Learning Spanish-Galician Translation Equivalents using a Comparable Corpus and a Bilingual Dictionary

Pablo Gamallo Otero[1] and José Ramom Pichel Campos[2]

[1] Departamento de Língua Espanhola, Faculdade de Filologia
Universidade de Santiago de Compostela, Galiza, Spain
[2] Departamento de Tecnologia Linguística da Imaxin|Software
Santiago de Compostela, Galiza

Abstract. So far, research on extraction of translation equivalents from comparable, non-parallel corpora has not been very popular. The main reason was the poor results when compared to those obtained from aligned parallel corpora. The method proposed in this paper, relying on *seed patterns* generated from external bilingual dictionaries, allows us to achieve similar results to those from parallel corpus. In this way, the huge amount of comparable corpora available via Web can be viewed as a never-ending source of lexicographic information. In this paper, we describe the experiments performed on a comparable, Spanish-Galician corpus.

1 Introduction

There exist many approaches to extract bilingual lexicons from parallel corpora [8, 16, 1, 22, 14]. These approaches share the same basic strategy: first, bitexts are aligned in pairs of segments and, second, word co-ocurrences are computed on the basis of that alignment. They usually reach high score values, namely about 90% precision with 90% recall. Unfortunately, parallel texts are not easily available, in particular for minority languages. To overcome this drawback, different methods to extract bilingual lexicons have been implemented lately using non-parallel, comparable corpora. These methods take up with the idea of using the Web as a huge resource of multilingual texts which can be easily organized as a collection of non-parallel, comparable corpora. A non-parallel and comparable corpus (hereafter "comparable corpus") consists of documents in two or more languages which are not translation of each other and deal with similar topics. However, the accuracy scores of such methods are not as good as those reached by the strategies based on aligned parallel corpora. So far, the highest values have not improved an 72% accuracy [18], and that's without taking into consideration the coverage of the extracted lexicon over the corpus.

This paper proposes a new method to extract bilingual lexicons from a POS tagged comparable corpus. Our method relies on the use of a bilingual dictionary

to identify bilingual correlations between pairs of lexico-syntactic patterns. Such patterns will be used as "seed expressions" as follows: a lemma of the target language will be taken as a possible translation of a lemma in the source language if both lemmas co-occur with a great number of seed patterns. Beside the external dictionary, we also identify seed patterns with cognates previously selected from the comparable corpus. We will work not only on monoword lemmas but also on multiwords. Our results improve the accuracy reached by Rapp (i.e. 72%), for a coverage of more than 80%. These encouraging results show that the huge amount of comparable corpora via Web can be seen as an endless resource of lexicographic knowledge.

The article is organized as follows. In Section 2, we will situate our approach with regard to the state of art in comparable corpora extraction. Section 3 will be focused on defining the different steps of our approach. Then, in 4, we will describe the experiments performed on a Spanish-Galician corpus as well as an evaluation protocol. Finally, we will enumerate some conclusions and discuss future work.

2 Related Work

There are not many approaches to extract bilingual lexicons from non-parallel corpora in comparison to those using a strategy based on aligned, parallel texts. The most popular method to extract word translations from non-parallel, comparable corpora is described and used in [6, 7, 18, 4, 19]. The starting point of this strategy is as follows: word w_1 is a candidate translation of w_2 if the words with which w_1 co-occurs within a particular window are translations of the words with which w_2 co-occurs within the same window. This strategy relies on a list of bilingual word pairs (called *seed words*) provided by an external bilingual dictionary. So, w_1 is a candidate translation of w_2 if they tend to co-occur with the same seed words. The main drawback of this method is the use of word windows to define coarse-grained contexts. According to the Harris's hypothesis [13], counting co-occurrences within a window of size N is less precise than counting co-occurrences within local syntactic contexts. In the most efficient approaches to thesaurus generation [12, 15], word similarity is computed using co-occurrences between words and specific syntactic contexts. Syntactic contexts are considered to be less ambiguous and more sense-sensitive than contexts defined as windows of size N. In order to define contexts with more fine-grained information, we build a list of bilingual lexico-syntactic templates. In [9], these templates were previously extracted from small samples of parallel corpus. In this paper, however, they are extracted directly from an external bilingual dictionary. As such templates represent unambiguous local contexts of words, they are discriminative and confident seed expressions to extract word translations from comparable texts. In [21], syntactic templates are also used for extraction of translations, but they were specified with semantic attributes introduced by hand. In [5], it is described a particular strategy based on a multilingual thesaurus instead of an external bilingual dictionary. Finally, some researchers have

