Building and using comparable corpora for domain-specific bilingual lexicon extraction

Darja Fišer
University of Ljubljana, Faculty of Arts,
Deparment of Translation
Aškerčeva 2
Ljubljana, Slovenia
darja.fiser@ff.uni-lj.si

Nikola Ljubešić
University of Zagreb, Faculty of Humanities
and Social Sciences
Ivana Lučića 3
Zagreb, Croatia
nikola.ljubesic@ffzg.hr

Špela Vintar
University of Ljubljana, Faculty of Arts,
Deparment of Translation
Aškerčeva 2
Ljubljana, Slovenia
spela.vintar@ff.uni-lj.si

Senja Pollak
University of Ljubljana, Faculty of Arts,
Deparment of Translation
Aškerčeva 2
Ljubljana, Slovenia
senja.pollak@ff.uni-lj.si

Abstract

This paper presents a series of experiments aimed at inducing and evaluating domain-specific bilingual lexica from comparable corpora. First, a small English-Slovene comparable corpus from health magazines was manually constructed and then used to compile a large comparable corpus on health-related topics from web corpora. Next, a bilingual lexicon for the domain was extracted from the corpus by comparing context vectors in the two languages. Evaluation of the results shows that a 2-way translation of context vectors significantly improves precision of the extracted translation equivalents. We also show that it is sufficient to increase the corpus for one language in order to obtain a higher recall, and that the increase of the number of new words is linear in the size of the corpus. Finally, we demonstrate that by lowering the frequency threshold for context vectors, the drop in precision is much slower than the increase of recall.

1 Introduction

Research into using comparable corpora in NLP has gained momentum in the past decade largely due to limited availability of parallel data for many language pairs and domains. As an alternative to already established parallel approaches (e.g. Och 2000, Tiedemann 2005) the comparable corpus-based approach relies on texts in two or more languages which are not parallel but nevertheless share several parameters, such as topic, time of publication and communicative goal (Fung 1998, Rapp 1999). The main advantage of this approach is the simpler, faster and more time efficient compilation of comparable corpora, especially from the rich web data (Xiao & McEnery 2006).

In this paper we describe the compilation process of a large comparable corpus of texts on health-related topics for Slovene and English that were published on the web. Then we report on a set of experiments we conducted in order to automatically extract translation equivalents for terms from the health domain. The parameters we tested and analysed are: 1- and 2-way translations of context vectors with a seed lexicon, the size of the corpus used for bilingual lexicon extraction, and the word frequency threshold for vector construction. The main contribution of this paper is a much-desired language- and domain-independent approach to bootstrapping bilingual lexica with minimal manual intervention as well as minimal reliance on the existing linguistic resources.

The paper is structured as follows: in the next section we give an overview of previous work relevant for our research. In Section 3 we present the construction of the corpus. Section 4 describes

Proceedings of the 4th Workshop on Building and Using Comparable Corpora, pages 19–26,
49th Annual Meeting of the Association for Computational Linguistics,
Portland, Oregon, 24 June 2011.

the experiments for bilingual lexicon extraction the results of which are reported, evaluated and discussed in Section 5. We conclude the paper with final remarks and ideas for future work.

2 Related work

Bilingual lexica are the key component of all cross-lingual NLP applications and their compilation remains a major bottleneck in computational linguistics. In this paper we follow the line of research that was inspired by Fung (1998) and Rapp (1999) who showed that texts do not need to be parallel in order to extract translation equivalents from them. Instead, their main assumption is that the term and its translation appear in similar contexts anyhow. The task of finding the appropriate translation equivalent of a term is therefore reduced to finding the word in the target language whose context vector is most similar to the source term's context vector based on their occurrence in a comparable corpus. This is basically a three-step procedure:

(1) Building context vectors. When representing a word's context, some approaches look at a simple co-occurrence window of a certain size while others include some syntactic information as well. For example, Otero (2007) proposes binary dependences previously extracted from a parallel corpus, while Yu and Tsujii (2009) use dependency parsers and Marsi and Krahmer (2010) use syntactic trees. Instead of context windows, Shao and Ng (2004) use language models. Next, words in co-occurrence vectors can be represented as binary features, by term frequency or weighted by different association measures, such as TF-IDF (Fung, 1998), PMI (Shezaf and Rappoport, 2010) or, one of the most popular, the log likelihood score. Approaches also exist that weigh co-occurrence terms differently if they appear closer to or further from the nucleus word in the context (e.g. Saralegi et al., 2008).
(2) Translating context vectors. Finding the most similar context vectors in the source and target language is not straightforward because a direct comparison of vectors in two different languages is not possible. This is why most researchers first translate features of source context vectors with machine-readable dictionaries and compute similarity measures on those. Koehn and Knight (2002) construct the seed dictionary automatically based on identical spelled words in the two languages. Similarly, cognate detection is used by Saralegi et al. (2008) by computing the longest common subsequence ratio. Déjean et al. (2005), on the other hand, use a bilingual thesaurus instead of a bilingual dictionary.
(3) Selecting translation candidates. After source context vectors have been translated, they are ready to be compared to the target context vectors. A number of different vector similarity measures have been investigated. Rapp (1999) applies city-block metric, while Fung (1998) works with cosine similarity. Recent work often uses Jaccard index or Dice coefficient (Saralegi et al., 2008). In addition, some approaches include a subsequent re-ranking of translation candidates based on cognates detection (e.g. Shao and Ng, 2004).

3 Corpus construction

A common scenario in the NLP community is a project on a specific language pair in a new domain for which no ready-made resources are available. This is why we propose an approach that takes advantage of the existing general resources, which are then fine-tuned and enriched to be better suited for the task at hand. In this section we describe the construction of a domain-specific corpus that we use for extraction of translation equivalents in the second part of the paper.

3.1 Initial corpus

We start with a small part of the Slovene PoS tagged and lemmatized reference corpus FidaPLUS (Arhar et al., 2007) that contains collections of articles from the monthly health and lifestyle magazine called *Zdravje*[1], which were published between 2003 and 2005 and contain 1 million words.
We collected the same amount of text from the most recent issues of the Health Magazine, which is a similar magazine for the English-speaking readers. We PoS-tagged and lemmatized the English part of the corpus with the TreeTagger (Schmid, 1994).

[1] http://www.zdravje.si/category/revija-zdravje [1.4.2010]

3.2 Corpus extension

We then extended the initial corpus automatically from the 2 billion-word ukWaC (Ferraresi et al., 2008) and the 380 million-word slWaC (Ljubešić and Erjavec, 2011), very large corpora that were constructed from the web by crawling the .uk and .si domain respectively.
We took into account all the documents from these two corpora that best fit the initial corpora by computing a similarity measure between models of each document and the initial corpus in the corresponding language. The models were built with content words lemmas as their parameters and TF-IDF values as the corresponding parameter values. The inverse document frequency was computed for every language on a newspaper domain of 20 million words. The similarity measure used for calculating the similarity between a document model and a corpus model was cosine with a similarity threshold of 0.2. This way, we were able to extend the Slovene part of the corpus from 1 to 6 million words and the English part to as much as 50 million words. We are aware of more complex methods for building comparable corpora, such as (Li and Gaussier, 2010), but the focus of this paper is on using comparable corpora collected from the web on the bilingual lexicon extraction task, and not the corpus extension method itself. Bilingual lexicon extraction from the extended corpus is described in the following section.

4 Bilingual lexicon extraction

In this section we describe the experiments we conducted in order to extract translation equivalents of key terms in the health domain. We ran a series of experiments in which we adjusted the following parameters:

(1) 1- and 2-way translation of context vectors with a seed dictionary;
(2) corpus size of the texts between the languages;
(3) the word frequency threshold for vector construction.

Although several parameters change in each run of the experiment, the basic algorithm for finding translation equivalents in comparable corpora is always the same:

(1) build context vectors for all unknown words in the source language that satisfy the minimum frequency criterion and translate the vectors with a seed dictionary;
(2) build context vectors for all candidate translations satisfying the frequency criterion in the target language;
(3) compute the similarity of all translated source vectors with the target vectors and rank translation candidates according to this score.

Previous research (Ljubešić et al., 2011) has shown that best results are achieved by using content words as features in context vectors and a context window of 7 with encoded position. The highest-scoring combination of vector association and similarity measures turned out to be Log Likelihood (Dunning, 1993) and Jensen-Shannon divergence (Lin, 1991), so we are using those throughout the experiments presented in this paper.

