Analyse des contextes et des candidats dans l'identification des équivalents terminologiques en corpus comparables

Audrey Laroche[1]
(1) RALI–DIRO, Université de Montréal, C.P. 6128, Succ. Centre-Ville Montréal (Québec) H3C 3J7, Canada
audrey.laroche@umontreal.ca

RÉSUMÉ

L'approche standard d'identification d'équivalents terminologiques à partir de corpus comparables repose sur la comparaison de mots contextuels en langues source et cible et sur l'utilisation d'un lexique bilingue. Nous analysons manuellement, selon des critères linguistiques (parties du discours, spécificité et relations sémantiques), les propriétés des mots contextuels et des erreurs commises par l'approche standard appliquée à la terminologie médicale pour suggérer des améliorations basées sur la sélection de mots contextuels.

ABSTRACT

Analysis of contexts and candidates in term-translation spotting in comparable corpora

The standard approach for identifying terminological equivalents from comparable corpora is based on the comparison of source and target language context words using a bilingual lexicon. We cary a manual analysis of the linguistic properties (parts of speech, specificity and semantic relations) of the context words and the inacurrate equivalents given by the standard approach applied to medical terminology, in order to suggest improvements based on the selection of context words.

MOTS-CLÉS : équivalents terminologiques, vecteurs contextuels, corpus comparables, terminologie médicale, étude qualitative.

KEYWORDS: terminological equivalents, contextual vectors, comparable corpora, medical terminology, qualitative study.

1 Introduction

L'identification automatique d'équivalents de termes (comme *accident vasculaire cérébral* et *stroke*) est un sujet qui intéresse de nombreux chercheurs. Les applications potentielles en sont multiples : aide à la rédaction de dictionnaires bilingues spécialisés ou à la traduction, recherche d'information multilingue, traduction automatique, etc. (Rapp, 1999; Li et Gaussier, 2010). La plupart des techniques proposées pour identifier des équivalents de termes reposent sur l'hypothèse selon laquelle le contexte d'un mot en langue source est similaire au contexte de son équivalent en langue cible (Rapp, 1999).

Les équivalents terminologiques sont, dans l'approche dite standard (Rapp, 1999), repérés en

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 407–414,
Grenoble, 4 au 8 juin 2012. ©2012 ATALA & AFCP

comparant, à l'aide d'un lexique bilingue de projection, leurs vecteurs contextuels extraits de corpus comparables [1]. Parmi les travaux récents s'inspirant de cette approche, mentionnons ceux de (Rubino, 2011; Rubino et Linarès, 2011), qui combinent par vote trois niveaux de représentation pour chaque terme (contexte, thème et graphie). (Morin et Prochasson, 2011) utilisent un lexique de projection spécialisé formé à partir de phrases parallèles extraites automatiquement de corpus comparables. (Prochasson et Fung, 2011) combinent des vecteurs contextuels et des modèles de cooccurrence entre mots pour trouver les équivalents de termes rares. (Li et Gaussier, 2010) et (Li *et al.*, 2011) proposent deux méthodes pour améliorer le degré de comparabilité des corpus, paramètre également étudié par (Prochasson, 2009). Toutes ces stratégies contribuent à améliorer la performance de l'identification d'équivalents terminologiques. Notons que la majorité des travaux mentionnés portent sur le domaine médical, qui fait aussi l'objet du présent article.

L'analyse de la performance de ces divers systèmes repose sur des mesures classiques comme la précision et le rappel. Dans le présent article, nous analysons qualitativement les mots contextuels pris en compte dans l'approche standard, notamment du point de vue de leur partie du discours, de leur degré de spécificité et de leur relation sémantique avec le terme en langue source. Cette analyse servira éventuellement à déterminer s'il est possible de sélectionner les mots du contexte pour augmenter la performance de l'identification d'équivalents. De plus, nous examinons les candidats équivalents erronés obtenus avec l'approche standard, de façon à catégoriser les erreurs typiques et ainsi proposer des heuristiques pouvant améliorer la performance.

