Strategies for Building High Quality Bilingual Lexicons from Comparable Corpora

Pablo Gamallo

1 Introduction

A comparable corpus consists of documents in two or more languages or varieties which are not a translation of each other and deal with similar topics. Comparable corpora are by definition multilingual and cross-lingual text collections. The use of comparable corpora to automatically extract bilingual lexicons has been growing in recent years (Tamura et al. 2012, Aker et al. 2013, Ansari et al 2014, Hazem / Morin 2014). The main advantage of using comparable corpora to perform this extraction task is that they are easily available and make use of the internet as a huge resource of multilingual texts. Comparable corpora are more easily available than parallel texts, especially for minority languages. However, their main drawback is the low performance of the extraction systems based on them. According to Nakagawa (2001), bilingual lexicon extraction from comparable corpora is an overly difficult and ambitious objective, and much more complex than extraction from parallel and aligned corpora.

It is possible, however, to use comparable corpora in a less ambitious way, not to build large and accurate bilingual lexicons from scratch, but just to filter out false bilingual pairs from those selected by other basic bilingual extraction methods. In this paper, we will focus on two basic methods for extracting bilingual lexicons: first, the construction of new bilingual dictionaries by transitivity using intermediary dictionaries and, second, the

selection of bilingual cognates by means of string similarity. These two strategies are aimed at building large bilingual lexicons and/or terminologies, even if their correctness and precision is low due to polysemy and false friend candidates. In order to discard false pairs and select only correct pairs of translation candidates, we take into account their context and distribution in comparable corpora. In this way, the bilingual pairs with a similar distribution in comparable corpora are considered to be correct pairs and subsequently are not removed from the lexicon. Let us see some examples to illustrate our methodology. Suppose we are deriving a new English-Galician dictionary by transitivity from two existing ones, English-Spanish and Spanish-Galician, with Spanish as the intermediary (or pivot) language between English and Galician. Let us take, for instance the English verb *subside*, which is translated by the polysemic Spanish word *bajar* in the English-Spanish dictionary. The Spanish polysemic word is, in turn, translated by the Spanish-Galician dictionary into two different verbs: *baixar* (*go down*) and *apear* (*take down*). Then, the derived English-Galician dictionary generates the bilingual pairs (*subside, baixar*) and (*subside, apear*). While the former translation is correct, the latter is clearly odd. The Galician verb *apear* does not mean *subside* in any context; it means *take down*, which is one of the senses of the Spanish word *bajar*. In order to filter out the false pair (*subside, apear*) while keeping (*subside, baixar*), we compute the distributional similarity of those pairs using comparable corpora. As *subside* appears in very different contexts than the Galician word *apear* (*take down*), this bilingual pair is removed from the derived dictionary. Let us now suppose that we are building an English-

Spanish list of bilingual terms by selecting those with high string similarity. This procedure is known as a bilingual cognates search. In the process of searching for bilingual cognates, the main problem that arises concerns false friends. For instance, the spelling of the English noun *code* is very close to that of the Spanish word *codo* (*elbow*). They are separated by an Edit Distance of only 1 (i.e. they differ by just one character). However, they cannot be considered to be translation candidates because their meanings are very different. To filter out fake bilingual pairs and select the correct ones, we again use their distribution similarity in comparable corpora. As both words do not appear in similar contexts, the pair is removed from the list of bilingual cognates.

In short, the specific objective of this paper is to describe two methods used to derive new bilingual lexicons using comparable corpora to select correct candidate pairs. The first method consists of using two existing bilingual dictionaries, (*A, B*) and (*B, C*), in order obtain a new pair (*A, C*) by simple transitivity and, then, in validating correctly generated bilingual correspondences by using dependency-based distributional similarity computed from comparable corpora. The second method consists of generating candidate cognates from comparable corpora and, then, in validating correct candidates by computing their dependency-based distributional similarity in those corpora. As the experiments conducted will show, the performance of these strategies in terms of precision is close to the precision achieved by extraction methods based on parallel corpora.

