# Learning Bilingual Lexicons from Comparable English and Spanish Corpora

**Pablo Gamallo Otero**

Department of Spanish
University of Santiago de Compostela
Spain
pablogam@usc.es

**Abstract**

Research on extraction of word translation from comparable, non-parallel texts has not been very popular because it produces poor results when compared to those obtained from aligned parallel corpora. Whereas for parallel texts, word translation extraction can reach about *99%*, the accuracy for comparable corpora has been around *72%* up to now. The current approach, which relies not on a bilingual dictionary but on the previous extraction of bilingual information from parallel corpora, makes a significant improvement to about *79%* of words translations identified correctly.

## 1. Introduction

Many approaches have tried to automatically acquire translation equivalents from bilingual corpora. These approaches can be organised in a continuum according to the type of bilingual input required to acquisition. The input bitext is ranged from well aligned parallel corpora (Melamed, 1997a; Tiedemann, 1999; Vintar, 2001; Guinovar, 2004; Gamallo, 2005) to unrelated non-parallel corpora (Rapp, 1999), going through intermediate levels such as noisy parallel (Melamed, 1997b; Fung, 1995b), pseudo-parallel (Utsuro, 2002), and related non-parallel corpora (hereafter ``comparable corpora") (Fung & McKeown, 1997; Fung & Yee, 1998; Gamallo & Pichel, 2005).

The approaches relying on (well aligned, noisy, or pseudo) parallel corpora are endowed with two positive properties: on the one hand, their aligned segments can be easily used as bilingual anchors to extract translation correspondences, and on the other, external bilingual dictionaries are not required. They can be considered as knowledge-poor approaches. By contrast, their main shortcoming is the fact that parallel corpora are not easily available.

As regards approaches relying on (comparable or unrelated) non-parallel corpora, they are more abundant, less expensive, and easily available via web. However, translation equivalents are not easily found because of the use of fuzzy and vague bilingual anchors. Moreover, in many cases, external bilingual dictionaries are required to remedy the lack of aligned segments and meaningful bilingual anchors within the texts. They are knowledge-rich approaches.

In this paper, we describe an unsupervised acquisition method (Section 3) provided with the positive features of related work. More precisely, our strategy aims at extracting translation equivalents from comparable corpora without requiring external bilingual resources. To find meaningful bilingual anchors within the corpus, we use some bilingual correspondences between lexico-syntactic templates previously extracted from small parallel texts.

So, our approach inherits three positive features: it makes uses of meaningful bilingual anchors, it is a knowledge-poor strategy, and it relies on easily available non-parallel corpora to train the extractor.

In Section 4.1, we perform several experiments using different lists of bilingual lexico-syntactic templates. In one of these experiments, we compare the use of bilingual templates extracted from parallel corpora with the use of those extracted from general-purpose bilingual dictionaries. The results will show that a domain-specific parallel corpus, even if very small, provides with more accurate information than general-purpose dictionaries to extract translation equivalents from comparable corpora. Unlike related work on extraction from comparable corpora, the evaluation protocol defined in this section introduces *recall* scores. This is one of the main contributions of this paper, since it will allow comparing our results with future work on non-parallel corpora. This simple comparison was not possible so far because, in related work, evaluation was performed measuring only the translation accuracy of the most frequent words in the training corpus. Finally, in Section 4.2 we will describe an experiment with an improved version of the extraction method.

## 2. Related Work

There are few approaches to extract bilingual lexicons from comparable corpora in comparison to those using a strategy based on aligned, parallel texts. The most popular method to extract word translations from comparable, non-parallel corpora is described and used in (Fung & McKeown. 1997; Fung & Yee, 1998; Rapp, 1999; Chiao & Zweigenbaum, 2002). 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. There are three drawbacks to this method. First, it is not a knowledge-poor approach since it needs external lexical resources such as a bilingual dictionary. Second, not all the words of the dictionary seem to be

reliable seed expressions. Polysemous words should be removed from the list of seed words since they may introduce semantic noise. And third, according to the Harris's hypothesis (Harris, 1985}, 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 (Grefenstette, 1994; Lin, 1998), 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 overcome these drawbacks, we used as seed expressions, not word entries from an external bilingual dictionary, but pairs of bilingual lexico-syntactic templates previously extracted from small samples of parallel corpus, (Gamallo & Pichel, 2005}.

