Bilingual Lexicon Extraction from Comparable Corpora Enhanced with Parallel Corpora

Emmanuel Morin, Emmanuel Ep Prochasson

Bilingual Lexicon Extraction from Comparable Corpora Enhanced with Parallel Corpora

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

Abstract

In this article, we present a simple and effective approach for extracting bilingual lexicon from comparable corpora enhanced with parallel corpora. We make use of structural characteristics of the documents comprising the comparable corpus to extract parallel sentences with a high degree of quality. We then use state-of-the-art techniques to build a specialized bilingual lexicon from these sentences and evaluate the contribution of this lexicon when added to the comparable corpus-based alignment technique. Finally, the value of this approach is demonstrated by the improvement of translation accuracy for medical words.

1 Introduction

Bilingual lexicons are important resources of many applications of natural language processing such as cross-language information retrieval or machine translation. These lexicons are traditionally extracted from bilingual corpora.

In this area, the main work involves parallel corpora, i.e. a corpus that contains source texts and their translations. From sentence-to-sentence aligned corpora, symbolic (Carl and Langlais, 2002), statistical (Daille et al., 1994), or hybrid techniques (Gaussier and Langé, 1995) are used for word and expression alignments. However, despite good results in the compilation of bilingual lexicons, parallel corpora are rather scarce resources, especially for technical domains and for language pairs not involving English. For instance, current resources of parallel corpora are built from the proceedings of international institutions such as the European Union (11 languages) or the United Nations (6 languages), bilingual countries such as Canada (English and French languages), or bilingual regions such as Hong Kong (Chinese and English languages).

For these reasons, research in bilingual lexicon extraction is focused on another kind of bilingual corpora. These corpora, known as comparable corpora, are comprised of texts sharing common features such as domain, genre, register, sampling period, etc. without having a source text-target text relationship. Although the building of comparable corpora is easier than the building of parallel corpora, the results obtained thus far on comparable corpora are contrasted. For instance, good results are obtained from large corpora — several million words — for which the accuracy of the proposed translation is between 76% (Fung, 1998) and 89% (Rapp, 1999) for the first 20 candidates. (Cao and Li, 2002) have achieved 91% accuracy for the top three candidates using the Web as a comparable corpus. But for technical domains, for which large corpora are not available, the results obtained, even though encouraging, are not completely satisfactory yet. For instance, (Déjean et al., 2002) obtained a precision of 44% and 57% for the first 10 and 20 candidates in a 100,000-word medical corpus, and 35% and 42% in a multi-domain 8 million-word corpus. For French/English single words, (Chiao and Zweigenbaum, 2002) using a medical corpus of 1.2 million words, obtained a precision of about 50% and 60% for the top 10 and top 20 candidates. (Morin et al., 2007) obtained a precision of 51% and 60% for the top 10 and 20 candidates in a 1.5

Proceedings of the 4th Workshop on Building and Using Comparable Corpora, pages 27–34,
49th Annual Meeting of the Association for Computational Linguistics,
Portland, Oregon, 24 June 2011.

million-word French-Japanese diabetes corpus.

The above work in bilingual lexicon extraction from comparable corpora relies on the assumption that words which have the same meaning in different languages tend to appear in the same lexical contexts (Fung, 1998; Rapp, 1999). Based on this assumption, a standard approach consists of building context vectors for each word of the source and target languages. The candidate translations for a particular word are obtained by comparing the translated source context vector with all target context vectors. In this approach, the translation of the words of the source context vectors depends on the coverage of the bilingual dictionary vis-à-vis the corpus. This aspect can be a potential problem if too few corpus words are found in the bilingual dictionary (Chiao and Zweigenbaum, 2003; Déjean et al., 2002).

In this article, we want to show how this problem can be partially circumvented by combining a general bilingual dictionary with a specialized bilingual dictionary based on a parallel corpus extracted through mining of the comparable corpus. In the same way that recent works in Statistical Machine Translation (SMT) mines comparable corpora to discover parallel sentences (Resnik and Smith, 2003; Yang and Li, 2003; Munteanu and Marcu, 2005; Abdul-Rauf and Schwenk, 2009, among others), this work contributes to the bridging of the gap between comparable and parallel corpora by offering a framework for bilingual lexicon extraction from comparable corpus with the help of parallel corpus-based pairs of terms.

