Extraction de lexiques bilingues à partir de Wikipédia

Rahma Sellami[1] *Fatiha Sadat*[2] *Lamia Hadrich Belguith*[1]
(1) ANLP Research Group – Laboratoire MIRACL
Faculté des Sciences Economiques et de Gestion de Sfax
B.P. 1088, 3018 - Sfax – TUNISIE
(2) Université du Québec à Montréal, 201 av. President Kennedy,
Montréal, QC, H3X 2Y3, Canada
Rahma.Sellami@fsegs.rnu.tn, sadat.fatiha@uqam.ca,
l.belguith@fsegs.rnu.tn

RESUME

Avec l'intérêt accru de la traduction automatique, le besoin de ressources multilingues comme les corpus comparables et les lexiques bilingues s'est imposé. Ces ressources sont peu disponibles, surtout pour les paires de langues qui ne font pas intervenir l'anglais. Cet article présente notre approche sur l'extraction de lexiques bilingues pour les paires de langues arabe-français et yoruba-français à partir de l'encyclopédie en ligne Wikipédia. Nous exploitons la taille gigantesque et la couverture de plusieurs domaines des articles pour extraire deux lexiques, qui pourront être exploités pour d'autres applications en traitement automatique du langage naturel.

ABSTRACT

Bilingual lexicon extraction from Wikipedia

With the increased interest of the machine translation, needs of multilingual resources such as comparable corpora and bilingual lexicon has increased. These resources are not available mainly for pair of languages that do not involve English.

This paper aims to describe our approach on the extraction of bilingual lexicons for Arabic-French and Yoruba-French pairs of languages from the online encyclopedia, Wikipedia. We exploit the large scale of Wikipedia article to extract two bilingual lexicons that will be very useful for natural language applications.

MOTS-CLES : Lexique bilingue, corpus comparable, Wikipédia, arabe-français, yoruba-français.

KEYWORDS : Bilingual lexicon, comparable corpora, Wikipedia, Arabic-French, Yoruba-French.

JEP-TALN-RECITAL 2012, Atelier TALAf 2012: Traitement Automatique des Langues Africaines, pages 107–117,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Les ressources linguistiques multilingues sont généralement construites à partir de corpus parallèles. Cependant, l'absence de ces corpus a incité les chercheurs à exploiter d'autres ressources multilingues, telles que les corpus comparables : ensembles de textes dans différentes langues, qui ne sont pas des traductions les uns des autres (Adafre et de Rijke, 2006), mais qui contiennent des textes partageant des caractères communs, tel que le domaine, la date de publication, etc. Car moins contrains, ils sont donc plus faciles à construire que les corpus parallèles.

Les lexiques bilingues constituent une partie cruciale dans plusieurs applications telles que la traduction automatique (Och et Ney, 2003) et la recherche d'information multilingue (Grefenstette, 1998).

Dans cet article, nous cherchons à exploiter l'aspect multilingue ainsi que la taille gigantesque de l'encyclopédie en ligne, Wikipédia, comme un grand corpus comparable pour l'extraction de deux lexiques bilingues (arabe-français et yoruba-français). (Morin, 2007) a montré que non seulement la taille du corpus comparable mais aussi sa qualité est importante pour l'extraction d'un dictionnaire bilingue. Nous proposons d'utiliser une méthode simple mais efficace, il s'agit d'exploiter les liens inter-langues entre les articles Wikipédia afin d'extraire des termes (simples ou composés) arabes et yoruba et leurs traductions en français, puis, utiliser une approche statistique pour aligner les mots des termes composés.

Les lexiques extraits seront utilisés pour l'extraction d'un corpus parallèle à partir de wikipèdia.

Le contenu de cet article se résume comme suit. La section 2 présente un bref aperçu des travaux antérieurs sur l'extraction de lexiques bilingues. La section 3 décrit certaines caractéristiques de Wikipédia que nous avons exploitées pour l'extraction de nos lexiques bilingues. La section 4 présente brièvement les langues arabe et yoruba. Nous présentons, dans la section 5, notre travail de construction des lexiques bilingues à partir de Wikipédia. Nous évaluons nos lexiques, dans la section 6. La section 7 conclu cet article et donne des pointeurs et extensions pour le futur.