focused on a different issue: disambiguation of candidate translations. According to [17], the process of building bilingual lexicons from non-parallel corpora is a too difficult and ambitious objective. He preferred to work on a less ambitious task: to choose between several translation alternatives previously selected from a bilingual dictionary.

3 The Approach

Our approach consists of three steps: (1) text processing, (2) building a list of seed bilingual patterns by using a bilingual dictionary and a set of cognates previously selected from the corpus, and (3), translation equivalents extraction from a comparable corpus making use of the list of seed patterns.

3.1 Text Processing

POS Tagging and Multiword Extraction First, the texts of both languages are lemmatized and POS tagged. Lemmatization also involves name entity recognition (i.e., identification of proper nouns). Proper nouns can be either mono or multiword units. Besides monowords lemmas and proper nouns, we also extract multiwords, that is, lemmas consisting of several lexical units with some degree of internal cohesion: e.g., "traffic jam", "tv channel", "take into account", etc. This type of expressions are extracted using basic patterns of POS tags such as N-PRP, N-A, V-N, etc. This task is performed on the comparable corpus, so we extract multiword candidates in both languages. Then, the list of multiword candidates is reduced with a basic statistical filter, which only selects those multiwords with a *SCP* coefficient higher than a empirically set threshold. Here, we follows the strategy described in [20].

Dependency Triplets and Lexico-Syntactic Patterns Once the corpus has been POS tagged and the multiwords have been extracted, we build a collocation database where each entry consists of a lemma (either monoword unit or multiword) and the lexico-syntactic patterns with which it co-occurs in the corpus. The database is built in two steps. First, we make use of regular expressions to identify binary dependencies. Regular expressions represent basic patterns of POS tags which are supposed to stand for syntactic dependencies between two lemmas. In our approach, we work with dependencies between verbs, nouns, and adjectives. Second, we extract lexico-syntactic patterns from the dependencies and count the co-occurrences of lemmas with those lexico-syntactic patterns. Let's take an example. Suppose our corpus contains the following tagged sentence:

> a_D man_N see_V yesterday_R a_D very_R big_A dog_N with_PRP a_D broken_A leg_N

Table 1. Dependency triplets and patterns of POS tags

Dependencies	Patterns of POS tags
$(see, subj, man)$	$(\mathbf{N})(?:A\|R)*(\mathbf{V})$
(see, obj, dog)	$(\mathbf{V})(?:R\|D\|R\|A\|N)*(\mathbf{N})$
$(dog, with, leg)$	$(\mathbf{N})(?:R\|A)*(\mathbf{PRP})(?:D\|R\|A\|N)*(\mathbf{N})$
(dog, mod, big) $(leg, mod, broken)$	$(\mathbf{A})(?:N)*(\mathbf{N})$
()	$(\mathbf{N})(?:N)*(\mathbf{N})$
()	$(\mathbf{V})(?:R)*(\mathbf{PRP})(?:D\|R\|A\|N)*(\mathbf{N})$

Table 2. Collocation database of lemmas and lexico-syntactic patterns

Lemmas	Lexico-Syntactic Patterns and freqs.
man	$< (see, subj, N), 1 >$
see	$< (V, subj, man), 1 >$, $< (V, obj, dog), 1 >$
big	$< (dog, mod, A), 1 >$
dog	$< (N, mod, big), 1 >$, $< (N, with, leg), 1 >$
broken	$< (leg, mod, A), 1 >$
leg	$< (N, mod, broken), 1 >$, $< (dog, with, N), 1 >$