2 Approche par projection de contextes

L'approche standard d'extraction d'équivalents terminologiques à partir de corpus comparables est basée sur la comparaison de vecteurs de mots contextuels tirés de corpus de langues source et cible. Chaque terme dont on cherche l'équivalent est caractérisé par ses mots voisins (la définition de *voisinage* variant d'une étude a l'autre) qui sont pondérés (à l'aide d'une mesure d'association) en fonction de leur fréquence de cooccurrence avec ce terme source. Les mots voisins sont projetés dans la langue cible à l'aide d'un lexique bilingue de projection (constitué de mots spécialisés ou généraux), formant ainsi un vecteur projection. Ce dernier est comparé aux vecteurs de mots contextuels des termes extraits du corpus en langue cible : les candidats équivalents sont ceux dont le vecteur contextuel ressemble le plus (selon une certaine mesure de similarité) au vecteur projection.

Cette approche par projection de contextes comporte plusieurs paramètres de base. (Laroche et Langlais, 2010) en ont fait une étude détaillée en utilisant des corpus comparables anglais et français tirés de Wikipédia avec NlGbAse (Charton et Torres-Moreno, 2010), ainsi que des lexiques de projection contenant des mots généraux (provenant de *Freelang*) et spécialisés (provenant du *Multilingual glossary of technical and popular medical terms* du Heymans Institute of Pharmacology). Les expériences portaient sur 5 000 termes nominaux simples et complexes [2]

1. Des corpus sont comparables s'ils ne sont pas des traductions l'un de l'autre, mais portent sur le même domaine.
2. Une approche « en amont » (Morin, 2007) est utilisée pour extraire les contextes (mots simples seulement) des termes complexes sources ; les mots des vecteurs projections sont simples ou complexes, selon leur traduction dans le lexique de projection. Des vecteurs contextuels sont extraits du corpus cible pour tous les mots simples et tous les bigrammes composés de deux mots lexicaux (99,5 % des équivalents de référence comptant au plus deux composantes).

du domaine médical tirés du MeSH [3]. Les meilleures valeurs de paramètres, selon les expériences de (Laroche et Langlais, 2010) sur 70 configurations différentes, sont :
– Longueur du contexte : la phrase [4].
– Mesure d'association : le ratio log-odds.
– Mesure de similarité entre vecteurs contextuels : le cosinus.
– Taille du lexique bilingue de projection : 9 000 termes (pour des corpus de 90 328 mots (source) et 38 929 mots (cible) en moyenne).
– Contenu du lexique de projection : mots généraux et termes spécialisés.

Tout comme (Prochasson, 2009), (Laroche et Langlais, 2010) ont observé que les mesures d'association et de similarité sont les paramètres ayant la plus grande influence sur la performance et que les différents paramètres s'influencent les uns les autres. D'autres paramètres importants ont fait l'objet d'articles, comme le degré de comparabilité des corpus (Li et Gaussier, 2010; Li *et al.*, 2011) et la fréquence d'occurrence des termes (Prochasson et Fung, 2011; Li *et al.*, 2011). Tel que mentionné en introduction, de nombreuses techniques ont récemment été proposées pour améliorer l'approche standard.

3 Analyse des mots contextuels

Nous avons analysé manuellement le contenu des vecteurs projections de 30 termes nominaux simples ou complexes choisis arbitrairement (mais commençant par la lettre *a* [5]), et ce, pour quatre tailles de lexiques de projection différentes (5 000, 7 000, 9 000 et 11 000 entrées, dont 2 000 du domaine médical et le reste de langue générale) ; l'analyse porte donc sur 120 vecteurs projections. Ces vecteurs ont été obtenus en utilisant les ressources, l'implémentation et la configuration paramétrique optimale de (Laroche et Langlais, 2010). Rappelons que les vecteurs projections correspondent aux mots voisins des termes en langue source qui sont projetés vers la langue cible à l'aide du lexique bilingue de projection. Le Tableau 1 présente le contenu de quelques-uns des vecteurs étudiés (obtenus avec le lexique de projection de 5 000 entrées).

Nous examinons la répartition des parties du discours de même que le degré de spécificité (langue spécialisée ou générale) et la sémantique des 20 plus forts mots contextuels des 30 termes français [6]. Ceci nous permet d'identifier des critères pouvant améliorer la qualité de l'identification d'équivalents. Cette analyse est complémentaire à l'inspection manuelle des mots contextuels dans l'approche standard menée par (Morin, 2007) pour 100 termes simples du domaine de la foresterie, qui est centrée sur l'apport des termes complexes dans les vecteurs contextuels en langue source et les vecteurs projections. Ses résultats ne peuvent pas être directement comparés aux nôtres, puisque notre implémentation n'extrait du corpus source que des termes simples pour peupler les vecteurs contextuels.