2 Etat de l'art

Dans un premier temps, les chercheurs construisent les lexiques bilingues à partir des corpus parallèles. Mais, en raison de l'absence de ces ressources, l'exploitation des corpus

comparables a attiré l'attention de plusieurs chercheurs. (Morin et Daille, 2004) présentent une méthode pour l'extraction de terminologie bilingue à partir d'un corpus comparable du domaine technique. Ils extraient les termes composés dans chaque langue puis ils alignent ces termes au niveau mot en utilisant une méthode statistique exploitant le contexte des termes. (Otero, 2007) a crée un lexique bilingue (anglais-espagnol), en se basant sur des informations syntaxiques et lexicales extraites à partir d'un petit corpus parallèle. (Sadat *et al.*, 2003) ont présenté une méthode hybride qui se base sur des informations statistiques (deux modèles de traduction bidirectionnels) combinées à des informations linguistiques pour construire une terminologie anglais-japonais. (Morin et Prochasson, 2011) ont présenté une méthode pour l'extraction d'un lexique bilingue spécialisé à partir d'un corpus comparable, agrémenté d'un corpus parallèle. Ils extraient des phrases parallèles à partir du corpus comparable, puis, ils alignent ces phrases au niveau mots pour en extraire un lexique bilingue. (Hazem *et al.*, 2011) proposent une extension de l'approche par similarité inter-langue abordée dans les travaux précédents. Ils présentent un modèle inspiré des métamoteurs de recherche d'information.

Dans ce qui suit, nous décrivons les travaux antérieurs qui ont exploité Wikipédia comme corpus comparable pour la construction d'un lexique bilingue.

(Adafre et de Rijke, 2006) a créé un lexique bilingue (anglais-néerlandais) à partir de Wikipedia dans le but de l'utiliser pour la construction d'un corpus parallèle à partir des articles de Wikipédia. Le lexique extrait est composé uniquement de titres des articles Wikipédia reliés par des liens inter-langues. Les auteurs ont montré l'efficacité de l'utilisation de ce lexique pour la construction d'un corpus parallèle. (Bouma *et al.*, 2006) ont construit un lexique bilingue pour la création d'un système de question réponse multilingue (français-néerlandais). En outre, (Decklerck *et al.*, 2006) ont extrait un lexique bilingue à partir des liens inter-langues de Wikipédia. Ce lexique a été utilisé pour la traduction des labels d'une ontologie. Ces travaux sont caractérisés par le fait qu'ils exploitent uniquement les liens inter-langues de Wikipédia. Par contre, (Erdmann *et al.*, 2008) analysent non seulement les liens inter-langues de wikipédia, mais exploitent aussi les redirections et les liens inter-wiki pour la construction d'un dictionnaire anglais-japonais. Les auteurs ont montré l'apport de l'utilisation de Wikipédia par rapport aux corpus parallèles pour l'extraction d'un dictionnaire bilingue. Cet apport apparait surtout au niveau de la large couverture des termes. (Sadat et Terrasa, 2010) proposent une approche pour l'extraction de terminologie bilingue à partir de Wikipédia. Cette approche consiste à extraire d'abord des paires de termes et

traductions à partir des différents types d'informations, des liens et des textes de Wikipédia, puis, à utiliser des informations linguistiques afin de réordonner les termes et leurs traductions pertinentes et ainsi éliminer les termes cibles inutiles.

3 Bref aperçu sur les langues arabe et yoruba

3.1 La langue arabe

L'arabe (العربية) est une langue originaire de la péninsule Arabique. Elle est parlée en Asie et en Afrique du Nord. L'Arabe est issue du groupe méridional des langues sémitiques. Elle s'écrit de droite à gauche tout en utilisant des lettres qui prennent des formes différentes suivant qu'elles soient isolées, au début, au milieu ou à la fin du mot.[1]
La langue arabe est morphologiquement riche ce qui pose le problème de l'ambiguïté au niveau de son traitement automatique, un mot en arabe peut encapsuler la signification de toute une phrase (ا ذ روذ ذ ذكّ تأ/est ce que vous souvenez de nous ?).