The remainder of this article is organized as follows. In Section 2, we first present the method for bilingual lexicon extraction from comparable corpora enhanced with parallel corpora and the associated system architecture. We then quantify and analyse in Section 3 the performance improvement of our method on a medical comparable corpora when used to extract specialized bilingual lexicon. Finally, in Section 4, we discuss the present study and present our conclusions.

2 System Architecture

The overall architecture of the system for lexical alignment is shown in Figure 1 and comprises parallel corpus- and comparable corpus-based alignments. Starting from a comparable corpus harvested from the web, we first propose to extract parallel sentences based on the structural characteristics of the documents harvested. These parallel sentences are then used to build a bilingual lexicon through a tool dedicated to bilingual lexicon extraction. Finally, this bilingual lexicon is used to perform the comparable corpus-based alignment. For a word to be translated, the output of the system is a ranked list of candidate translations.

2.1 Extracting Parallel Sentences from Comparable Corpora

Parallel sentence extraction from comparable corpora has been studied by a number of researchers (Ma and Liberman, 1999; Chen and Nie, 2000; Resnik and Smith, 2003; Yang and Li, 2003; Fung and Cheung, 2004; Munteanu and Marcu, 2005; Abdul-Rauf and Schwenk, 2009, among others) and several systems have been developed such as BITS (Bilingual Internet Test Search) (Ma and Liberman, 1999), PTMiner (Parallel Text Miner) (Chen and Nie, 2000), and STRAND (Structural Translation Recognition for Acquiring Natural Data) (Resnik and Smith, 2003). Their work relies on the observation that a collection of texts in different languages composed independently and based on sharing common features such as content, domain, genre, register, sampling period, etc. contains probably some sentences with a source text-target text relationship. Based on this observation, dynamic programming (Yang and Li, 2003), similarity measures such as Cosine (Fung and Cheung, 2004) or word and translation error ratios (Abdul-Rauf and Schwenk, 2009), or maximum entropy classifier (Munteanu and Marcu, 2005) are used for discovering parallel sentences.

Although our purpose is similar to these works, the amount of data required by these techniques makes them ineffective when applied to specialized comparable corpora used to discover parallel sentences. In addition, the focus of this paper is not to propose a new technique for this task but to study how parallel sentences extracted from a comparable corpus can improve the quality of the candidate translations. For theses reasons, we propose to make use of structural characteristics of the documents comprising the comparable corpus to extract auto-

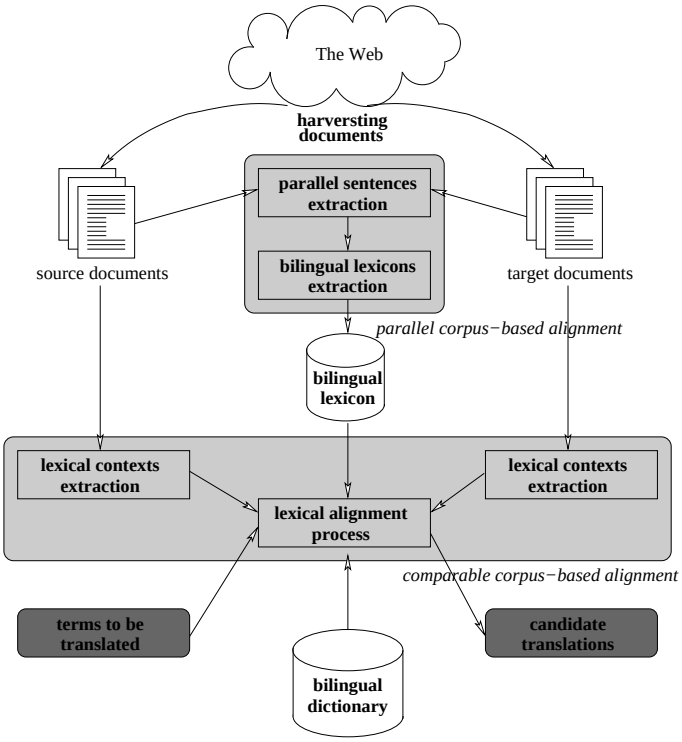

Figure 1: Overview of the system for lexical alignment

matically parallel sentences.