The first step consists in identifying dependencies between lemmas using basic patterns of POS tags. Dependencies are noted as triplets: $(head, rel, dependent)$. Table 1 shows the 5 triplets extracted from the sentence above using different patterns of POS tags. The 5 extracted triplets instantiate 4 schemes of dependencies: adjective-noun, noun-verb, verb-noun, and noun-prep-noun. The sentence does not contain triplets instantiating the noun-noun and verb-prep-noun dependencies. Wildcards $(?:D|R|A|N)*$ stand for optional modifiers, that is, they represent sequences of determiners, adverbs, adjectives, or nouns that are not considered for triplets.

In the second step, the extracted triplets allow us to easily build the collocation database depicted in Table 2. The first line of the table describes the entry *man*. This noun co-occurs once with one lexico-syntactic pattern, which represents the subject position of the verb *see*. The second line describes the entry *see*, which co-occurs once with two lexico-syntactic patterns: a verb co-occurring with *man* in the subject position and a verb co-occurring with *dog* in the object position. The remaining lines describe the collocation information of the other nouns and adjectives appearing in the sentence above.

Notice we always extract 2 complementary lexico-syntactic patterns from a triplet. For instance, from $(dog, with, leg)$, we extract:

- $(N, with, leg)$
- $(dog, with, N)$

This is in accordance with the notion of co-requirement defined in [10]. In this work, two syntactically dependent words are no longer interpreted as a standard

"predicate-argument" structure, where the predicate is the active function imposing syntactic and semantic conditions on a passive argument, which matches such conditions. On the contrary, each word of a binary dependency is perceived simultaneously as a predicate and an argument. In the example above, $(dog, with, N)$ is seen as a unary predicate that requires nouns denoting parts of dogs (e.g. legs), and simultaneously, $(N, with, leg)$ is another unary predicate requiring as argument entities having legs (e.g. dogs).

To simplify the process of extracting binary relations, long-distance dependencies are not taken into account. So, we do not propose the attachment between the verb "see" and the prepositional phrase "with a broken leg". In fact, our use of regular expressions over POS-tags emulates a parsing strategy based on the Right-Association heuristic. It is a robust analysis, and about 75% of the triplets are correctly extracted. Note that the patterns of tags in Table 1 work well for English texts. To extract triplets from texts in Romance languages such as Spanish, French, Portuguese, or Galician, we need to do, at least, 3 tiny changes: nouns as optional modifiers are not taken into account; a new pattern with dependent adjectives at the right position of nouns is required; the noun in the left position of a noun-noun dependency must be considered the head of the triplet. The experiments that will be described later were performed over Spanish and Galician corpora.

3.2 Generating Seed Lexico-Syntactic Patterns

To extract translation equivalents from a comparable corpus, a list of "seed" expressions is required. In our approach, the seed expressions used as cross-language pivot contexts are not bilingual pairs of words as in related work, but bilingual pairs of lexico-syntactic patterns (or "seed patterns"). The process of building a list of seed patterns consists of two steps: first, we generate a large list from an external bilingual dictionary (and from a set of cognates). Second, this list is reduced by filtering out those pairs of patterns that do not occur in the comparable corpus. We also remove those that are sparse or unbalanced in the corpus.

Patterns from Bilingual Dictionaries In order to generate bilingual correlations between lexico-syntactic patterns, we make use of bilingual dictionaries. Let's suppose that an English-Spanish dictionary translates the noun *import* into the Spanish counterpart *importación*. To generate bilingual pairs of lexico-syntactic patterns from these two nouns, we follow basic rules such as: (1) if *import* is the subject of a verb, then its Spanish equivalent, *importación*, is also the subject; (2) if *import* is modified by an adjective at the left position, then its Spanish equivalent is modified by an adjective at the right position; (3) if *import* is restricted by a prepositional complement headed by the preposition *in*, then its Spanish counterpart is restricted by a prepositional complement headed by the preposition *en*. The third rule needs a closed list of English prepositions and their more usual Spanish translations. For each entry (noun, verb, or adjective), we only generated a subset of all possible patterns. Table 3 depicts the

patterns generated from the bilingual pair *import-importación* and a restricted set of rules.