This paper organizes, integrates and expands the work presented in two previous articles: (Gamallo / Pichel 2010, Gamallo / Garcia 2012) with further experiments. The rest of the article is organized as follows: Section 2 describes the different steps underlying the method of building lexicons by transitivity with a pivot language, then,, Section 3 describes the cognate-based strategy. In Section 4, we describe some experiments aimed at generating new bilingual dictionaries and evaluating the performance of our different strategies. Finally, some conclusions are presented in Section 5.

2 Pruning lexicons built by transitiviy

Our strategy[1] consists of two main tasks: both the generation of candidate bilingual correspondences by transitivity and their validation by using translation equivalents extracted from comparable corpora. This strategy is especially well suited to creating new language resources for minority languages (e.g., Galician) from languages such as English or Spanish, which have many more resources. The method does not require the minority language to be provided with many and large linguistic resources: only some raw text is required. This is enough to automatically build a new non-noisy, bilingual lexicon.

2.1 *Basic assumptions*

The crucial aspect of the method is the process of validating bilingual correspondences derived by transitivity, by means of translation equivalents

[1] The strategy was implemented in a prototype available at: http://gramatica.usc.es/~gamallo/prototypes/BilingualExtraction.tar.gz

extracted from comparable corpora. We observed that if a bilingual pair derived by transitivity also appears in the list of pairs extracted by distributional similarity from comparable corpora, then the pair is correct. This observation is supported by the following linguistic conjectures:

- In handcrafted bilingual dictionaries, each bilingual correspondence consists of two terms that share two different aspects of their lexical meaning: both of them have similar *conceptual* and *distributional* properties. It follows that the two terms both refer to similar entities or concepts (conceptual properties) and combine with similar entities or concepts (distributional properties).
- In noisy bilingual dictionaries derived by transitivity, most generated correspondences consist of bilingual pairs that share similar *conceptual properties*, but do not always have the same *distributional properties*. This is because the different senses of a polysemous word are related by conceptual aspects but not in distributional terms. Only homonymous words in the pivot language give rise to completely wrong and unrelated bilingual correspondences generated by transitivity.
- In bilingual lexicons automatically extracted from comparable corpora, the extracted correspondences consist of bilingual pairs with the same *distributional properties*, but which do not always share similar *conceptual properties*.

It follows on that correct bilingual pairs are those that only share both conceptual and distributional properties. Then, the intersection of the dictionaries derived by transitivity (conceptual similarity) with those which

are extracted from comparable corpora (distributional similarity) give rise to correct bilingual pairs, i.e., to pairs that share both conceptual and distributional similarity. This intersection results in a bilingual, non-noisy lexicon.

The distributional hypothesis states that two words are semantically related if they share similar linguistic contexts. In a bilingual framework this hypothesis may allow identifying translation candidates. The procedure works as follows: a word *w2* in the target language is a candidate translation of *w1* in the source language if the context expressions with which *w2* co-occurs tend to be translations of the context expressions with which *w1* co-occurs. The basis of the method is to find the target words that have the most similar distributions with a given source word. The starting point of this strategy is a list of bilingual expressions that are used to build the context vectors defining all words in both languages. This list is usually provided by an external bilingual dictionary.

2.2 *The method*

Let us look at the toy example in Figure 1. The objective is to generate non-noisy English-Portuguese pairs using Spanish as a pivot language.

In the English-Spanish dictionary, the Spanish noun *titular* has two different bilingual correspondences: *(headline, titular)* and *(holder, titular)*. This noun is then a polysemous word which also appears with two translations in the Portuguese-Spanish dictionary: *(titular, manchete)* and *(titular, titular)*. The two senses of the Spanish polysemous word are conceptually related: the two refer to small text labels, the headline of an article or a

person's name, used to identify either the specific article of a journal or a specific card owner.