3. Les ressources utilisées dans l'étude en question sont disponibles sur `http://olst.ling.umontreal.ca/~audrey/coling2010`

4. Selon (Prochasson, 2009; Rubino et Linarès, 2011), la longueur optimale dépend de la fréquence du terme source.

5. *Abcès* (*abscess*), *acétylène* (*acetylene*), *acétylcholine* (*acetylcholine*), *accident vasculaire cérébral* (*stroke*), *acide lactique* (*lactic acid*), *acides* (*acids*), *adhésifs* (*adhesives*), *aine* (*groin*), *albinisme* (*albinism*), *allèles* (*alleles*), *alliages* (*alloys*), *aloès* (*aloe*), *amiante* (*asbestos*), *amidon* (*starch*), *amnésie* (*amnesia*), *amphétamines* (*amphetamines*), *analyse harmonique* (*Fourier analysis*), *anatomie* (*anatomy*), *anesthésie* (*anesthesia*), *anorexie mentale* (*anorexia nervosa*), *antigenes* (*antigens*), *antioxydants* (*antioxidants*), *anxiété* (*anxiety*), *apnée* (*apnea*), *appendicite* (*appendicitis*), *artère pulmonaire* (*pulmonary artery*), *artères* (*arteries*), *articulations* (*joints*), *atrophie* (*atrophy*).

6. (Laroche et Langlais, 2010) étudient en détail la pondération statistique des mots contextuels.

Terme source	Vecteur projection
albinisme	*syndrome*, *Parkinsonism*, *deficit*, *hypoplasia*, *anomaly*, *ocular*, *corpus luteum*, *pigments*, *absence*, *mutation*, *origin*, *retina*, *nystagmus*, *nobody*, *pigmentation*, *abatement*, *synthesis*, *lyophilisate*
artères	*aorta*, *pulmonary*, *Parkinsonism*, *cardiac*, *coronary*, *circulation*, *infarction*, *fact*, *risk patient*, *vascular*, *fat*, *afterload*, *members*, *network*, *hypertension*, *myelosuppression*, *myocardium*, *arterial*, *lyophilisate*, *fabrics*
artère pulmonaire	*pulmonary*, *cardiac*, *aorta*, *afterload*, *arterial*, *duct*, *ventricular*, *coronary*, *venous*, *hypertension*, *Parkinsonism*, *systolic*, *function*, *stenosis*, *absence*, *fat*, *fact*, *circulation*, *diameter*, *anomaly*

TABLE 1 – Contenu des vecteurs projections

3.1 Parties du discours

Les noms sont fortement majoritaires dans les vecteurs projections. Par exemple, avec le lexique de projection de 5 000 entrées, 80,9 % des mots contextuels dans les vecteurs sont des noms. Par comparaison, ce même lexique de projection compte 68,7 % de noms. Rappelons que dans nos expériences, les termes dont on cherche l'équivalent sont des noms ou des syntagmes nominaux (de deux composants) ; il serait intéressant de voir, avec de nouveaux équivalents de référence, si la majorité des mots contextuels seraient aussi des noms si les termes dont on cherche l'équivalent étaient d'une autre partie du discours.

D'autre part, il n'y a pas de différence significative quant aux proportions des parties du discours des mots contextuels selon que les termes dont on cherche l'équivalent sont simples (par ex. *artères*) ou complexes (par ex. *artère pulmonaire*), ces deux types de termes étant traités de la même façon (c'est-à-dire d'un seul bloc) dans notre implémentation.

3.2 Spécificité

Pour analyser le degré de spécificité des mots contextuels dans les vecteurs projections, nous avons classé chacun d'entre eux dans l'une de trois catégories : « domaine médical », « langue générale » ou « ambigu ». Certains mots sont ambigus parce qu'ils appartiennent à la fois à la langue générale et à la langue médicale. Par exemple, dans le vecteur projection d'*anatomie* se trouve le mot *fingers*, que nous avons considéré comme ambigu. Cette tâche est relativement difficile étant donné que nous ne sommes ni spécialiste du domaine médical, ni terminologue, ni anglophone ; certaines tendances se dessinent tout de même (Figure 1).