In fact, specialized comparable corpora are generally constructed via the consultation of specialized Web portals. For instance, (Chiao and Zweigenbaum, 2002) use CISMeF[1] for building the French part of their comparable corpora and CliniWeb[2] for the English part, and (Déjean and Gaussier, 2002) use documents extracted from MEDLINE[3] to build a German/English comparable corpus. Consequently, the documents collected through these portals are often scientific papers. Moreover, when the language of these papers is not the English, the paper usually comprises an abstract, keywords and title in the native language and their translations in the English language. These characteristics of scientific paper is useful for the efficient extraction of parallel sentences or word translations from the documents forming a specialized comparable corpus for which one part will inevitably be in English.

In this study, the documents comprising the French/English specialized comparable corpus were taken from the medical domain within the subdomain of 'breast cancer'. These documents have been automatically selected from the Elsevier website[4] among the articles published between 2001 and 2008 for which the title or the keywords of the articles contain the multi-word term 'cancer du sein' in French and 'breast cancer' in English. We thus collected 130 documents in French and 118 in English and about 530,000 words for each language. Since the 130 French documents previously collected are scientific papers, each document contains a French abstract which is accompanied by its English translation. We exploit this structural characteristic of the French documents in order to build a small specialized parallel corpus directly correlated to the subdomain of 'breast cancer' involved in the comparable corpus.

2.2 Parallel Corpus-Based Alignment

We use the *Uplug*[5] collection of tools for alignment (Tiedemann, 2003) to extract translations from our

[1] http://www.chu-rouen.fr/cismef/
[2] http://www.ohsu.edu/cliniweb/
[3] http://www.ncbi.nlm.nih.gov/PubMed
[4] http://www.elsevier.com
[5] http://stp.ling.uu.se/cgi-bin/joerg/Uplug

specialized parallel corpus. The output of such a tool is a list of aligned *parts of sentences*, that has to be post-process and filtered in our case. We clean the alignment with a simple yet efficient method in order to obtain only word translations. We associate every source word from a source sequence with every target word from the target sequence. As an example, *uplug* efficiently aligns the English word *breast cancer* with the French word *cancer du sein* (the data are described in Section 3.1). We obtain the following lexical alignment:

- cancer (fr) → (en) breast, cancer
- du (fr) → (en) breast, cancer
- sein (fr) → (en) breast, cancer

With more occurrences of the French word *cancer*, we are able to align it with the English words {breast, cancer, cancer, cancer, the, of, breast, cancer}. We can then filter such a list by counting the translation candidates. In the previous example, we obtain: cancer (fr) → breast/2, the /1, of/1, cancer/4. The English word *cancer* is here the best match for the French word *cancer*. In many cases, only one alignment is obtained. For example, there is only one occurrence of the French word *chromosome*, aligned with the English word *chromosome*.

In order to filter translation candidates, we keep 1:1 candidates if their frequencies are comparable in the original corpus. We keep the most frequent translation candidates (in the previous example, *cancer*) if their frequencies in the corpus are also comparable. This in-corpus frequency constraint is useful for discarding candidates that appear in many alignments (such as functional words). The criterion for frequency acceptability is:

$$min(f_1, f_2)/max(f_1, f_2) > 2/3$$

with f_1 and f_2 the frequency of words to be aligned in the parallel corpus.

By this way, we build a French/English specialized bilingual lexicon from the parallel corpus. This lexicon, called breast cancer dictionary (BC dictionary) in the remainder of this article, is composed of 549 French/English single words.

2.3 Comparable Corpus-Based Alignment

The comparable corpus-based alignment relies on the simple observation that a word and its translation tend to appear in the same lexical contexts. Based on this observation, the alignment method, known as the *standard approach*, builds context vectors in the source and the target languages where each vector element represents a word which occurs within the window of the word to be translated (for instance a seven-word window approximates syntactic dependencies). In order to emphasize significant words in the context vector and to reduce word-frequency effects, the context vectors are normalized according to association measures. Then, the translation is obtained by comparing the source context vector to each translation candidate vector after having translated each element of the source vector with a general dictionary.

The implementation of this approach can be carried out by applying the four following steps (Fung, 1998; Rapp, 1999):

1. We collect all the lexical units in the context of each lexical unit i and count their occurrence frequency in a window of n words around i. For each lexical unit i of the source and the target languages, we obtain a context vector v_i which gathers the set of co-occurrence units j associated with the number of times that j and i occur together $occ(i, j)$. In order to identify specific words in the lexical context and to reduce word-frequency effects, we normalize context vectors using an association score such as Mutual Information (MI) or Log-likelihood, as shown in equations 1 and 2 and in Table 1 (where $N = a + b + c + d$).