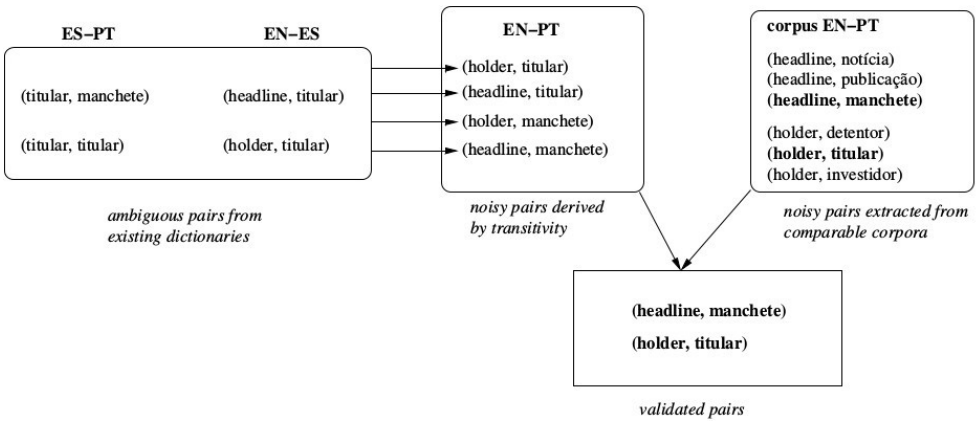

Figure 1: Example of the validation process

The English-Portuguese pairs derived by transitivity are: *(headline, manchete), *(holder, manchete)*, **(headline, titular)*, *(holder, titular)*, where "*" stands for incorrect pairs. So, derivation by transitivity overgenerates bilingual pairs when one of the source words has multiple senses. According to our conjectures, even if all these generated pairs are somehow conceptually related, only those that are also distributionally similar can be considered to be correct. To identify the correct pairs, we make use of the translation equivalents extracted from an English-Portuguese comparable corpus, which allow us to validate those distributionally related pairs. In our experiments (described later in Section 4), the Portuguese translation candidates of *headline* extracted by our system are the following nouns: *notícia* (*news*), *publicação* (*publication*),

manchete (*headline*), etc. These words are distributionally similar, but only the latter (in bold) describes the same concept as *headline*.

The noun *titular* (*holder*) was not extracted because its word distribution is very different to that of *headline*. On the other hand, the Portuguese translation candidates of *holder* extracted by our systems are nouns like *detentor* (*detainer*), *titular* (*holder*), *investidor* (*investor*), etc. All of them have a similar distribution (agents of verb actions), but only the second one in Figure 1 (in bold) refers to the same concept as *holder*. The term *manchete* (*headline*) was not extracted because its distribution is very different from that of *holder*. This intersection results in a bilingual, non-noisy lexicon. So, the final intersection between the noisy pairs generated by transitivity and those derived from comparable corpora yields correct bilingual pairs.

It is worth mentioning that computing distributional similarity from comparable corpora requires some external bilingual resources to generate seed contexts. These seed contexts are conceived as anchors to (pseudo)-align the bilingual text corpora.

3 Pruning bilingual cognates

3.1 *Basic Assumptions*

An efficient strategy used to build high quality bilingual lexicons between closely related languages is to search for bilingual cognates in highly comparable corpora. *Bilingual cognates* are considered here to be those words in two languages with similar spelling and similar meaning. There

are, at least, three different aspects involved in the correctness of these kinds of lexicons:

Corpus similarity: The more comparable the corpora are, the more efficient the extraction performed on them is. In this way, we will discover similar articles in Wikipedia with very high degree of comparability (pseudo-parallel texts).

Distributional similarity: Words with similar distribution in comparable corpora are likely to be translation equivalents. As in the previous strategy (transitivity), we will use distributional similarity for validation, namely to validate the correct cognates.

Spelling similarity: Two words with the same or almost the same spelling are good candidates to be bilingual cognates. We will use the Edit distance to identify similar words in terms of spelling. To minimize the low coverage of the lexicons acquired using this method, it is convenient to conduct the experiments on a family of related languages. Indeed, only languages belonging to the same linguistic family share many cognates.

Only this third aspect (spelling similarity) is exclusive to cognate extraction. Distributional similarity is the validation method we used in the two proposed strategies (transitivity and cognates). Corpus comparability is an aspect that affects the two strategies, however, we only measure comparability in cognate extraction because most cognates are technical terms appearing in domain specific texts. So, finding and selecting bilingual texts in the same technical domain is a crucial issue. The degree of comparability is very important in cognate extraction, but this is not enough to obtain non-noisy bilingual lexicons of cognates. The combination of the

aforementioned three types of similarity, including distributional similarity, is required to generate high-quality cognate lexicons.