As the lexico-syntactic templates represent unambiguous local contexts of words, they are discriminative and confident seed expressions to extract word translations from comparable texts. In (Tanaka, 2002), syntactic templates are also used for extraction, but they were specified with semantic attributes introduced by hand.

There exist other approaches to bilingual lexicon extraction which do not use external bilingual dictionaries (Fung, 1995a; Rapp, 1995; Diab & Finch, 2001). Yet, (Fung, 1995a) failed to reach an acceptable accuracy rate for actual use, (Rapp, 1995) had strong computational limitations, and (Diab & Finch, 2001) was applied only to non-parallel texts in the same language. On the other hand, (Dejean et al., 2002) describes 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 (Nakagawa, 2001), 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.

In this paper, we aim at building bilingual lexicons from comparable corpora, with the help of some lexico-syntactic information extracted from small parallel texts. No external lexical resource will be used. In order to extend the work described in (Gamallo & Pichel, 2005), more accurate experiments will be performed, a more appropriate evaluation protocol will be defined, and an improved extraction strategy will be evaluated.

## 3. The Approach

Our method consists of three steps: (1) text processing, (2) extraction of bilingual lexico-syntactic templates from parallel corpora, and (3) extraction of word translations from comparable texts using bilingual templates.

### 3.1. Text Processing

Both parallel and comparable corpora are POS tagged using Freeling (Padró, 2004). Then, we use basic pattern matching techniques to identify potential binary dependencies. From each binary dependency, two complementary lexico-syntactic templates are selected. Table 1 shows some representative examples. A lexico-syntactic template defines a set of semantically related words. Given a binary dependency:

*of* (import, sugar)

two templates are selected: <import of [NOUN]>, which represents the set of nouns that can appear after "import of", for instance, "sugar", "goods", "oil", etc. On the other hand, <[NOUN] of sugar> represents the set of nouns appearing before "of sugar": "import", "export", "sell", etc. We follow the notion of *co-requirement* introduced in (Gamallo et al.).

| **Binary Dependencies** | **Templates** |
|---|---|
| *of* (import, sugar) | <import of [NOUN]><br><[NOUN] of sugar> |
| *robj* (approve, law) | <approve [NOUN]><br><[VERB] law> |
| *lobj* (approve, president) | <president [VERB]><br><[NOUN] approve> |
| *modA* (legal, document) | <legal [NOUN]><br><[NOUN] document> |
| *modN* (area, protection) | <protection [NOUN]><br><[NOUN] area> |

Table 1: Some binary dependencies and their corresponding templates.

Note that *lobj* represents the relationship between a verb and the noun immediately appearing at its left; *robj* is the relationship between a verb and the noun appearing at its right. On the other hand, *modA* is the relationship between a noun and its adjective modifier and *modN* is the relation between two nouns: the head and its modifier.

### 3.2. Extracting Bilingual Templates from Parallel Corpora

Once the lexico-syntactic templates have been identified, we aim at extracting bilingual correspondences between templates from aligned, small parallel corpora. For this purpose, we use a very simple learning method. Similarity between pairs of bilingual templates is computed by taking into account their co-occurrence in each aligned segment. We use Dice coefficient as similarity measure. Finally, each template of the source language is linked to the most similar template of the target language provided that the Dice coefficient is higher than an empirically set threshold. Table 2 depicts some bilingual correlations extracted from an English-Spanish parallel corpus.