2. Using a bilingual dictionary, we translate the lexical units of the source context vector. If the bilingual dictionary provides several translations for a lexical unit, we consider all of them but weight the different translations according to their frequency in the target language.

3. For a lexical unit to be translated, we compute the similarity between the translated context vector and all target vectors through vector distance measures such as Cosine or Weighted

Jaccard (WJ) (see equations 3 and 4 where $assoc_j^i$ stands for "association score").

4. The candidate translations of a lexical unit are the target lexical units ranked following the similarity score.

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

Table 1: Contingency table

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

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

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

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

This approach is sensitive to the choice of parameters such as the size of the context, the choice of the association and similarity measures. The most complete study about the influence of these parameters on the quality of bilingual alignment has been carried out by Laroche and Langlais (2010).

3 Experiments and Results

In the previous section, we have introduced our comparable corpus and described the method dedicated to bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. In this section, we then quantify and analyse the performance improvement of our method on a medical comparable corpus when used to extract specialized bilingual lexicon.

3.1 Experimental Test bed

The documents comprising the French/English specialized comparable corpus have been normalised through the following linguistic pre-processing steps: tokenisation, part-of-speech tagging, and lemmatisation. Next, the function words were removed and the words occurring less than twice in the French and the English parts were discarded. Finally, the comparable corpus comprised about 7,400 distinct words in French and 8,200 in English.

In this study, we used four types of bilingual dictionary: i) the Wiktionary[6] free-content multilingual dictionary, ii) the ELRA-M0033[7] professional French/English bilingual dictionary, iii) the MeSH[8] metha-thesaurus, and iv) the BC dictionary (see Section 2.2). Table 2 shows the main features of the dictionaries, namely: the number of distinct French single words in the dictionary (# SWs dico.), the number of distinct French single words in the dictionary after projection on the French part of the comparable corpus (# SWs corpus), and the number of translations per entry in the dictionary (# TPE). For instance, 42% of the French context vectors could be translated with the Wiktionary (3,099/7,400).

Table 2: Main features of the French/English dictionaries

Name	# SWs dict.	# SWs corpus	# TPE
Wiktionary	20,317	3,099	1.8
ELRA	50,330	4,567	2.8
MeSH	18,972	833	1.6
BC	549	549	1.0

In bilingual terminology extraction from specialized comparable corpora, the terminology reference list required to evaluate the performance of the alignment programs are often composed of 100 single-word terms (SWTs) (180 SWTs in (Déjean and Gaussier, 2002), 95 SWTs in (Chiao and Zweigenbaum, 2002), and 100 SWTs in (Daille and Morin, 2005)). To build our reference list, we selected 400 French/English SWTs from the UMLS[9]

[6] http://www.wiktionary.org/
[7] http://www.elra.info/
[8] http://www.ncbi.nlm.nih.gov/mesh
[9] http://www.nlm.nih.gov/research/umls

meta-thesaurus and the *Grand dictionnaire terminologique*[10]. We kept only the French/English pair of SWTs which occur more than five times in each part of the comparable corpus. As a result of filtering, 122 French/English SWTs were extracted.

3.2 Experimental Results

In order to evaluate the influence of the parallel corpus-based bilingual lexicon induced from the comparable corpus on the quality of comparable corpus based-bilingual terminology extraction, four experiments were carried out. For each experiment, we change the bilingual dictionary required for the translation phase of the standard approach (see Section 2.3):

1. The first experiment uses only the Wiktionary. Since the coverage of the Wiktionary from the comparable corpus is small (see Table 2), the results obtained with this dictionary yield a lower boundary.

2. The second experiment uses the Wiktionary added to the BC dictionary. This experiment attempts to verify the hypothesis of this study.

3. The third experiment uses the Wiktionary added to the MeSH thesaurus. This experiment attempts to determine whether a specialised dictionary (in this case the MeSH) would be more suitable than a specialized bilingual dictionary (in this case the BC dictionary) directly extracted from the corpus.

4. The last experiment uses only the ELRA dictionary. Since the coverage of the ELRA dictionary from the comparable corpus is the best (see Table 2), the results obtained with this one yield a higher boundary.