3.2 *The method*

Our cognate-based method follows the idea that the use of distributional similarity to extract bilingual cognates from very comparable corpora should generate correct translations. Considering this idea, we designed a strategy adapted to the Wikipedia structure. Among the different web sources of comparable corpora, Wikipedia is likely to be the largest repository of similar texts in many languages. We only require the appropriate computational tools to make them comparable. The proposed method is based on the Wikipedia structure, even though it can be easily generalized to be adapted to other sources of comparable corpora.

The output is bilingual terminology containing many domain-specific terms found on Wikipedia. The method consists of four steps:

Corpus alignment: First, we identify the Wikipedia articles in two languages whose titles are translations of each other.

Degree of comparability: Then, to calculate a degree of comparability between two aligned articles, we apply a similarity measure and select the most comparable pairs of bilingual articles (Gamallo / González 2011b).

Candidates for translation equivalents: From each very comparable pair of articles, we calculate the distributional similarity and select the most similar word pairs, which are considered to be candidates for translation equivalents (Gamallo / Pichel 2008, Gamallo 2007). We also take

into account multiwords. Two bilingual dictionaries are required to generate lexico-syntactic seeds, which are used as text anchors in relation to both languages.

Selecting cognates: Finally, using the Edit Distance (Levenshtein version), we check whether the candidates are *cognates* and select the most similar ones as true translation equivalents.

4 Experiments

To verify whether our methods are useful, we used them to generate bilingual dictionaries in two different tasks. First, as was described in Section 2, two new bilingual dictionaries were built by transitivity: (*English, Galician*) and (*English, Portuguese*) dictionaries. Secondly, we produced bilingual terminology (*Spanish, Portuguese*) with bilingual cognates by making use of the strategy defined in Section 3. These resources were evaluated. All the dictionaries and terminologies we produced are freely available[2].

4.1 *Derivation by transitivity*

In this task, we built two new free dictionaries: (*English, Galician*) and (*English, Portuguese*). The existing dictionaries used as sources for

[2] http://fegalaz.usc.es/~gamallo/dicos_comparable.tgz

deriving the new ones by transitivity are the following (only nouns, adjectives, and verbs are considered):

English-Spanish: For this pair, we used two resources: the free dictionary from Apertium v0.8[3], which contains 10,828 bilingual entries, and the Collins[4] dictionary, which contains 48,637 entries.

Spanish-Portuguese: In this case, we just used the free resource of Apertium v1.1[5], which contains 10,281 entries.

Spanish-Galician: We also used Apertium v1.0[6], which contains, for this pair of languages, 27,640 entries.

Dictionaries	Total number	Ambiguous entries	Unambiguous entries
(*English, Galician*)	25, 790	18, 623	7, 167
(*English, Portuguese*)	12, 306	7, 179	5, 127

Table 1: Noisy dictionaries derived by transitivity

The Apertium dictionaries contain few multi-words, just some idioms. For this reason, we did not carry out multi-word extraction within this particular task. Using the strategy described above in Section 2, we generated the noisy bilingual dictionaries showed in Table 1 by transitivity.

Note that the third column of the table shows the number of ambiguous entries, which are actually the noisy entries. The words making up each

[3] http://sourceforge.net/projects/apertium/files/apertium-en-es

[4] http://www.collinslanguage.com/

[5] http://sourceforge.net/projects/apertium/files/apertium-es-pt

[6] http://sourceforge.net/projects/apertium/files/apertium-es-gl

entry pair are conceptually related even if many of them are not correct bilingual correspondences. The next step is to validate the noisy part of the dictionary by making use of translation equivalents extracted from comparable corpora, i.e. by making use of distributional similarity.