4.1 Translation of context vectors

In order to be able to compare two vectors in different languages, a seed dictionary to translate features in context vectors of source words is needed. We tested our approach with a 1-way translation of context features of English vectors into Slovene and a 2-way translation of the vectors from English into Slovene and vice versa where we then take the harmonic mean of the context similarity in both directions for every word pair.
A similar 2-way approach is described in (Chiao et al, 2004) with the difference that they average on rank values, not on similarity measures. An empirical comparison with their method is given in the automatic evaluation section.
A traditional general large-sized English-Slovene dictionary was used for the 1-way translation, which was then complemented with another general large-sized Slovene-English dictionary by the same author in the 2-way translation setting. Our technique relies on the assumption that additional linguistic knowledge is encoded in the independent dictionary in the opposite direction and was indirectly inspired by a common approach to filter out the noise in bilingual lexicon extraction from parallel corpora with source-to-target and target-to-source word-alignment.

Only content-word dictionary entries were taken into account. No multi-word entries were considered either. And, since we do not yet deal with polysemy at this stage of our research, we only extracted the first sense for each dictionary entry. The seed dictionaries we obtained in this way contained 41.405 entries (Eng-Slo) and 30.955 entries (Slo-Eng).

4.2 Corpus size

Next, we tested the impact of the extended corpus on the quality and quantity of the extracted translation equivalents by gradually increasing the size of the corpus from 1 to 6 million words.

Not only did we increase corpus size for each language equally, we also tested a much more realistic setting in which the amount of data available for one language is much higher than for the other, in our case English for which we were able to compile a 50 million word corpus, which is more than eight times more than for Slovene.

4.3 Word frequency threshold

Finally, we tested the precision and recall of the extracted lexica based on the minimum frequency of the words in the corpus from as high as 150 and down to 25 occurrences. This is an important parameter that shows the proportion of the corpus lexical inventory our method can capture and with which quality.

5 Evaluation of the results

At this stage of our research we have limited the experiments to nouns. This speeds up and simplifies our task but we believe it still gives an adequate insight into the usefulness of the approach for a particular domain since nouns carry the highest domain-specific terminological load.

5.1 Automatic evaluation

Automatic evaluation of the results was performed against a gold standard lexicon of health-related terms that was obtained from the top-ranking nouns in the English health domain model of the initial corpus and that at the same time appeared in the comprehensive dictionary of medical terms mediLexicon[2] and were missing from the general bilingual seed dictionary. The gold standard contains 360 English single-word terms with their translations into Slovene. If more than one translation variant is possible for a single English term, all variants appear in the gold standard and any of these translations suggested by the algorithm is considered as correct.

Below we present the results of three experiments that best demonstrate the performance and impact of the key parameters for bilingual lexicon extraction from comparable corpora that we were testing in this research. The evaluation measure for precision used throughout this research is mean reciprocal rank (Vorhees, 2001) on first ten translation candidates. Recall is calculated as the percentage of goldstandard entries we were able to calculate translation candidates for. Additionally, a global recall impact of our methods is shown as the overall number of entries for which we were able to calculate translation candidates. Unless stated otherwise, the frequency threshold for the generation of context vectors in the experiments was set to 50.

We begin with the results of 1- and 2-way context vector translations that we tested on the initial 1-million-word corpus we constructed from health magazines as well as on a corpus of the same size we extracted from the web. We compared the results of our method with that proposed in (Chiao et al, 2004) strengthening our claim that it is the additional information in the reverse dictionary that makes the significant impact, not the reversing itself.

As Table 1 shows, using two general dictionaries (2-way two dict) significantly improves the results as a new dictionary brings additional information. That it is the dictionary improving the results is proven by using just one, inverted dictionary in the 2-way manner, which produced worse results than the 1-way approach (2-way inverse dict). The approach of Chiao et al (2004) is also based on new dictionary knowledge since using only one inverted dictionary with their 2-way method yielded results that were almost identical to the 1-way computation. Using rank, not similarity score in averaging results proved to be a good approach (2-way Chiao two dict), but not as efficient as our approach which uses similarity scores (2-way two dict). Our approach yields higher precision and is also easier to compute. Namely, for every candidate pair only the reverse similarity score has

[2] http://www.medilexicon.com [1.4.2010]

to be computed, and not all similarity scores for every inverse pair to obtain a rank value.