Table 3 shows the coverage of the four bilingual lexical resources involved in the previous experiments in the comparable corpus. The first column indicates the number of single words belonging to a dictionary found in the comparable corpus (# SWs corpus). The other column indicates the coverage of each dictionary in the ELRA dictionary (Coverage ELRA). Here, 98.9% of the single words belonging to the Wiktionary are included in the ELRA dictionary whereas less than 95% of the single words belonging to the Wiktionary+BC and Wiktionary+MeSH dictionaries are included in the ELRA dictionary. Moreover, the MeSH and BC dictionaries are two rather distinct specialized resources since they have only 117 single words in common.

Table 3: Coverage of the bilingual lexical resources in the comparable corpus

Name	# SWs corpus	Coverage ELRA
Wiktionary	3,099	98.8%
Wiktionary + BC	3,326	94.8%
Wiktionary + MeSH	3,465	94.9%
ELRA	4,567	100%

In the experiments reported here, the size of the context window n was set to 3 (i.e. a seven-word window), the association measure was the Mutual Information and the distance measure the Cosine (see Section 2.3). Other combinations of parameters were assessed but the previous parameters turned out to give the best performance.

Figure 2 summarises the results obtained for the four experiments for the terms belonging to the reference list according to the French to English direction. As one could expect, the precision of the result obtained with the ELRA dictionary is the best and the precision obtained with the Wiktionary is the lowest. For instance, the ELRA dictionary improves the precision of the Wiktionary by about 14 points for the Top 10 and 9 points for the top 20. These results confirm that the coverage of the dictionary is an important factor in the quality of the results obtained. Now, when you add the BC dictionary to the Wiktionary, the results obtained are also much better than those obtained with the Wiktionary alone and very similar to those obtained with the ELRA dictionary alone (without taking into account the top 5). This result suggests that a standard general language dictionary enriched with a small specialized dictionary can replace a large general language dictionary.

Furthermore, this combination is more interesting than the combination of the MeSH dictionary with

[10] `http://www.granddictionnaire.com/`

the Wiktionary. Since the BC dictionary is induced from the corpus, this dictionary is directly correlated to the theme of breast cancer involved in the corpus. Consequently the BC dictionary is more suitable than the MeSH dictionary i) even if the MeSH dictionary specializes in the medical domain and ii) even if more words in the comparable corpus are found in the MeSH dictionary than in the BC dictionary.

This last observation should make us relativize the claim: the greater the number of context vector elements that are translated, the more discriminating the context vector will be for selecting translations in the target language. We must also take into account the specificity of the context vector elements in accordance with the thematic of the documents making up the corpus studied in order to improve bilingual lexicon extraction from specialized comparable corpora.

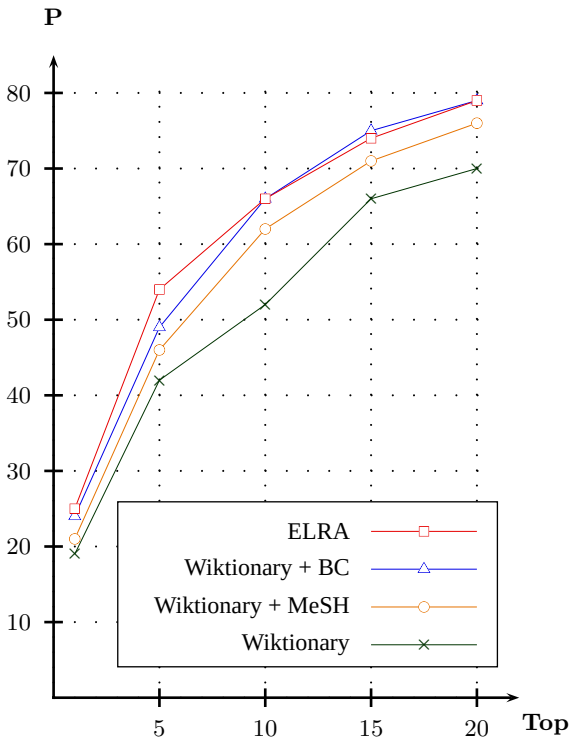

Figure 2: Precision of translations found according to the rank

4 Conclusion and Discussion

In this article, we have shown how the quality of bilingual lexicon extraction from comparable corpora could be improved with a small specialized bilingual lexicon induced through parallel sentences included in the comparable corpus. We have evaluated the performance improvement of our method on a French/English comparable corpus within the sub-domain of breast cancer in the medical domain. Our experimental results show that this simple bilingual lexicon, when combined with a general dictionary, helps improve the accuracy of single word alignments by about 14 points for the Top 10 and 9 points for the top 20. Even though we focus here on one structural characteristic (i.e. the abstracts) of the documents comprising the comparable corpus to discover parallel sentences and induced bilingual lexicon, the method could be easily applied to other comparable corpora for which a bilingual dictionary can be extracted by using other characteristics such as the presence of parallel segments or paraphrases in the documents making up the comparable corpus.