Table 3. Bilingual correlations between patterns generated from the translation pair: import-importación.

English	**Spanish**
$(import, of\|to\|in\|for\|by\|with, N)$	$(importación, de\|a\|en\|para\|por\|con, N)$
$(N, of\|to\|in\|for\|by\|with, import)$	$(N, de\|a\|en\|para\|por\|con, importación)$
$(V, obj, import)$	$(V, obj, importación)$
$(V, subj, import)$	$(V, subj, importación)$
$(V, of\|to\|in\|for\|by\|with, import)$	$(V, de\|a\|en\|para\|por\|con, importación)$
$(import, mod, A)$	$(importación, mod, A)$

In order to have a larger list of bilingual patterns, we also use a complementary strategy based on the identification of cognates from the comparable texts. We call "cognates" two lemmas written in the same way. We select those cognates appearing in the texts that are not in the bilingual dictionary. Most of them are proper names and dates. As they can be treated as entries of a bilingual dictionary, we are able to generate more bilingual lexico-syntactic patterns using the same basic rules described above.

Filtering The list generated in the previous process may contain lexico-syntactic patterns that do not occur in the comparable corpus, i.e., in the collocation database created in the first step of the approach (Subsection 3.1). Such patterns are removed. In addition, we also filter out those bilingual pairs that have one of these two properties: being sparse or being unbalanced in the comparable corpus. A bilingual pair of patterns is sparse if it has high dispersion. Dispersion is defined as the number of different lemmas occurring with a bilingual pair divided by the total number of lemmas in the comparable corpus. A bilingual pair is unbalanced when one of the patterns is very frequent while the other one is very rare. We use empirically set thresholds to separate sparse and unbalanced bilingual patterns from the rest. The final list of selected patterns will be used as seed expressions in the following step.

3.3 Identifying Translation Equivalents

The final step consists in extracting translation equivalents of lemmas with the help of the list of seed patterns. To compute the similarity between a lemma in the source language and a lemma in the target language, we conceive lemmas as vectors whose dimensions are the seed patterns. The value for each dimension is selected from the co-occurrence information stocked in the *collocation database* (see above Subsection 3.1). For instance, let's suppose that the collocation database contains the English lemma *uranium* co-occurring 14 times with

the lexico-syntactic pattern $(import, of, N)$. As this English pattern was associated to the Spanish pattern $(importación, of, N)$ in the list of seed patterns, then, we have to build a vector for *uranium* whose value is 14 in the dimension defined by this pair of patterns. Note that all Spanish lemmas co-occurring 14 times with $(importación, of, N)$ require vectors with the same value in the same dimension.

Similarity between lemmas l_1 and l_2 is computed using the following version of the *Dice* coefficient:

$$Dice(l_1, l_2) = \frac{2 * \sum_i min(f(l_1, p_i), f(l_2, p_i))}{f(l_1) + f(l_2)} \quad (1)$$

where $f(l_1, p_i)$ represents the number of times the lemma l_1 co-occurs with a seed pattern p_i. In some experiments, instead of co-occurrences we used log-likelihood as association value between lemmas and patterns. This weighted version of the measure did not improve the results in a significant way, since unbalanced and sparse patterns were filtered out before computing similarity. So, all the experiments described and evaluated in the next section were performed using only co-occurrences as association value.

As a result, each lemma of the source language is associated a list of candidate translations. This list is ranked by degree of similarity.