3.2 La langue yoruba

Le yoruba (yorùbá) est une langue tonale appartenant à la famille des langues nigéro-congolaises. Le yorouba, langue maternelle d'environ 20% de la population nigériane, est également parlé au Bénin et au Togo. Au Nigéria, il est parlé dans la plus grande partie des états d'Oyo, Ogun, Ondo, Osun, Kwara et Lagos, et à l'ouest de l'état de Kogi.
La langue se subdivise en de nombreux dialectes. Il existe néanmoins aussi une langue standard[2].
Le yoruba s'écrit au moyen de plusieurs alphabet fondées sur l'alphabet latin muni d'accents pour noter les tons (dont la charge fonctionnelle est très importante), et de points souscrits pour noter les voyelles ouvertes.
La voyelle est le centre de la syllabe. Le ton apparaît comme une caractéristique inhérente à la voyelle ou à la syllabe. Il y a autant de syllabes que de tons. Le symbolisme se présente comme suit : ton haut: (/), ton bas: (\), ton moyen: (-).
Ces tons déterminent le sens du mot, une forme peut avoir plusieurs sens (ex. Igba/deux cent, Igba/calebasse, Ìgba/temps, etc)[3].

[1] http://fr.wikipedia.org/wiki/Arabe [consulté le 26/04/2012].
[2] http://fr.wikipedia.org/wiki/Yoruba_(langue) [consulté le 18/04/2012].
[3] http://www.africananaphora.rutgers.edu/downloads/casefiles/YorubaGS.pdf [consulté le 24/04/2012].

La morphologie de la langue yoruba est riche, faisant, par exemple, un large emploi du redoublement (ex. Eso/fruit, so/donner de fruits, jò/ dégoutter , òjo/pluie).

4 Caractéristiques de Wikipédia

Lors de l'extraction de terminologies bilingues à partir de corpus parallèles ou comparables, il est difficile d'atteindre une précision et une couverture suffisantes, en particulier pour les mots moins fréquents tels que les terminologies spécifiques à un domaine (Erdmann, 2008). Pour notre travail de construction de lexiques bilingues, nous proposons d'exploiter Wikipédia, une ressource multilingue dont la taille est gigantesque et qui est en développement continu.

Dans ce qui suit, nous décrivons certaines caractéristiques de Wikipédia, ces caractéristiques font de Wikipèdia une ressource précieuse pour l'extraction de ressources bilingues.

Actuellement, Wikipédia contient 21 368 483 articles dont 1 221 995 articles français, 170771 articles en langue arabe et 29 884 articles en langue yoruba[4]. Ces articles couvrent plusieurs domaines. Nous exploitons l'aspect multilingue et gigantesque de cette ressource afin d'extraire des lexiques bilingues de large couverture.

La structure de Wikipédia est très dense en liens ; ces liens relient soit des articles d'une seule langue soit des articles rédigés en langues différentes.

Les liens Wikipédia peuvent être classés en :

- Lien inter-langue : un lien inter-langue relie deux articles en langues différentes. Un article a au maximum un seul lien inter-langue pour chaque langue, ce lien a comme syntaxe *[[code de la langue cible : titre de l'article en langue cible]]* avec « *code de la langue cible* » identifie la langue de l'article cible et « *titre de l'article en langue cible* » identifie son titre (ex. *[[yo:Júpítérì]]*). Puisque les titres des articles Wikipédia sont uniques, la syntaxe des liens inter-langue est suffisante pour identifier les articles en langues cibles.
- Redirection : une redirection renvoie automatiquement le visiteur sur une autre page. La syntaxe Wikipédia d'une redirection est : *#REDIRECTION[[page de destination]]*. Les pages de redirection sont notamment utilisées pour des abréviations (ex. *SNCF* redirige vers *Société Nationale des Chemins de Fer*), des synonymes (ex. *e-*

[4] http://meta.wikimedia.org/wiki/List of Wikipedias [consulté le 01/03/2012].

mail, courriel, mél et messagerie électronique redirigent vers *courrier électronique*), des noms alternatifs (ex. *Karol Wojtyła* redirige vers *Jean-Paul II*), etc.

- Lien inter-wiki : c'est un lien vers une autre page de la même instance de Wikipédia. Le texte du lien peut correspondre au titre de l'article qui constitue la cible du lien (la syntaxe en sera alors : *[[titre de l'article]]*), ou différer du titre de l'article-cible (avec la syntaxe suivante : *[[titre de l'article|texte du lien]]*).