Acknowledgments

The research leading to these results has received funding from the European Communitys Seventh Framework Programme (FP7/2007-2013) under Grant Agreement no 248005.

References

Sadaf Abdul-Rauf and Holger Schwenk. 2009. On the use of comparable corpora to improve SMT performance. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL'09)*, pages 16–23, Athens, Greece.

Yunbo Cao and Hang Li. 2002. Base Noun Phrase Translation Using Web Data and the EM Algorithm. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 127–133, Tapei, Taiwan.

Michael Carl and Philippe Langlais. 2002. An Intelligent Terminology Database as a Pre-processor for Statistical Machine Translation. In L.-F. Chien, B. Daille, K. Kageura, and H. Nakagawa, editors, *Proceedings of the 2nd International Workshop on Computational Terminology (COMPUTERM'02)*, pages 15–21, Tapei, Taiwan.

Jiang Chen and Jian-Yun Nie. 2000. Parallel Web Text Mining for Cross-Language Information Retrieval. In *Proceedings of Recherche d'Information Assistée par Ordinateur (RIAO'00)*, pages 62–77, Paris, France.

Yun-Chuang Chiao and Pierre Zweigenbaum. 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, Tapei, Taiwan.

Yun-Chuang Chiao and Pierre Zweigenbaum. 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, and P. Ruch, editors, *The New Navigators: from Professionals to Patients, Actes Medical Informatics Europe*, volume 95 of *Studies in Health Technology and Informatics*, pages 397–402.

Béatrice Daille and Emmanuel Morin. 2005. French-English Terminology Extraction from Comparable Corpora. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCLNP'05)*, pages 707–718, Jeju Island, Korea.

Béatrice Daille, Éric Gaussier, and Jean-Marc Langé. 1994. Towards Automatic Extraction of Monolingual and Bilingual Terminology. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, volume I, pages 515–521, Kyoto, Japan.

Hervé Déjean and Éric Gaussier. 2002. Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica, Alignement lexical dans les corpus multilingues*, pages 1–22.

Hervé Déjean, Fatia Sadat, and Éric Gaussier. 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)*, pages 218–224, Tapei, Taiwan.

Pascale Fung and Percy Cheung. 2004. Mining Very-Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM. In D. Lin and D. Wu, editors, *Proceedings of Empirical Methods on Natural Language Processing (EMNLP'04)*, pages 57–63, Barcelona, Spain.

Pascale Fung. 1998. A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-parallel Corpora. In D. Farwell, L. Gerber, and E. Hovy, editors, *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA'98)*, pages 1–16, Langhorne, PA, USA.

Éric Gaussier and Jean-Marc Langé. 1995. Modèles statistiques pour l'extraction de lexiques bilingues. *Traitement Automatique des Langues (TAL)*, 36(1–2):133–155.

Audrey Laroche and Philippe Langlais. 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.

Xiaoyi Ma and Mark Y. Liberman. 1999. Bits: A Method for Bilingual Text Search over the Web. In *Proceedings of Machine Translation Summit VII*, Kent Ridge Digital Labs, National University of Singapore.

Emmanuel Morin, Béatrice Daille, Koichi Takeuchi, and Kyo Kageura. 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)*, pages 664–671, Prague, Czech Republic.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics*, 31(4):477–504.

Reinhard Rapp. 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.

Philip Resnik and Noah A. Smith. 2003. The Web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

J. Tiedemann. 2003. *Recycling Translations - Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing*. Ph.D. thesis, Studia Linguistica Upsaliensia 1.

Christopher C. Yang and Kar Wing Li. 2003. Automatic construction of English/Chinese parallel corpora. *Journal of the American Society for Information Science and Technology*, 54(8):730–742.