4.1.1 *Comparable corpora*

To validate the English-Galician and English-Portuguese correspondences with ambiguous words, we used the distributional-based strategy described in Gamallo (2007). Text corpora were syntactically analyzed using a multilingual dependency based parser, DepPattern (Gamallo / González 2011a). The comparable corpora were basically produced with news crawled from different online journals: 70*Mb* of English news from The New York Times[7], Reuters Agency[8], and The Guardian[9]; 70*Mb* of Galician news from Vieiros[10] and Galicia-Hoxe[11]; 21*Mb* of Portuguese news from Jornal de Notícias[12] and Público[13] . These monolingual corpora were used to build both (*English, Galician*) and (*English, Portuguese*) comparable corpora. Automatic extraction of translation equivalents were carried out on those comparable corpora by making use of the unambiguous entries generated by transitivity. These

[7] http://www.nytimes.com/

[8] http://trec.nist.gov/data/reuters/reuters.html

[9] http://www.theguardian.com/

[10] http://www.vieiros.com/

[11] http://www.galiciahoxe.com/

[12] http://www.galiciahoxe.com/

[13] http://www.publico.pt/

entries were used to generate lexico-syntactic seeds. Two sets of translation candidates were built: 700,000 candidates from (*English, Galician*) and 500,000 candidates from (*English, Portuguese*). Corpus-based lexicons are much bigger than those directly derived by transitivity because each word is associated to its *N* most appropriate translation candidates (where *N* = 10 in our experiments) by using distributional similarity. So, they contain much more noisy correspondences than those generated by transitivity. Ideally, they should contain, at least, a good bilingual correspondence for each word. This good correspondence will be used to validate dubious pairs derived by transitivity.

4.1.2 *Validation*

To check the validity of the dubious correspondences within the ambiguity-based lexicons (i.e. containing ambiguous entries), we make their intersection with the corpus-based lexicons. This way, we filter out odd bilingual correspondences so as to just select the correct ones, which share both conceptual and distributional similarity.

The second column in table 2 shows the validation number resulting from unifying both the distributional dictionaries and the ambiguous entries obtained by inter-section. In the (*English, Galician*) dictionary, we validated 4,248 correct entries which represent almost 23% of entries found in the ambiguity-based dictionaries (18,623 entries). In the (*English, Portuguese*) dictionary we validated 2,411 out of 7,179 ambiguous entries, which is 33.5% of all ambiguous entries generated by transitivity. These results are very similar to those obtained by Nerima / Wehrli (2008) using parallel corpora. These authors reported an experiment to derive an

English-German dictionary by transitivity, where the ambiguity-based correspondences were validated using parallel corpora. The result of this checking process allowed them to validate 6,282 correspondences, which represent 26% of all candidate correspondences with ambiguous words. Even though we use non-parallel or comparable corpora, our results cannot be considered as being worse, which is very promising.

The third column in Table 2 shows the number of unambiguous entries, i.e., those with one-to-one bilingual correspondences. Note that unambiguous entries are not required to be validated because they must all be correct considering that the handcrafted dictionaries taken as lexical source are also correct.

Dictionaries	**Ambiguous (validated)**	**Unambiguous entries**	**Total entries**
(*English, Galician*)	4, 248	7, 167	**11, 415**
(*English, Portuguese*)	2, 411	5, 127	**7, 538**

Table 2: Final non-noisy dictionaries

At the end of the process, the resulting non-noisy dictionary is the union of the validated correspondences with the lexicons containing unambiguous words. The last column shows the total number of non-noisy correspondences that our method was able to automatically generate. To be precisel, we generated 11,415 entries in the (*English, Galician*) dictionary, which represent 44% of the total correspondences found in the original and noisy dictionary generated by transitivity (25,790). On the other hand, we generated 7,538 entries in the (*English, Portuguese*) dictionary, which

represent 62% of the total correspondences found in the original and noisy dictionary generated by transitivity (12,206).

4.1.3 *Evaluation of the dictionaries generated by transitivity*

To evaluate the correctness of the lexicons, we have selected several samples of 200 word pairs for each dictionary and for each subset of entries to be evaluated. More precisely, we evaluated the performance of both the list of unambiguous words as well as the process of validating ambiguous words with comparable corpora.

As expected, the set of unambiguous words generated by transitivity is 100% correct in terms of precision for both language pairs. No error was found.