5 Extraction des lexiques bilingues à partir de Wikipédia

5.1 Extraction des termes

Nous avons extrait deux lexiques bilingues en exploitant la syntaxe des liens inter-langues de Wikipédia. En effet, les liens inter-langues relient deux articles en langues différentes dont les titres sont en traduction mutuelle. En outre, ces liens sont créés par les auteurs des articles, nous supposons que les auteurs ont correctement positionné ces liens. Aussi, un article en langue source est lié à un seul article en langue cible, donc, nous n'avons pas à gérer d'éventuels problèmes d'ambiguïté au niveau de l'extraction des paires de titres.

Nous avons téléchargé la base de données Wikipédia arabe (janvier 2012)[5] et yoruba (mars 2012)[6] sous format XML et nous avons extrait 104 104 liens inter-langue arabe et 15 345 liens inter-langue yoruba vers les articles français. Chaque lien correspond à une paire de titres arabe-français et yoruba-français. Certains titres sont composés de termes simples et d'autres sont composés de termes composés de plusieurs mots.

5.2 Alignement des mots

Dans le but d'avoir un lexique composé uniquement des termes simples, nous avons procéder à une étape d'alignement des mots.

Cette étape présente plusieurs difficultés dont : Premièrement, les alignements ne sont pas nécessairement contigus : deux mots consécutifs dans la phrase source peuvent être alignés avec deux mots arbitrairement distants de la phrase cible. On appelle ce phénomène distorsion. Deuxièmement, un mot en langue source peut être aligné à plusieurs mots en langue cible ; ce qui est défini en tant que fertilité.

[5] http://download.wikipedia.com/arwiki/20120114/ [consulté le 01/03/2012].

[6] http://dumps.wikimedia.org/yowiki/20120316/ [consulté le 15/03/2012].

Nous avons procédé à une étape d'alignement des mots des paires de titres en nous basant sur une approche statistique, nous avons utilisé les modèles IBM [1-5] (Brown *et al.*, 1993) combinés avec les modèles de Markov cachés HMM (Vogel *et al.*,1996) vu que ces modèles standard se sont avérés efficaces dans les travaux d'alignement de mots.

Les modèles IBM sont des modèles à base de mots, c'est-à-dire que l'unité de traduction qui apparaît dans les lois de probabilité est le mot.

Les cinq modèles IBM permettent d'estimer les probabilités *P(fr |ar)* et *P(fr |yo)* de façon itérative, tel que *fr* est un mot français, *ar* est un mot arabe et *yo* est un mot yoruba. Chaque modèle s'appuie sur les paramètres estimés par le modèle le précédant et prend en compte de nouvelles caractéristiques telles que la distorsion, la fertilité, etc.

Le modèle de Markov caché (nommé usuellement HMM) (Vogel *et al.*, 1996) est une amélioration du modèle IBM2. Il modélise explicitement la distance entre l'alignement du mot courant et l'alignement du mot précédent.

Nous avons utilisé l'outil open source Giza++ (Och et Ney, 2003) qui implémente ces modèles pour l'alignement des mots et nous avons extrait les traductions candidates à partir d'une table de traductions créée par Giza++. Chaque ligne de cette table contient un mot en langue arabe (*ar*) (respectivement yoruba (*yo*)), une traduction candidate (*fr*) et un score qui calcule la probabilité de traduction *P*(fr|ar) (resp. yoruba *P(fr|yo)*).

Après l'étape d'alignement, nous avons extrait 65 049 mots arabes et 155 348 paires de traductions candidates en français. En ce qui concerne le lexique yoruba-français, nous avons extrait 11 235 mots yoruba et 20 089 paires de traductions candidates en français.

Afin d'améliorer la qualité de nos lexiques, nous avons procédé à une étape de filtrage qui élimine les traductions candidates ayant un score inférieur à un seuil.

تد شد تت	Flou	1.0000000
تد شد تت	Diffusion	0.1666667
ل لرجال	Équipes	0.1250000
ل لرجال	féminin	0.0067568
ل لرجال	masculin	0.6690141

FIGURE 1 – Extrait de la table de traduction *ar-fr*

Rómù	Rome	0.7500
Rómù	romaine	0.33333
aládánidá	naturelles	1.00000
Àwùjọ	Société	0.66666
Àwùjọ	Communauté	0.20000

FIGURE 2 – Extrait de la table de traduction *yo-fr*

6 Evaluation

Puisque notre intérêt est centré sur les liens inter-langues de Wikipédia, les lexiques extraits ne contiennent pas des verbes.