En moyenne, 57,5 % des mots contextuels sont du domaine médical lorsque le lexique de projection compte 5 000 entrées, bien que celui-ci contienne 36 % d'entrées du domaine médical : les mots contextuels projetés et qui ont un score d'association fort avec le terme source ont donc tendance à être des mots spécialisés. La quantité de termes spécialisés dans les vecteurs projections dépend tout de même de la proportion de termes spécialisés dans le lexique de projection, comme le montre la Figure 1 (36 % des entrées sont du domaine médical dans le lexique de taille 5 000, contre 16 % dans le lexique de 11 000 entrées).

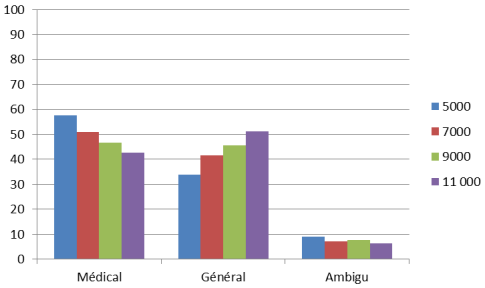

FIGURE 1 – Répartition des domaines dans les vecteurs projections (en %) selon la taille du lexique de projection

Avec notre échantillon de 30 termes en langue source, il n'y a pas de corrélation entre la proportion de mots contextuels qui appartiennent au domaine médical et le rang du bon équivalent. Toutefois, en utilisant l'approche standard pour identifier les équivalents de 122 termes simples du domaine médical, (Morin et Prochasson, 2011) ont montré que les résultats sont meilleurs lorsque le lexique de projection contient des termes spécialisés en plus de mots de la langue générale. Puisque le contenu du lexique de projection a une influence directe sur le contenu des vecteurs projections, il est possible qu'avec un échantillon de taille comparable nous observions un effet analogue ; de plus amples analyses seraient nécessaires pour le vérifier.

3.3 Relations sémantiques

Les mots faisant partie du vecteur projection sont souvent sémantiquement liés au terme source. Selon nos observations sur un sous-ensemble de 10 termes, les mots contextuels peuvent être en relation de collocation (*amnésie* et *infantile*), de synonymie (*anorexie mentale* et *anorexia*) ou d'hyponymie (*artères* et *aorta*) avec le terme source, ils peuvent lui être liés morphologiquement (*anatomie* et *anatomical*), avoir un lien sémantique spécifique au domaine médical (par exemple *est un symptôme de* ou autres relations plus difficiles à caractériser comme celle entre *albinisme* et *pigmentation*), n'être pas liés sémantiquement au terme source ou encore être trop génériques. Les relations spécifiques au domaine médical sont les plus fréquentes (moyenne de 7,3/20 mots contextuels en utilisant le lexique bilingue de projection de 5 000 entrées), suivies des relations de collocation (moyenne de 4,7/20). Les mêmes mots contextuels de sémantisme faible (comme *nobody*) ou qui ne sont pas liés sémantiquement au terme source reviennent dans plusieurs vecteurs ; les premiers pourraient être éliminés en les incluant dans un antidictionnaire, et la présence des seconds peut s'expliquer par certaines faiblesses de nos ressources, tel qu'expliqué dans la prochaine section.

3.4 Cas problématiques

Dans environ 10 % des vecteurs projections examinés, les domaines d'où proviennent les mots contextuels sont très disparates. Par exemple, le vecteur projection obtenu pour *aine* (avec le lexique de projection de 5 000 entrées) est formé des mots : *grand mal*, *meadow*, *radio*, *music*, *region*, *quarter*, *network*, *yearly*, *policy*, *family*, *country*, *origin*, *population*, *venous*, *dance*, *diffusion*, *orange*, *contest*, *prince*, *hockey*. Les candidats équivalents pour *aine* sont à leur tour très diversifiés sémantiquement. Plusieurs éléments peuvent expliquer cette disparité : la façon dont le corpus a été construit, son contenu, la polysémie.

