties française et anglaise. Le corpus comparable fournit finalement 4 000 mots simples et 5 100 mots composés en français et 4 000 mots simples et 4 100 mots composés en anglais.

5.2 Dictionnaire bilingue

Le dictionnaire français-anglais nécessaire à l'étape de transfert 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 dictionnaire de langue générale qui ne contient que peu de termes en rapport avec le domaine médical.

5. `http://www.elsevier.com`
6. Pour le français et l'anglais, nous avons utilisé l'étiqueteur de Brill (Brill, 1994).
7. Pour le français, nous avons utilisé le lemmatiseur FLEMM : `http://www.univ-nancy2.fr/pers/namer/` et pour l'anglais un lemmatiseur construit à partir de la base CELEX.
8. Nous avons choisi d'utiliser ACABIT (Daille, 2003), un outil ouvert qui permet de traiter des corpus volumineux et dont la conception est fondamentalement multilingue, avec des implémentations pour le français et l'anglais.

5.3 Liste de référence

La liste de référence contient une liste de termes complexes extraits automatiquement dans l'une des parties monolingues du corpus comparable. Cette liste est utilisée pour comparer la couverture et la précision des trois différentes approches : la simple projection dans un dictionnaire bilingue, l'approche compositionnelle et l'approche compositionnelle enrichie par des informations contextuelles. Les unités terminologiques qui sont extraites par ACABIT sont des termes complexes dont les structures syntaxiques correspondent soit à un structure canonique soit à structure de variation. Les structures sont exprimées en utilisant des étiquettes syntaxiques [9]. Pour le français les principales structures sont N N, N Prep N et N Adj et pour l'anglais N N, Adj N et N Prep N. Les variantes prises en compte sont morphologiques et syntaxiques pour les deux langues. ACABIT considère comme une variante morphologique la modification morphologique de l'un des composants de la forme de base, comme une variante syntaxique l'insertion d'un autre mot dans les composants de la forme de base. Par exemple, dans la partie française du corpus comparable le terme candidat *cancer du sein* apparaît sous les formes suivantes :

- **forme de base** de structure N Prep N : *cancer du sein* ;
- **variante flexionnelle** : *cancers du sein* ;
- **variante syntaxique** (par insertion d'un modifieur dans la forme de base) : *cancer primitif du sein* ;
- **variante syntaxique** (par expansion par coordination de la forme de base) : *cancer des ovaires et du sein*.

Pour la liste de référence, nous avons sélectionné les termes complexes français extraits par ACABIT ayant un nombre d'occurrences supérieur ou égal à 5. Cette liste de référence est composée de 976 termes complexes français. Dans cette liste, 90 % des candidats termes fournis par le processus d'extraction terminologique après regroupement sont composés de deux mots pleins.

6 Expériences et résultats

Dans cette section, nous évaluons les performances des différentes approches en fonction de la qualité des traductions obtenues.

6.1 Projection du dictionnaire et approche compositionnelle

Dans un premier temps, nous commençons par compter le nombre de termes qui peuvent être directement traduits par le dictionnaire bilingue. Nous évaluons ensuite la qualité des traductions fournies par l'approche compositionnelle. La table 1 présente les résultats obtenus pour la traduction du français vers l'anglais. La première colonne indique le nombre de termes complexes français qui sont traduits. Puisque l'approche compositionnelle, comme la traduction dictionnairique, peut donner plusieurs traductions pour un terme à traduire, la colonne suivante indique le nombre de termes complexes français pour lesquels une ou plusieurs traductions sont obtenues en anglais. La troisième colonne indique quant à elle le nombre de bonnes traductions en anglais. Enfin, les deux dernières colonnes indiquent la précision aux rangs 1 (Top_1) et 5 (Top_5)

9. Les symboles sont Adj (Adjectif), N (Nom), Prep (Préposition).

pour chaque stratégie. La précision au rang n (Top_n) représente le nombre de traductions correctes trouvées dans la liste ordonnée des n premières traductions candidates. Ici, les traductions candidates sont ordonnées selon leur fréquence d'apparition dans la partie anglaise du corpus comparable. Les résultats de cette première expérience montrent que sur les 976 termes complexes de la liste de référence, 51 termes complexes sont présents dans le dictionnaire et 140 termes complexes sont traduits au moyen de l'approche compositionnelle avec une précision de 79,1 % pour le Top_5 (les termes complexes présents dans le dictionnaire ne sont pas utilisés par l'approche compositionnelle). Ici, nous sommes incapables de proposer une traduction pour 785 termes complexes de la liste de référence.