4 Experiments and Evaluation

4.1 The Comparable Corpus

The experiments were performed on a Spanish and Galician comparable corpus, which is constituted by news from on-line journals published between 2005 and 2006. As the Spanish corpus, we used 13 million words of two newspapers: *La Voz de Galicia* and *El Correo Gallego*, and as Galician corpus 10 million words of *Galicia-Hoxe*, *Vieiros* and *A Nosa Terra*. the Spanish and Galician texts were lemmatized and POS tagged using a multilingual free software: Freeling [3]. Since the orientation of the newspaper is quite similar the two corpora can be considered as more or less comparable.

4.2 The Bilingual Dictionary

The bilingual dictionary used to generate part of the seed patterns is the lexical resource integrated in an open source machine translation system, OpenTrad, for Spanish-Galician [2]. The final objective of our experiments is to update that resource in order to improve the results of the machine translation system, which is used by *La Voz de Galicia*, the sixth newspaper in Spain concerning the number of readers. The dictionary contains about $25,000$ Spanish-Galician entries.

The amount of bilingual patterns generated from the entries of the dictionary is 539,561. In addition, we generated 754,469 further patterns from bilingual cognates. In sum, we got 1,294,030. However, after the filtering process, the list of seed patterns is reduced to only 127,604. This is the number of dimensions of lemma vectors.

4.3 The Evaluation Protocol

To evaluate the efficiency of our method in the process of extracting translation equivalents, we elaborate an evaluation protocol with the following characteristics. Accuracy is computed taking into account coverage at tree levels: 90%, 80%, and 50%. In our work, to choose a level of coverage, we need to rank lemmas of the source language by frequency and select those whose frequency covers a specific percentage of the total frequency in the corpus. More precisely, given a ranked list of all lemmas found in the corpus, a level of coverage is the frequency in the corpus of an ordered set of lemmas in the list divided by the frequency of all lemmas. This ratio was computed separately for three different POS categories: nouns, adjectives, and verbs. This way, a 90% of coverage for nouns means that the frequency of the nouns considered for evaluation is 90% with regard to the total frequency of all nouns in the corpus.

To compute accuracy, we need first to choose a specific POS category and a particular level of coverage. Then, we randomly extract 150 test lemmas of this category from the list of lemmas whose occurrences in the corpus achieve the level of coverage at stake. We compute two types of accuracy: *accuracy-1* is defined as the number of times a correct translation candidate of the test lemma is ranked first, divided by the number of test lemmas. *Accuracy-10* is the number of correct candidates appearing in the top 10, divided by the number of test lemmas. Indirect associations are judged to be incorrect.

As far as we know, no definition of coverage (nor recall) has ever been proposed in related work. In most evaluation protocols of previous work, authors only give information on the number of occurrences of the test words in the corpus. In some work, test words are the N most frequent expressions in the training corpus [7], while in other experiments, they are word types or lemmas with a frequency higher than N (where N is often $>= 100$) [11, 4]. In fact, as absolute frequencies are dependent on the corpus size, they are not very useful if we try to compare the precision or accuracy among different approaches. By considering levels of coverage, which are independent of the corpus size, we try to overcome such a limitation.

4.4 Results

Table 4 shows the evaluation of our approach. For each POS category (including multiword nouns), and for each level of coverage (90%, 80%, and 50%), we compute *accuracy-1* and *accuracy-10*.

As far as nouns are concerned, the three levels of coverage (90%, 80%, and 50%) correspond to three lists of lemmas containing 9,798, 3,534, and 597 nouns,

Table 4. Evaluation of our approach

Category	Cov(%)	Acc-1	Acc-10	lemmas
Noun	90%	.55	.60	9,798
Noun	80%	.81	.90	3,534
Noun	50%	.95	.99	597
Adj	90%	.61	.70	1,468
Adj	80%	.81	.87	639
Adj	50%	.94	.98	124
Verb	90%	.92	.99	745
Verb	80%	.97	.100	401
Verb	50%	.100	.100	86
multi-lex	50%	.59	.62	2,013