Nous avons évalué, manuellement, la qualité de notre lexique bilingue en calculant la mesure de précision et en se référant à un expert.

$$precision = \frac{\text{nombre de traductions extraites correctes}}{\text{nombre de traductions extraites}}$$

Nous avons calculé la précision en se basant sur les traductions candidates de 50 mots arabes et yoruba et nous avons fait varier le seuil de 0 à 1 pour en identifier la valeur optimale en fonction de la précision.

La figure 3 présente les valeurs de précision des deux lexiques en variant le seuil.

Remarquons qu'en augmentant le seuil, la précision est améliorée. Sa valeur passe de 0.46 (avec un seuil égale 0) à 0.74 (quand le seuil égale à 1) pour le lexique yoruba-français et de 0.22 à 0.75 pour le lexique arabe-français.

La figure 4 montre que la couverture du lexique français-yoruba et presque stable, elle varie entre 14045 (quand le seuil égale à 0) et 11184 (quand le seuil égale à 1). Ces valeurs sont très inférieures par rapport à celles du lexique arabe-français, ceci est dû principalement au faible nombre des articles Wikipèdia yoruba.

La figure 3 montre que les meilleures valeurs de précision sont atteintes à partir d'un seuil égal à 0.6 pour le lexique arabe-français. Mais, remarquons dans la figure 4, qu'à partir de ce seuil, la couverture du lexique est affaiblie. Ceci est expliqué par le fait que plusieurs fausses traductions ont été éliminées à partir de ce seuil.

Les erreurs du lexique yoruba-français sont dues principalement au fait que certains titres wikipèdia sont introduits en anglais (ex. density/densité) et aux erreurs d'alignements (ex. Tanaka/Giichi).

Les erreurs de traduction du lexique arabe-français sont dues principalement au fait que certains titres arabes sont introduits en langue autre que l'arabe (ex. cv/cv), en majorité en langue anglaise. Certaines traductions candidates sont des translitérations et pas des traductions (ex. انتفاضة/Intifada). Aussi, nous avons détecté des erreurs d'alignement (ex. ذ ف س ية/diagnostique). D'autres erreurs sont dues au fait que les paires de titres des articles ne sont pas des traductions précises mais il s'agit juste de la même notion (ex. عيد/Noël).

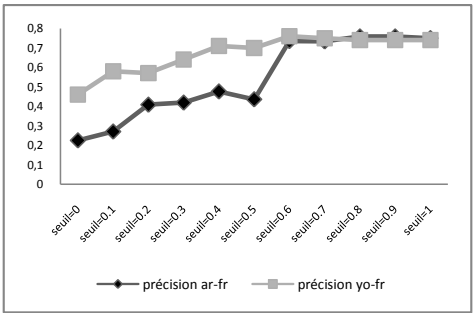

FIGURE 3 –Variation de la précision des lexiques *yo-fr* et *ar-fr* selon le seuil

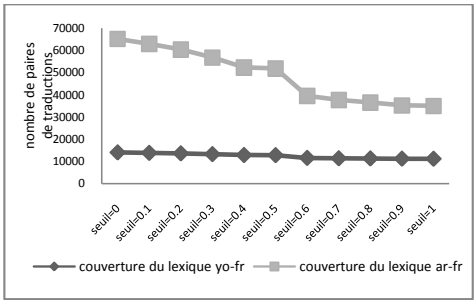

FIGURE 4 – Variation de la couverture des lexiques *yo-fr* et *ar-fr* selon le seuil

7 Conclusion

L'exploitation de Wikipédia pour la construction de ressources linguistiques multilingues fait l'objet de plusieurs travaux de recherches, comme la construction des corpus parallèles, des lexiques multilingues et des ontologies multilingues.

Dans cet article, nous avons décrit notre travail préliminaire d'extraction de lexiques (arabe-français et yoruba-français) à partir de Wikipédia. En effet, notre but majeur est d'exploiter Wikipédia en tant que corpus comparable pour la traduction automatique statistique.