	# termes français	# termes anglais	# traductions correctes	Top_1	Top_5
projection du dictionnaire	51	69	69	100 %	100 %
approche compositionnelle	140	172	136	73,2 %	79,1 %

TABLE 1 – Projection du dictionnaire et approche compositionnelle

6.2 Approche compositionnelle enrichie par des informations contextuelles

Nous appliquons maintenant l'approche compositionnelle enrichie par des informations contextuelles sur les 785 termes non traduits de la liste de référence. Dans cette expérience, les paramètres utilisés sont les suivants : la taille de la fenêtre contextuelle w est fixée à 3 (c'est-à-dire une fenêtre de sept mots), la mesure d'association est l'information mutuelle et la mesure de distance vectorielle est le Cosinus. D'autres combinaisons de paramètres ont été évaluées, mais les précédents paramètres sont ceux qui donnent les meilleurs résultats. La table 2 présente le pourcentage de termes français pour lesquels la bonne traduction est obtenue parmi les $Top_{1, 5, 10, \text{ et } 20}$ traductions candidates pour une traduction du français vers l'anglais. En partant des 785 termes complexes non traduits, nous traduisons 514 termes complexes français par la méthode compositionnelle enrichie avec une précision de 33,6 % pour le Top_1 et 51,6 % pour le Top_{20}. Ces résultats indiquent que la majorité des termes complexes correctement traduits sont en fait obtenus pour le Top_5.

# traductions	Top_1	Top_5	Top_{10}	Top_{20}
514	33,6 %	48,9 %	50,7 %	51,6 %

TABLE 2 – Précision des traductions pour la méthode compositionnelle enrichie par des informations contextuelles

En ce qui concerne les termes complexes correctement traduits, nous trouvons une grande majorité de termes complexes français impliquant un adjectif relationnel. Par exemple, le terme complexe français *dépistage mammographique* n'est pas traduit par l'approche compositionnelle

de base puisque l'adjectif relationnel français *mammographique* n'est pas trouvé dans le dictionnaire bilingue. En revanche, la traduction attendue *mammographic screening* est trouvée pour le Top_3 avec l'approche basée sur le contexte dans la mesure où nous avons associé le vecteur de contexte français de *mammographique* avec le vecteur de contexte anglais de *mammographic* et la paire français/anglais *dépistage/screening* est bien présente dans le dictionnaire. Les autres termes complexes correctement traduits sont principalement des termes avec une structure compositionnelle pour lesquels un élément n'est pas trouvé dans le dictionnaire comme : *amélioration significative/significant benefit* (Top_1) et *analyse multivariée/multivariate analysis* (Top_4) ou sans structure compositionnelle comme *curage axillaire/axillary dissection* (Top_{11}). Pour ce qui est des termes complexes mal traduits, nous identifions principalement deux situations. D'une part, nous trouvons comme traductions candidates des termes sémantiquement proches du terme à traduire comme *retrospective study* pour *étude comparative*. D'autre part, nous ne proposons parfois qu'une sous-partie du terme complexe anglais tel que *node dissection* pour *curage ganglionnaire* (*lymph node dissection*). Cette dernière situation nécessite la prise en compte de la fertilité pour pouvoir être traitée.

7 Travaux connexes

La plupart des travaux d'alignement lexical à partir de corpus comparables qui s'appuient sur le contexte traitent uniquement les mots simples (Fung, 1998; Rapp, 1999; Chiao et Zweigenbaum, 2002; Gaussier *et al.*, 2004; Laroche et Langlais, 2010, parmi d'autres).

Les travaux portant sur la traduction des termes complexes adoptent plutôt l'approche compositionnelle simple comme celle présentée en section 3 ou améliorée par des propriétés morphologiques ou du repli (Grefenstette, 1999; Tanaka, 2002; Robitaille *et al.*, 2006; Vintar, 2010).