Therefore, only the 2-way translation setting averaging on similarity scores is used in the rest of the experiments. It is interesting that the results on the web corpus have a higher precision but a lower recall (0.355 on the initial corpus and 0.198 on the web corpus). Higher precision can be explained with the domain modelling technique that was used to extract web data, which may have contributed to a terminologically more homogenous collection of documents in the health domain. On the other hand, the lower recall can be explained with the extracted web documents being less terminologically loaded than the initial corpus.

Corpus	1-way	2-way inverse dict	2-way Chiao two dict	2-way two dict
1 M initial	0.591	0.566	0.628	0.641
1 M web	0.626	0.610	0.705	0.710

Table 1: Precision regarding the corpus source and the translation method

The second parameter we tested in our experiments was the impact of corpus size on the quality and amount of the extracted translation equivalents. For the first 6 million words the Slovene and English parts of the corpus were enlarged in equal proportions and after that only the English part of the corpus was increased up to 18 million words.

Corpus size	P	R	No. of translated words	Not already in dict
1	0.718	0.198	1246	244
6	0.668	0.565	4535	1546
18	0.691	0.716	9122	4184

Table 2: Precision, recall, number of translated words and number of new words (not found in the dictionary) obtained with different corpus sizes

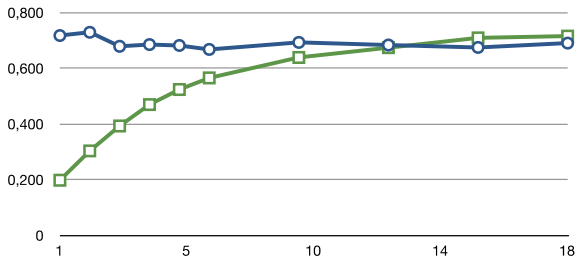

Figure 1: Precision and recall as a function of corpus size

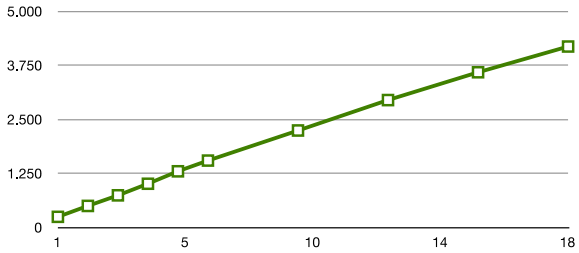

Figure 2: The number of new words (not found in the seed dictionary) as a function of corpus size

Figure 1 shows that precision with regard to the gold standard is more or less constant with an average of 0.68 if we disregard the first two measurements that are probably bad estimates since the intersection with the gold standard is small (as shown in Table 1) and evens out as the size of the corpus increases.

When analyzing recall against the gold standard we see the typical logarithmic recall behavior when depicted as a function of corpus size. On the other hand, when we consider the number of new translation equivalents (i.e. the number of source words that do not appear in the seed dictionary), the function behaves almost linearly (see Figure 2). This can be explained with the fact that in the dictionary the most frequent words are best represented. Because of that we can observe a steady increase in the number of words not present in the seed lexicon that pass the frequency threshold with the increasing corpus size.

Finally, we study the impact of the word frequency threshold for context vector generation on the quality and amount of the extracted translation equivalents on the six million corpora in both languages.

Frequency	P	No. of translated words	F1
25	0.561	7203	0.719
50	0.668	4535	0.648
75	0.711	3435	0.571
100	0.752	2803	0.513
125	0.785	2374	0.464
150	0.815	2062	0.424

Table 3: Precision, number of new words and F1 obtained with different frequency thresholds

As can be seen in Table 3, by lowering the frequency criterion, the F1 measure increases showing greater gain in recall than loss in precision. For calculating recall, the number of new words passing the frequency criterion is normalized with the assumed number of obtainable lexicon entries set to 7.203 (the number of new words obtained with the lowest frequency criterion).

This is a valuable insight since the threshold can be set according to different project scenarios. If, for example, lexicographers can be used in order to check the translation candidates and choose the best ones among them, the threshold may well be left low and they will still be able to identify the correct translation very quickly. If, on the other hand, the results will be used directly by another application, the threshold will be raised in order to reduce the amount of noise introduced by the lexicon for the following processing stages.