As far as the validation process is concerned, Table 3 shows the results of our evaluation. Precision is the number of correct pairs validated by our system divided by all validated pairs. We found just two errors (99.0% precision) in the (*English, Galician*) sample and one error (99.5% precision) in the (*English, Portuguese*) one. Recall is the number of correct pairs validated by our system divided by all correct pairs found in the set of ambiguous entries. We estimated the number of correct ambiguous entries by using a sample of 200 ambiguous pairs for each language pair before validation. As we found that the percentage of correct ambiguous pairs is respectively 80% and 79% in the (*English, Galician*) and (*English, Portuguese*) dictionaries, the total number of correct ambiguous pairs likely to be extracted in (*English, Galician*) is 14,898, and 5,671 in (*English, Portuguese*). As Table 3 shows, the recall is still far from reasonable, in particular when the source dictionaries

contain many ambiguous pairs that are not very frequent words, as in the case of (*Spanish, Galician*). This is in accordance with the results described in related work (Saralegi et al. 2011, Saralegi et al. 2012), where the authors provide good precision but recall is seriously damaged. However, the correctness of the derived lexicons is similar to the dictionaries built by hand by lexicographers, since they are close to 100% correct. Moreover, in spite of the low recall, the size of the (*English, Galician*) dictionary is larger that the smaller source dictionary: namely, the *English-Spanish* lexicon integrated in the machine translation system Apertium (Armentano-Oller et al. 2006). It follows that our automatically generated dictionaries are both good and large enough to be inserted in rule-based machine translation systems.

Dictionaries	Precision	Recall	F-score
(*English, Galician*)	99.0%	28.2%	43.8
(*English, Portuguese*)	99.5%	42.3%	59.3%

Table 3: Evaluation of the validated bilingual correspondences

In fact, one of the direct applications of the two new generated dictionaries is their integration into an open source machine translation system: Apertium. More precisely, the main objective of our experiments is to update the bilingual lexicons of Apertium in order to improve the results of the machine translation system.

4.2 *Bilingual cognates*

We conducted another experiment aimed at learning a large set of new bilingual cognates from the Portuguese and Spanish versions of

Wikipedia. To minimize the low coverage of the lexicons acquired by the cognate-based method, it is convenient to use it on families of related languages which share many cognates. This is why our experiments were carried out using Portuguese and Spanish, two Latin languages which are closely related. The extraction is focused on nouns, adjectives, and verbs, as well as on multi-words.

4.2.1 Existing resources

Our method requires a list of seed lexico-syntactic patterns, whose constituent lemmas are taken from existing bilingual resources. We used two different existing dictionaries:

- Apertium: The general purpose bilingual dictionary (*Spanish, Portuguese*) available in Apertium, and already used in the previous experiment. It contains 9,854 bilingual entries with nouns, adjectives, and verbs.
- Wikipedia: We created a new (Spanish, Portuguese) dictionary using the interlanguage links of Wikipedia. Since Wikipedia is an encyclopedia dealing with named entities and terms, this new dictionary only contains names and domain-specific terminology. It has up to 253,367 bilingual entries.

4.2.2 *Size of the extracted lexicons*

The total size reached by the union of both resources is 263,362 different bilingual correspondences, which will be used as seed pairs. Note that the two dictionaries are complementary: we only found 263 entries in common.

After applying our method to the whole Portuguese and Spanish Wikipedia, we extracted 27,843 new bilingual correspondences. None of them were in the two input dictionaries.

Table 4 depicts the final results. In the first row, we show the extractions of single words while the second row is focused on multi-words. Single words and multiwords are distributed by PoS categories: nouns, adjectives, and verbs. As far as multi-words are concerned, adjectives are not considered. The total extractions considering both multi-word terms and single words are shown in the third row. Notice that the total size of the new bilingual dictionary at 27,843 entries, is much larger than that of the general purpose dictionary of Apertium, which contains only 9,854 bilingual correspondences.

	nouns	**adject.**	**verbs**	**total**
single words	9,374	5,725	2,215	17,314
multi-word terms	9,585	-	944	10,529
all terms	18,959	5,725	3,159	**27,843**