Les cooccurrents sont au cœur de toutes les tâches de désambiguïsation contextuelle. Pour la traduction automatique statistique, Koehn et Knight (2002) construisent un lexique bilingue à partir de corpus comparables composé de mots possédant une graphie proche dans les deux langues. L'utilisation de cooccurrents permet d'améliorer la précision de ce lexique bilingue de 15 %. Une approche similaire est utilisée par Haghighi et al. (2008) qui se contentent de réduire l'espace de recherche et par Ismail et Manandhar (2010) qui s'appuient sur des cooccurrents spécifiques au domaine.

Munteanu et Marcu (2006) extraient des segments de texte parallèles à partir de corpus comparables. À l'aide d'un corpus parallèle, ils extraient pour chaque mot du texte source ses traductions candidates et obtiennent ainsi un premier dictionnaire bilingue où chaque traduction candidate est associée à un poids. Un segment de texte parallèle rencontré dans un corpus comparable sera une suite de mots apparaissant dans un texte source pour laquelle chacun des mots a une traduction dans le lexique pondéré. Cette méthode correspond à la méthode compositionnelle de base où est substitué à un dictionnaire bilingue un dictionnaire construit à partir d'un corpus aligné. Cette méthode diffère de plus sur le type et la taille du corpus utilisé ainsi que sur la nature des segments textuels qui ne correspondent pas forcément à une entrée lexicale.

Enfin, Shezaf et Rappoport (2010) filtrent à l'aide de cooccurrents les entrées d'un dictionnaire bilingue bruité construit par pivot à l'aide d'une lingua franca. L'utilisation de ces cooccurrents recueillis sur un corpus comparable augmente la qualité du dictionnaire bilingue de 20 %.

8 Conclusion

Dans cet article, nous avons proposé une méthode mixte pour aider à la construction de terminologies bilingues à partir de corpus comparables. Nous avons montré que la méthode compositionnelle utilisée par les programmes d'alignement de termes complexes pouvait être largement améliorée à l'aide de cooccurrents collectés à partir de corpus comparables. Cette méthode permet de traduire de nombreux termes complexes dont les traductions ne pouvaient pas être trouvées par la seule méthode compositionnelle. Prochainement, nous souhaitons généraliser cette méthode pour prendre en compte d'autres types de termes complexes qui existent dans d'autres langues tels que les composés morphologiques allemands ou chinois. Nous étudierons comment éviter de générer des traductions incomplètes et comment résoudre ainsi le problème de fertilité pour les termes complexes. Enfin, nous envisageons de modifier le protocole d'évaluation en acceptant plusieurs traductions possibles dans le cas de synonymes ou de traductions proches qui n'ont pas été détectés lors de l'extraction terminologique.

Remerciements

Ce travail a bénéficié de l'aide du septième programme cadre de la Commission européenne (FP7/2007-2013) (Grant Agreement no 248005).

Références

BALDWIN, T. et TANAKA, T. (2004). Translation by Machine of Complex Nominals : Getting it Right. *In Proceedings of the ACL 2004 Workshop on Multiword Expressions : Integrating Processing*, pages 24–31, Barcelona, Spain.

BOWKER, L. et PEARSON, J. (2002). *Working with Specialized Language : A Practical Guide to Using Corpora*. Routledge, London/New York.

BRILL, E. (1994). Some Advances in Transformation-Based Part of Speech Tagging. *In Proceedings of the 12th National Conference on Artificial Intelligence (AAAI'94)*, pages 722–727, Seattle, WA, USA.

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

CHIAO, Y.-C. et ZWEIGENBAUM, P. (2002). Looking for Candidate Translational Equivalents in Specialized, Comparable Corpora. *In Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 1208–1212.

DAILLE, B. (2003). Terminology Mining. *In* PAZIENZA, M. T., éditeur : *Information Extraction in the Web Era*, pages 29–44. Springer.

FUNG, P. (1998). A Statistical View on Bilingual Lexicon Extraction : From Parallel Corpora to Non-parallel Corpora. *In* FARWELL, D., GERBER, L. et HOVY, E., éditeurs : *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA'98)*, pages 1–16, Langhorne, PA, USA.

GAUSSIER, E., RENDERS, J.-M., MATVEEVA, I., GOUTTE, C. et DÉJEAN, H. (2004). A Geometric View on Bilingual Lexicon Extraction from Comparable Corpora. *In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL'04)*, pages 526–533, Barcelona, Spain.