5.2 Manual evaluation

For a more qualitative inspection of the results we performed manual evaluation on a random sample of 100 translation equivalents that are not in the general seed dictionary or present in our gold standard. We were interested in finding out to what extent these translation equivalents belong to the health domain and if their quality is comparable to the results of the automatic evaluation.

Manual evaluation was performed on translation equivalents extracted from the comparable corpus containing 18 million English words and 6 million Slovene words, where the frequency threshold was set to 50. 51% of the manually evaluated words belonged to the health domain, 23% were part of general vocabulary, 10% were proper names and the rest were acronyms and errors arising from PoS-tagging and lemmatization in the ukWaC corpus. Overall, in 45% the first translation equivalent was correct and additional 11% contained the correct translation among the ten best-ranked candidates.

For 44 % of the extracted translation equivalents no appropriate translation was suggested. Among the evaluated health-domain terms, 61% were translated correctly with the first candidate and for the additional 20% the correct translation appeared among the first 10 candidates.

Of the 19% health-domain terms with no appropriate translation suggestion, 4 terms, that is 21% of the wrongly translated terms, were translated as direct hypernyms and could loosely be considered as correct (e.g. the English term *bacillus* was translated as *mikroorganizem* into Slovene, which means microorganism). Even most other translation candidates were semantically closely related, in fact, there was only one case in the manually inspected sample that provided completely wrong translations.

Manual evaluation shows that the quality of translations for out-of-goldstandard terms is consistent with the results of automatic evaluation. A closer look revealed that we were able to obtain translation equivalents not only for the general vocabulary but especially terms relevant for the health domain, and furthermore, that their quality is also considerably higher than for the general vocabulary which is not of our primary interest in this research.

The results could be further improved by filtering out the noise obtained from errors in PoS-tagging and lemmatization and, more importantly, by identifying proper names. Multi-word expressions should also be tackled as they present problems, especially in cases of 1:many mappings, such as the English single-word term *immunodeficiency* that is translated with a multi-word expression in Slovene (*imunska pomanjkljivost*).

6 Conclusions

In this paper we described the compilation process of a domain-specific comparable corpus from already existing general resources. The corpus compiled from general web corpora was used in a set of experiments to extract translation equivalents

for the domain vocabulary by comparing contexts in which terms appear in the two languages.

The results show that a 2-way translation of context vectors consistently improves the quality of the extracted translation equivalents by using additional information given from the reverse dictionary. Next, increasing the size of only one part of the comparable corpus brings a slight increase in precision but a very substantial increase in recall.

If we are able to translate less than 20% of the gold standard with a 1 million word corpus, the recall is exceeds 70% when we extend the English part of the corpus to 15 million words. Moreover, the increase of the number of new words we obtain in this way keeps being linear for even large corpus sizes. We can also expect the amount of available text to keep rising in the future.

This is a valuable finding because a scenario in which much more data is available for one of the two languages in question is a very common one.

Finally, we have established that the word frequency threshold for building context vectors can be lowered in order to obtain more translation equivalents without a big sacrifice in their quality. For example, a 10% drop in precision yields almost twice as many translation equivalents.

Manual evaluation has shown that the quality of health-related terms that were at the center of our research is considerably higher than the rest of the vocabulary but has also revealed some noise in POS-tagging and lemmatization of the ukWaC corpus that consequently lowers the results of our method and should be dealt with in the future.

A straightforward extension of this research is to tackle other parts of speech in addition to nouns. Other shortcomings of our method that will have to be addressed in our future work are multi-word expressions and multiple senses of polysemous words and their translations. We also see potential in using cognates for re-ranking translation candidates as they are very common in the health domain.

Acknowledgments

Research reported in this paper has been supported by the ACCURAT project within the EU 7th Framework Programme (FP7/2007-2013), grant agreement no. 248347, and by the Slovenian Research Agency, grant no. Z6-3668.

References

Arhar, Š., Gorjanc, V., and Krek, S. (2007). FidaPLUS corpus of Slovenian - The New Generation of the Slovenian Reference Corpus: Its Design and Tools. In *Proceedings of the Corpus Linguistics Conference* (CL2007), Birmingham, pp. 95-110.