Nous avons remarqué, au cours de l'analyse, certaines lacunes concernant le lexique bilingue de projection (dans lequel les mots contextuels retenus figurent nécessairement). Dans ce lexique, l'équivalent de *maladie* est *Parkinsonism*, celui de *double* est *doubleblind*, celui de *mal* est *grand mal*, celui de *corps* est *corpus luteum*, celui de *produit* est *lyophilisate*, celui de *risque* est *risk patient*. Ces équivalents du lexique bilingue de projection sont beaucoup plus spécifiques que les termes français. Le mot *Parkinsonism* se retrouve ainsi dans plusieurs vecteurs projections (comme ceux des trois exemples du Tableau 1), mais, étant donné qu'il est rare dans les corpus anglais, il ne figure virtuellement pas dans les vecteurs contextuels des candidats équivalents. La performance de l'identification d'équivalents pourrait sans doute être améliorée si le lexique bilingue de projection était d'une meilleure qualité (Morin et Prochasson, 2011).

Enfin, les mots contextuels sont parfois redondants, étant donné que, dans notre implémentation, ils ne sont pas lemmatisés (ceci afin de ne pas faire dépendre l'approche standard d'outils externes qui ne sont pas entraînés sur des corpus médicaux). Pratiquement tous les chercheurs lemmatisent leur corpus avant de former les vecteurs contextuels, et notre examen manuel des mots projetés leur donne raison.

4 Analyse des candidats équivalents

Nous avons analysé manuellement les 20 premiers candidats équivalents de 30 termes français (les mêmes que ceux de la section 3) obtenus avec différents paramètres dans les expériences de (Laroche et Langlais, 2010), pour un total de 360 listes de 20 candidats. Avec la configuration paramétrique optimale, pour 18 des 30 termes examinés, le bon équivalent n'est pas au premier rang. On y trouve plutôt 3 candidats correspondant à une composante du terme complexe attendu (*acid* pour *acide lactique*), 1 qui ne diffère de l'équivalent de référence que par la morphologie (*joint* pour *articulations*), 11 qui sont sémantiquement liés à l'équivalent de référence (*eating* pour *anorexie mentale*), 2 qui sont des mots génériques (*causes* pour *albinisme* et *species* pour *antioxydants*) et 1 qui n'a aucun lien avec l'équivalent de référence (*combo* pour *aine*).

Pour toutes les valeurs de paramètres testées dans (Laroche et Langlais, 2010), il y a systématiquement environ deux candidats sur 20 qui sont des termes complexes, et ce, peu importe si l'équivalent de référence est simple ou complexe. Les termes équivalents d'une langue à l'autre n'ont pas toujours la même complexité (*accident vasculaire cérébral* et *stroke*) (Morin et Daille, 2004). Mais le fait que, peu importe les valeurs des paramètres, le système récupère le même nombre de candidats équivalents complexes suggère qu'il pourrait être amélioré, par exemple en extrayant préalablement les termes (simples et complexes) dans les corpus sources et cibles (Daille et Morin, 2005). Par ailleurs, dans plusieurs cas, les composantes des termes complexes

(ex. *fatty* et *acids*) sont situées à de meilleurs rangs que l'équivalent de référence (*fatty acids*) dans la liste des candidats équivalents. Ceci peut être attribué à notre implémentation, dans laquelle des vecteurs contextuels en langue cible sont construits à la fois pour les termes complexes (bigrammes) et pour chacune de leurs composantes. Une heuristique pourrait donner plus de poids au terme complexe dans ces cas.

Dans presque toutes les listes de candidats équivalents, de un à cinq (environ) candidats parmi les 20 sont morphologiquement liés à l'équivalent de référence. L'étiquetage des parties du discours et la lemmatisation (que font plusieurs chercheurs) permettraient de regrouper les candidats dont seule la flexion varie (par ex., *acids* et *acid*) pour ensuite proposer comme équivalent le candidat qui a le même nombre que le terme source. Le fait que l'approche par projection permette de trouver des candidats liés morphologiquement (comme *oxygen*, *oxidative*, *antioxidant* et *reactive oxygen* pour *antioxydants*) montre que les indices contextuels sont pertinents pour trouver automatiquement les flexions et les dérivations d'un mot donné.

Tous nos équivalents de référence sont des noms ou des syntagmes nominaux ; or, environ 25 % des candidats équivalents ont une autre partie du discours. De façon générale, les parties du discours ne sont pas toujours identiques entre les termes équivalents dans les langues distinctes (Névéol et Ozdowska, 2006), mais, étant donnés notre paire de langues source et cible (français et anglais) et notre domaine (le lexique médical), il serait justifié de réordonner les candidats équivalents pour favoriser les termes nominaux.