GREFENSTETTE, G. (1999). The World Wide Web as a Resource for Example-Based Machine Translation Tasks. *In ASLIB'99 Translating and the Computer 21*, London, UK.

HAGHIGHI, A., LIANG, P., BERG-KIRKPATRICK, T. et KLEIN, D. (2008). Learning Bilingual Lexicons from Monolingual Corpora. *In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL'08)*, pages 771–779, Columbus, Ohio, USA.

ISMAIL, A. et MANANDHAR, S. (2010). Bilingual lexicon extraction from comparable corpora using in-domain terms. *In Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 481–489, Beijing, China.

KEENAN, E. L. et FALTZ, L. M. (1985). *Boolean Semantics for Natural Language*. D. Reidel, Dordrecht, Holland.

KOEHN, P. et KNIGHT, K. (2002). Learning a translation lexicon from monolingual corpora. *In Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition*, pages 9–16, Philadelphia, Pennsylvania, USA.

LAROCHE, A. et LANGLAIS, P. (2010). Revisiting Context-based Projection Methods for Term-Translation Spotting in Comparable Corpora. *In Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 617–625, Beijing, China.

LI, B. et GAUSSIER, E. (2010). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. *In Proceedings of the 23rd International Conference on Computational Linguistics*, COLING'10, pages 644–652, Beijing, China.

MCENERY, A. et XIAO, Z. (2007). Parallel and comparable corpora : What is happening ? *In* ANDERMAN, G. et ROGERS, M., éditeurs : *Incorporating Corpora : The Linguist and the Translator*, Multilingual Matters. Clevedon.

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

MORIN, E. et DAILLE, B. (2010). Compositionality and Lexical Alignment of Multi-word terms. *In Language Resources and Evaluation*, volume 44, pages 79–95. Springer.

MUNTEANU, D. S. et MARCU, D. (2006). Extracting parallel sub-sentential fragments from non-parallel corpora. *In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL'2006)*, Sydney, Australia.

NAKAGAWA, H. et MORI, T. (2003). Automatic term recognition based on statistics of compound nouns and their components. *Terminology*, 9(2):201–219.

NOMURA, M. et M., I. (1989). *Gakujutu Yogo Goki-Hyo*. National Language Research Institute, Tokyo.

PEKAR, V., MITKOV, R., BLAGOEV, D. et MULLONI, A. (2006). Finding translations for low-frequency words in comparable corpora. *Machine Translation*, 20(4):247–266.

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)*, pages 519–526, College Park, MD, USA.

ROBITAILLE, X., SASAKI, X., TONOIKE, M., SATO, S. et UTSURO, S. (2006). Compiling French-Japanese Terminologies from the Web. *In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*, pages 225–232, Trento, Italy.

SAG, I. A., BALDWIN, T., BOND, F., COPESTAKE, A. A. et FLICKINGER, D. (2002). Multiword Expressions : A Pain in the Neck for NLP. *In Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'02)*, pages 1–15, Mexico City, Mexico.

SAVARY, A. et JACQUEMIN, C. (2003). Reducing Information Variation in Text. *In* GREFENSTETTE, G., éditeur : *Text- and Speech-Triggered Information Access*, Lecture Notes in Computer Science, pages 141–181. Springer Verlag.

SHAROFF, S., BABYCH, B. et HARTLEY, A. (2009). 'Irrefragable answers' using comparable corpora to retrieve translation equivalents. *Language Resources and Evaluation*, 43(1):15–25.

SHEZAF, D. et RAPPOPORT, A. (2010). Bilingual lexicon generation using non-aligned signatures. *In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL'10)*, pages 98–107, Uppsala, Sweden.

TANAKA, T. (2002). Measuring the Similarity between Compound Nouns in Different Languages Using Non-parallel Corpora. *In Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 1–7, Taipei, Taiwan.

TANAKA, T. et BALDWIN, T. (2003). Noun-Noun Compound Machine Translation : A Feasibility Study on Shallow Processing. *In Proceedings of the ACL 2003 workshop on Multiword Expressions : Analysis, Acquisition and Treatment*, pages 17–24, Sapporo, Japan.

VINTAR, S. (2010). Bilingual term recognition revisited : The bag-of-equivalents term alignment approach and its evaluation. *Terminology*, 16(2):141–158.