Déjean, H., Gaussier, E., Renders, J.-M. and Sadat, F. (2005). Automatic processing of multilingual medical terminology: Applications to thesaurus enrichment and cross-language information retrieval. *Artificial Intelligence in Medicine*, 33(2): 111–124.

Doe, J. (2011): *Bilingual lexicon extraction from comparable corpora: A comparative study.*

Doe, J. (2011): *Compiling web corpora for Croatian and Slovene*.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics - Special issue on using large corpora,* 19(1).

Fung, P. (1998). A statistical view on bilingual lexicon extraction: From parallel corpora to nonparallel corpora. In *Proc. of the 3rd Conference of the Association for Machine Translation in the Americas*, pp. 1–17.

Fung, P., Prochasson, E. and Shi, S. (2010). Trillions of Comparable Documents. In *Proc. of the 3rd workshop on Building and Using Comparable Corpora* (BUCC'10), Language Resource and Evaluation Conference (LREC2010), Malta, May 2010, pp. 26–34.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Norwell, MA.

Koehn, P. and Knight, K. (2002). Learning a translation lexicon from monolingual corpora. In *Proc. of the workshop on Unsupervised lexical acquisition* (ULA '02) at ACL 2002, Philadelphia, USA, pp. 9–16.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1): 145-151.

Ljubešić, N. and Erjavec, T. (2011). hrWaC and slWaC: Compiling Web Corpora for Croatian and Slovene. (submitted to International Workshop on Balto-Slavonic Natural Language Processing).

Ljubešić, N., Fišer D., Vintar Š. and Pollak S. Bilingual Lexicon Extraction from Comparable Corpora: A Comparative Study. (accepted for WoLeR 2011 at ESSLLI International Workshop on Lexical Resources).

Marsi, E. and Krahmer, E. (2010). Automatic analysis of semantic similarity in comparable text through syntactic tree matching. In *Proc. of the 23rd International Conference on Computational Linguistics* (Coling 2010), pages 752–760.

Och, F. J. and Ney, H. (2000). Improved Statistical Alignment Models. In *Proc. of the 38th Annual Meeting of the Association for Computational Linguistics*, Hongkong, China, pp. 440–447.

Otero, P. G. (2007). Learning Bilingual Lexicons from Comparable English and Spanish Corpora. In *Proc. of the Machine Translation Summit* (MTS 2007), pp. 191–198.

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proc. of the 37th annual meeting of the Association for Computational Linguistics* (ACL '99), pp. 519–526.

Schmid, H. (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proc. of International Conference on New Methods in Language Processing*.

Saralegi, X., San Vicente, I. and Gurrutxaga, A. (2008). Automatic Extraction of Bilingual Terms from Comparable Corpora in a Popular Science Domain. In *Proc. of the 1st Workshop on Building and Using Comparable Corpora* (BUCC) at LREC 2008.

Shao, L. and Ng, H. T. (2004). Mining New Word Translations from Comparable Corpora. *In Proc. of the 20th International Conference on Computational Linguistics* (COLING '04), Geneva, Switzerland.

Shezaf, D. and Rappoport, A. (2010). Bilingual Lexicon Generation Using Non-Aligned Signatures. In *Proc.of the 48th Annual Meeting of the Association for Computational Linguistics* (ACL 2010), Uppsala, Sweden, pp. 98–107.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D. and Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proc. of the 5th International Conference on Language Resources and Evaluation* (LREC 2006), pp. 2142–2147.

Tiedemann, J. (2005). Optimisation of Word Alignment Clues. *Natural Language Engineering*, 11(03): 279–293.

Vorhees, E. M. (2001). *Overview of the TREC-9 Question Answering Track*. In Proceedings of the Ninth Text REtrieval Conference (TREC-9), 2001.

Xiao, Z., McEnery, A. (2006). Collocation, semantic prosody and near synonymy: a cross-linguistic perspective. *Applied Linguistics* 27(1): 103–129.

Yu, K. and Tsujii, J. (2009). *Bilingual dictionary extraction from Wikipedia.* In *Proc. of the 12th Machine Translation Summit* (MTS 2009), Ottawa, Ontario, Canada.