La méthode que nous proposons est efficace malgré sa simplicité. Il s'agit d'extraire les titres arabes, yorubas et français des articles de Wikipédia, en se basant sur les liens inter-langues puis d'aligner les mots de ces titres en se basant sur une approche statistique. Nous avons atteint des valeurs de précision et de couverture encourageantes qui dépassent respectivement 0.7 et 60 000 paires de traductions pour le lexique arabe-français et 0.7 et 14 000 paires de traductions pour le lexique yoruba-français.

Comme travaux futurs, nous envisageons d'élargir la couverture de nos lexiques en exploitant d'autres liens Wikipédia comme les redirections et les liens inter-wiki. Nous envisageons aussi d'utiliser ces lexiques pour l'extraction des corpus parallèles (arabe-français et yoruba-français) à partir de Wikipédia. Ces corpus seront utilisés au niveau de l'apprentissage des systèmes de traduction automatique statistique arabe-français et yoruba-français.

Références

ADAFRE, S. F. ET DE RIJKE, M. (2006). Finding Similar Sentences across Multiple Languages in Wikipedia. *In Proceedings of the EACL Workshop on NEW TEXT Wikis and blogs and other dynamic text sources*, pages 62–69.

BOUMA, G., FAHMI, I., MUR, J., G. VAN NOORD, VAN DER, L., ET TIEDEMANN, J. (2006). Using Syntactic Knowledge for QA. *In Working Notes for the Cross Language Evaluation Forum Workshop*.

BROWN PETER, F., PIETRA, V. J., PIETRA, S. A., ET MERCER, R. L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. IBM T.J. Watson Research Center, pages 264-311.

DECLERCK, T., PEREZ, A. G., VELA, O., , Z., ET MANZANO-MACHO, D. (2006). Multilingual Lexical Semantic Resources for Ontology Translation. *In Proceedings of International Conference on Language Ressources and Evaluation* (LREC), pages 1492 – 1495.

ERDMANN, M., NAKAYAMA, K., HARA, T. ET NISHIO, S. (2008). A bilingual dictionary extracted from the wikipedia link structure. *In Proceedings of International Conference on Database Systems for Advanced Applications (DASFAA) Demonstration Track*, pages 380-392.

ERDMANN, M. (2008). Extraction of Bilingual Terminology from the Link Structure of Wikipedia. MSc. Thesis, Graduate School of Information Science and Engineering, Osaka University.

GREFENSTETTE, G. (1998). The Problem of Cross-language Information Retrieval. Cross-language Information Retrieval. Kluwer Academic Publishers.

HAZEM, A., MORIN, E. ET SEBASTIAN P. S. (2011). Bilingual Lexicon Extraction from Comparable Corpora as Metasearch. *In Proceedings of the 4th Workshop on Building and*

Using Comparable Corpora, pages 35–43, 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon.

MORIN, E. (2007). Synergie des approches et des ressources déployées pur le traitement de l'écrit. Ph.D. thesis, Habilitation à Diriger les Recherches, Université de Nantes.

MORIN, E. ET DAILLE, B. (2004). Extraction de terminologies bilingues à partir de corpus comparables d'un domaine spécialisé. *Traitement Automatique des Langues (TAL)*, pages 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*, pages 27–34.

OCH, F.J. ET NEY, H. (2003). A systematic comparison of various statistical alignment models. Computational Linguistics, pages 19–51, March.

OTERO, PABLO G. (2007). Learning bilingual lexicons from comparable english and spanish corpora. *In Proceedings of Machine Translation Summit XI*, pages 191–198.

SADAT, F., YOSHIKAWA, M. ET UEMURA, S. 2003. Bilingual terminology acquisition from comparable corpora and phrasal translation to cross-language information retrieval. *In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume*, pages 141–144. Association for Computational Linguistics.

SADAT, F. ET TERRASSA, A. (2010). Exploitation de Wikipédia pour l'Enrichissement et la Construction des Ressources Linguistiques. *TALN 2010*, Montréal.

VOGEL, S., NEY H. ET C. TILLMANN (1996). HMM-based word alignment in statistical translation. *In Preceding of the Conference on Computational Linguistics*, pages 836–841, Morristown, NJ, USA.