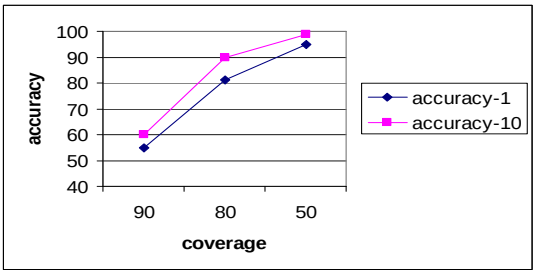
Table 5. Evaluation of the baseline method

Category	Cov(%)	Acc-1	Acc-10	lemmas
Noun	80%	.26	.54	3,534
Adj	80%	.43	.70	639

respectively. As nouns, we include all sort of proper names. Figure 1 depicts the progression of the two accuracies (1 and 10) at the three levels of coverage. With a coverage of 80%, accuracy is quite acceptable: between .80 and .90. At this level of coverage, the frequency of the test lemmas is ≥ 129. In fact, such a minimum frequency is not far from the thresholds proposed by related works, where the smallest frequency of test words was, in most cases, 100. However, in those works the best accuracy merely achieves 72% [18].

Regarding accuracy of adjectives and verbs, there is a significant difference in their results. Whereas the accuracy of verbs is close to .100 at the coverage of 80%, adjectives only reach about .80 of accuracy with the same coverage. The

Fig. 1. Accuracy of nouns at 3 levels of coverage



main drawback with adjectives comes from the difficulties of the POS tagger to correctly disambiguate between adjectives and past participles.

As far as multiword nouns are concerned, accuracy is about .60 at the coverage of 50%. The main drawback regarding multiwords is their low frequency in the corpus. The minimum frequency of the 2,013 lemmas evaluated at this level is very low, 40, which prevents us from getting acceptable results. However, our results are better than those obtained by similar approaches using multiword terms, with .52 accuracy in the best case [6][3].

Finally, Table 5 depicts the results of a baseline method. The baseline strategy relies on seed words and windows of size 2 (i.e., 4 context positions) instead of on lexico-syntactic patterns. In fact, with this baseline, we tried to simulate some aspects of the approach described by [18]. To permit comparing this approach to ours, we used as similarity coefficient the dice measure defined above. As in [18], our baseline method only search for translation equivalents of nouns and adjectives. In table 5, we can observe the accuracy obtained using the baseline method when the coverage is situated at 80%. This accuracy is significatively lower that the scores reached by our approach. So, lexico-syntactic patterns seem to be more precise than contexts based on windows of size N. Notice that the accuracy in our simulation is lower than that obtained by Rapp (about 72%). Such a difference can be explained by the size of our training corpus, 10 times smaller than the corpus used by Rapp.

5 Conclusions and Future Work

Few approaches to extract word translations from comparable, non-parallel texts have been proposed so far. The main reason is that results are not yet very encouraging. Whereas for parallel texts, most work on word translation extraction reaches more than 90%, the accuracy for non-parallel texts has been around 72% up to now. The main contribution of the approach proposed in this paper is to use bilingual pairs of lexico-syntactic patterns as seed expressions. This makes a significant improvement to about 80/90% of word translations identified correctly if only the best candidate is considered, and about 90/95% if we consider the top 10. These results are not very far from those obtained by approaches based on parallel texts. Such results show that non-parallel, comparable corpora can be considered as an interesting source of lexicographic knowledge. Moreover, there is still a good margin to improve results. Given that comparable corpora are growing daily as the web is getting larger, it could be easy to update and enrich bilingual lexicons and translation memories in an incremental way. Our current work is precisely to retrieve monthly further documents from the web in order to make the training corpus larger and update our bilingual lexicon. This way, we aim at improving the specific bilingual resource used by OpenTrad, a Spanish-Galician machine translation system.

[3] The merit of this work is to extract translation equivalents from two very different languages: English and Japanese.

Acknowledgments

This work has been supported by the Galician Government, within the project ExtraLex, ref: PGIDIT07PXIB204015PR.