Table 4: Size of the extracted lexicons

4.2.3 *Evaluation of the cognate-based extraction*

To evaluate the precision of the extracted dictionary, a test set of 450 bilingual pairs was randomly selected, consisting of three balanced subsets: 150 bilingual pairs of nouns, 150 bilingual pairs of verbs, and 150 bilingual pairs of adjectives. These included nominal and verbal multi-words, where

their head is either a noun or a verb. The results are depicted in Table 5. Precision is the number of correct pairs divided by the number of evaluated pairs. Recall is the number of correct pairs divided by the number of all correct candidates extracted before the final validation performed with distributional similarity. So, we consider that the total number of correct candidates is that provided by distributional similarity, just before being validated with the cognate-based strategy. To compute recall, we took into account the number of correct pairs extracted by the distributional similarity that were not selected by the cognate-based similarity. For this purpose, we used a new test set with 200 pairs (separated by PoS categories). We found that only 9% of those pairs which were not validated were correct. It follows that our cognate-based similarity is losing few correct cases, giving rise to high recall.

The best performance was achieved by using adjectives: 95% precision and 94% f-score. By contrast, verb extraction only achieves 89% precision. The performance for adjectives is better than that for verbs and nouns probably because adjectives are not on the list of multi-words.

	Precision	**Recall**	**F-score**
Nouns	91.0%	*88.7%*	*89.5*
Verbs	89%	86.8%	87.9%
Adjectives	95.9%	94.8%	94.5.9%
Total	**92%**	**89.5%**	**90.7%**

Table 5: Evaluation of the extracted bilingual cognates

The precision of the total bilingual lexicon, with 27,843 entries, is 92%. This performance outperforms state-of-the-art work on extraction from comparable corpora, whose best scores were about 70% accuracy in Rapp (1999) and 60-83% in Aker et al. (2013). The correctness of the generated translation equivalents is similar to that achieved using parallel corpora. It follows that our method permits the mimimisation of the effort to build a new bilingual dictionary of two related languages.

It is worth mentioning that the results of the comparability measure have not been directly evaluated. In order to know the error rate underlying the automatic alignment of similar texts, a quantitative analysis will be required in future work.

4.2.4 *Error analysis*

We found 39 errors out of 450 evaluated extractions. Most of them (58%) were due to foreign words, namely English words appearing in the input text as part of titles or citations. For instance, the translation pair "about/about" was incorrectly learned from two Portuguese and Spanish texts containing such a word within a non-translated English expression. It would not be difficult to avoid this kind of problem if we use automatic language identification to find parts of the input text written in other languages.

The second most common error type (8%) was caused by prefixes appearing in one of the two correlated words, for instance:

americanismo / anti-americanismo (*anti-americanism*)
anti-fascista / fascista (*anti-fascist*)
hispanoárabe / neo-hispano-árabe (*neo-hispano-arabic*)

Note that it would be possible to filter out those cases by making use of a list of productive prefixes.

In Table 6, we show some types of errors found in the evaluation. As the two most common errors (foreign words and prefixes), which represent 66% of the total number of errors, can be easily filtered out, the total achievable accuracy of our system could be 97%.

Error types	**Frequency (%)**
foreign words	58%
prefixes	8%
typos	8%
multi-words	5%
PoS-tagging	3%

Table 6: Types of errors ranked by frquency

5 Conclusions and future work

In this article, we described two different methods to build high quality bilingual lexicons using comparable corpora. In order to overcome the poor results and low precision inherent to most extraction approaches based on comparable corpora, we made use of two restrictions: transitivity and cognates. The performance of our approach, in terms of precision, is close to the precision achieved by the extraction methods based on parallel corpora.

We made use of dictionaries already integrated into rule-based machine translation systems such as Apertium. It follows that an application of our method will be helpful for the production of new language pairs treated by a machine translation system, namely those pairs included in minority languages. Further evaluations of the results obtained with machine translation systems could be considered as an indirect evaluation of the correctness of the dictionaries produced by our extraction strategies. The number of bilingual dictionaries required by a multilingual translator increases as a quadratic function of the number of languages the system aims to translate (Wehrli et al. 2009). So, the process of automatically deriving new bilingual resources can drastically reduce the amount of work required for this task. Moreover, as our extraction methods only require comparable corpora, it will not be difficult to generate new bilingual dictionaries and terminologies for those languages with less resources or with fewer parallel texts available.