Enfin, dans pratiquement toutes les listes de candidats équivalents observées, au moins 10 (et souvent au-delà de 15) des 20 candidats sont sémantiquement liés à l'équivalent de référence. Ces candidats seraient pertinents pour construire des thésaurus, des dictionnaires de synonymes ou d'analogies, etc. Parmi les autres candidats, ceux qui sont des mots génériques comme *process* et *levels* pourraient être inclus dans un antidictionnaire. Si les candidats équivalents ne sont pas génériques, mais appartiennent à des domaines très différents, cela indique que le terme source est polysémique et devrait être désambiguïsé.

5 Conclusion

L'approche standard par projection de contextes pour l'identification d'équivalents terminologiques en corpus comparables est une technique sur laquelle se basent plusieurs travaux récents qui proposent des stratégies pour améliorer la précision. Nous avons analysé manuellement, selon des critères linguistiques, le contenu des vecteurs projections et les listes de candidats équivalents obtenus avec l'implémentation de (Laroche et Langlais, 2010) appliquée à 30 termes du domaine médical. Les mots contextuels ont souvent la même partie du discours que le terme source, ils ont tendance à être spécialisés et à être liés au terme source par des relations spécifiques au domaine et par la collocation ; la qualité des ressources a une influence directe sur celle des vecteurs projections. Parmi les candidats équivalents, ceux qui ont le même nombre, le même degré de complexité et la même partie du discours que le terme source sont à privilégier (du moins pour le domaine médical). Le fait que la très grande majorité des candidats soient sémantiquement liés à l'équivalent de référence confirme l'intérêt de l'approche basée sur la projection de contextes. Les travaux futurs vérifieront l'influence sur la performance des heuristiques proposées ici et de la sélection de mots contextuels en fonction des caractéristiques que nous avons fait ressortir.

Remerciements

Nous remercions Raphaël Rubino pour les ressources ainsi que Philippe Langlais, Patrick Drouin et les relecteurs pour leurs commentaires pertinents. Nous reconnaissons le soutien du FQRSC.

Références

CHARTON, E. et TORRES-MORENO, J.-M. (2010). Nlgbase : A free linguistic resource for natural language processing systems. *In Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*.

DAILLE, B. et MORIN, E. (2005). French-English terminology extraction from comparable corpora. *In 2nd International Joint Conference on Natural Language Processing*, pages 707–718.

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*, pages 617–625.

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*, pages 644–652.

LI, B., GAUSSIER, E., MORIN, E. et HAZEM, A. (2011). Degré de comparabilité, extraction lexicale bilingue et recherche d'information interlingue. *In Actes de TALN 2011*.

MORIN, E. (2007). Apport des termes complexes à l'acquisition lexicale multilingue à partir de corpus comparables spécialisés : entre intuition et réalité. *In Actes, 7ème Rencontres Terminologie et Intelligence Artificielle*, pages 11–20.

MORIN, E. et DAILLE, B. (2004). Extraction de terminologies bilingues rtir de corpus comparables. *Traitement automatique des langues*, 45(3):103–122.

MORIN, E. et PROCHASSON, E. (2011). Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. *In Proceedings of the 4th Workshop on Building and Using Comparable Corpora : Comparable Corpora and the Web*, pages 27–34.

NÉVÉOL, A. et OZDOWSKA, S. (2006). Terminologie médicale bilingue anglais/français : usages cliniques et bilingues. *Glottopol*, 8.

PROCHASSON, E. (2009). *Alignement multilingue en corpus comparables spalisés : Caractérisation terminologique multilingue*. Thèse de doctorat, Université de Nantes.

PROCHASSON, E. et FUNG, P. (2011). Rare word translation extraction from aligned comparable documents. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, pages 1327–1335.

RAPP, R. (1999). Automatic identification of word translations from unrelated English and German corpora. *In 37th Annual Meeting of the Association for Computational Linguistics*, pages 66–70.

RUBINO, R. (2011). *Traduction automatique statistique et adaptation à un domaine spécialisé*. Thèse de doctorat, Université d'Avignon et des Pays de Vaucluse.

RUBINO, R. et LINARÈS, G. (2011). Une approche multi-vue pour l'extraction terminologique bilingue. *In Conférence en Recherche d'Infomations et Applications*, pages 97–111.