References

1. Lars Ahrenberg, Mikael Andersson, and Magnus Merkel. A simple hybrid aligner for generating lexical correspondences in parallel texts. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 29–35, Montreal, 1998.
2. Carme Armentano-Oller, Rafael C. Carrasco, Antonio M. Corb-Bellot, Mikel L. Forcada, Mireia Ginest-Rosell, Sergio Ortiz-Rojas, Juan Antonio Prez-Ortiz, Gema Ramrez-Snchez, Felipe Snchez-Martnez, and Miriam A. Scalco. Open-source portuguese-spanish machine translation. In *Lecture Notes in Computer Science, 3960*, pages 50–59, 2006.
3. X. Carreras, I. Chao, L. Padró, and M. Padró. An open-source suite of language analyzers. In *4th International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, 2004.
4. Y-C. Chiao and P. Zweigenbaum. Looking for candidate translational equivalents in specialized, comparable corpora. In *19th COLING'02*, 2002.
5. H. Dejean, E. Gaussier, and F. Sadat. Bilingual terminology extraction: an approach based on a multilingual thesaurus applicable to comparable corpora. In *COLING 2002*, Tapei, Taiwan, 2002.
6. Pascale Fung and Kathleen McKeown. Finding terminology translation from non-parallel corpora. In *5th Annual Workshop on Very Large Corpora*, pages 192–202, Hong Kong, 1997.
7. Pascale Fung and Lo Yuen Yee. An ir approach for translating new words from nonparallel, comparable texts. In *Coling'98*, pages 414–420, Montreal, Canada, 1998.
8. Willian Gale and Kenneth Church. Identifying word correspondences in parallel texts. In *Workshop DARPA SNL*, 1991.
9. Pablo Gamallo. Learning bilingual lexicons from comparable english and spanish corpora. In *Machine Translation SUMMIT XI*, Copenhagen, Denmark, 2007.
10. Pablo Gamallo, Alexandre Agustini, and Gabriel Lopes. Clustering syntactic positions with similar semantic requirements. *Computational Linguistics*, 31(1):107–146, 2005.
11. Pablo Gamallo and José Ramom Pichel. An approach to acquire word translations from non-parallel corpora. In *12th Portuguese Conference on Artificial Intelligence (EPIA'05)*, Evora, Portugal, 2005.
12. Gregory Grefenstette. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, USA, 1994.
13. Z. Harris. Distributional structure. In J.J. Katz, editor, *The Philosophy of Linguistics*, pages 26–47. New York: Oxford University Press, 1985.
14. Oi Yee Kwong, Benjamin K. Tsou, and Tom B. Lai. Alignment and extraction of bilingual legal terminology from context profiles. *Terminology*, 10(1):81–99, 2004.
15. Dekang Lin. Automatic retrieval and clustering of similar words. In *COLING-ACL'98*, Montreal, 1998.

16. Dan Melamed. A portable algorithm for mapping bitext correspondences. In *35th Conference of the Association of Computational Linguistics (ACL'97)*, pages 305–312, Madrid, Spain, 1997.
17. Hiroshi Nakagawa. Disambiguation of single noun translations extracted from bilingual comparable corpora. *Terminology*, 7(1):63–83, 2001.
18. Reinhard Rapp. Automatic identification of word translations from unrelated english and german corpora. In *ACL'99*, pages 519–526, 1999.
19. Li Shao and Hwee Tou Ng. Mining new word translations from comparable corpora. In *20th International Conference on Computational Linguistics (COLING 2004)*, pages 618–624, Geneva, Switzerland, 2004.
20. J. F. Silva, G. Dias, S. Guilloré, and G. P. Lopes. Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units. In *Progress in Artificial Intelligence*, pages 113–132. LNAI, Springer-Verlag, 1999.
21. T. Tanala. Measuring the similarity between compound nouns in different languages using non-parallel corpora. In *19th COLING'02*, pages 981–987, 2002.
22. Jorg Tiedemann. Extraction of translation equivalents from parallel corpora. In *11th Nordic Conference of Computational Linguistics*, Copenhagen, Denmark, 1998.