6 References

Aker, Ahmet, Monica Paramita, and Robert Gaizauskas. 2013. "Extracting bilingual terminologies from comparable corpora". In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 4–9, Sofia, Bulgaria.

Ansari, Ebrahim, M. H. Sadreddini, Alireza Tabebordbar, and Mehdi Sheikhalishahi. 2014. "Combining different seed dictionaries to extract lexicon from comparable corpus". *Indian Journal of Science and Technology*, 7(9):1279–1288.

Armentano-Oller, Carme, Rafael C. Carrasco, Antonio M. Corbí-Bellot, Mikel L. Forcada, Mireia Ginestí-Rosell, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Gema Ramírez-Sánchez, Felipe Sánchez-Martínez, and Miriam A. Scalco. 2006. "Open-source Portuguese-Spanish machine translation". *Lecture Notes in Computer Science, 3960*, pages 50–59.

Gamallo, Pablo. 2007. Learning Bilingual Lexicons from Comparable English and Spanish Corpora. In *Machine Translation SUMMIT XI*, Copenhagen, Denmark.

Gamallo, Pablo and Marcos Garcia. 2012. "Extraction of bilingual cognates from Wikipedia". *Lecture Notes in Computer Science* 7243, pages 63–72.

Gamallo, Pablo and Isaac González. 2011a. "A grammatical formalism based on patterns of part-of-speech tags". *International Journal of Corpus Linguistics*, 16(1), pages 45–71.

Gamallo, Pablo and Isaac González. 2011b. "Measuring comparability of multilingual corpora extracted from wikipedia". In *Workshop on Iberian Cross-Language NLP tasks (ICL-2011)*, Huelva, Spain.

Gamallo, Pablo and José Ramom Pichel. 2008. "Learning Spanish-Galician Translation Equivalents Using a Comparable Corpus and a Bilingual Dictionary". *Lecture Notes in Computer Science*, 4919, pages 413–423.

Gamallo, Pablo and José Ramom Pichel. 2010. "Automatic generation of bilingual dictionaries using intermediary languages

and comparable corpora". In *CICLING, LNCS, Vol. 6008*, pages 473–483, Iasi, Romania. Springer-Verlag.

Hazem, Amir and Emmanuel Morin. 2014. "Improving bilingual lexicon extraction from comparable corpora using window-based and syntax-based models". *Lecture Notes in Computer Science*, 8404, pages 310–323.

Nakagawa, Hiroshi. 2001. "Disambiguation of single noun translations extracted from bilingual comparable corpora". *Terminology*, 7(1), pages 63–83.

Nerima, Luka and Eric Wehrli. 2008. "Generating bilingual dictionaries by transitivity". In *LREC'08*, pages 2584–2587, Marrakesh, Marocco.

Rapp, Reinhard. 1999. "Automatic Identification of Word Translations from Unrelated English and German Corpora". In *Proceedings of ACL'99*, pages 519–526.

Saralegi, X., I. Manterola, and I. San Vicente. 2011. "Analyzing methods for improving precision of pivot-based bilingual dictionaries". In *Empirical Methods in Natural Language Processing (EMNLP-2011)*, pages 846–856, Edinburgh, Scotland, UK.

Saralegi, X., I. Manterola, and I. San Vicente. 2012. "Building a basque-chinese dictionary by using english as pivot". In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC-2012)*, Istanbul, Turkey.

Tamura, Akihiro, Taro Watanabe, and Eiichiro Sumita. 2012. "Bilingual lexicon extraction from comparable corpora using label

propagation". In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 24–36, Jeju Island, Korea.

Wehrli, Eric, Luka Nerima, and Yves Scherrer. 2009. "Deep linguistic multilingual translation and bilingual dictionaries". In *4th Workshop on Statistical Machine Translation*, pages 90–94, Athens, Greece.