For a threshold of *0.4*, the accuracy of the extracted bilingual pairs is about *90%*. Even if there is *10%* noise, these pairs of templates will be taken as bilingual anchor points in the following step.

| *ENGLISH* | *SPANISH* | SIM |
|---|---|---|
| *<agricultural [NOUN]>* | *<[NOUN] agrícola>* | *0.66* |
| *<inport of [NOUN]>* | *<importación de [NOUN]>* | *0.82* |
| *<[NOUN] air>* | *<aire de [NOUN]>* | *0.74* |
| *<fight against [NOUN]>* | *<lucha contra [NOUN]>* | *0.64* |
| *<[NOUN] against poverty>* | *<[NOUN] contra pobreza>* | *0.81* |
| *<[NOUN] alloy>* | *<aliado de [NOUN]>* | *0.43* |

Table 2: Some bilingual correlations between templates.

### 3.3. Extracting a Bilingual Lexicon from Comparable corpora

We start by filtering out those bilingual pairs of templates that have one of these two properties: being sparse or being unbalanced in the comparable corpus. A bilingual pair of templates 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 templates is very frequent while the other one is very rare. We use empirically set thresholds to separate sparse and unbalanced bilingual templates from the rest (which hereafter we will call *seed templates*).

Once the set of seed templates has been selected, the next step is to extract candidate translations of lemmas. We follow two different strategies: the one relies on a standard method to rank candidate translations, and the other one is based on a context-based algorithm. In both cases, *hapax legomena* (i.e., words occurring only once) were removed from the comparable corpus.

#### 3.3.1. Standard Strategy

The basic idea underlying this approach is as follows: a lemma $l_2$ in the target language can be the translation of a lemma $l_1$ in the source language, if $l_1$ tends to occur in the same seed templates $l_2$ occurs in. Each lemma is represented as a vector of seed templates. So, to compute similarity between a lemma and a possible translation, we compare the seed templates they share and do not share.

| *president* | *presidente* |
|---|---|
| *<albanian [NOUN]>* | *<[NOUN] albanés>* |
| *<[NOUN] of republic>* | *<[NOUN] de república>* |
| *<former [NOUN]>* | *<anterior [NOUN]>* |
| *<congratulate [NOUN]>* | *<felicitar a [NOUN]>* |

Table 3: Some bilingual pairs of templates shared by "president" and "presidente".

Table 3 shows some of the seed templates shared by the English noun "president" and its Spanish counterpart, "presidente". As a result, each lemma of the source language is associated a list of candidate translations. The list is ranked by degree of similarity.

#### 3.3.2. Context-Based Strategy

Following the algorithm described in (Gamallo, 2005), we do not compare a lemma of the source language to all the lemmas of the target language, but only to those that appear in the same seed templates. So, each template represents the context or domain within which we look for candidate translations. The algorithm is as follows:

Given a lemma of the source language, $l_1$, and a seed template $t_1$, we compare $l_1$ to all the lemmas of the target language occurring with $t_i$. The most similar one is selected and put in a list of candidate translations of $l_1$. Then, the same process is repeated with the rest of seed templates of the set. Finally, the list of candidate translations is ranked by taking into account the number of templates in which each candidate has been taken as the most similar to $l_1$. The same is done for the other lemmas of the source language.

The motivation of this algorithm is to capture word translations in sense-sensitive contexts. Let's see an example. The noun "fuel" can be translated in Spanish either as "carburante" or "combustible", each appearing in different contexts: "ayuda a carburantes" (*aide on fuel*), "combustible nuclear" (*nuclear fuel*). These translations represent two different uses of ``fuel": "carburante" has a very restrictive use since it means fuel for vehicles, while "combustible" means fuel for the rest of engines. In the ranked list of candidate translations of "fuel", obtained using the standard strategy, the best translation was "combustible", following by a more generic term, "energía" (*energy*), and then "carburante". However, using the context-based method, the final ranked list was modified. As "energía" often appears in the same contexts as "combustible", we found few contexts were "energy" appears as the best translation of "fuel". On the other hand, as "carburante" does not appear in the same contexts as "combustible", it was selected in many contexts as the most similar to "fuel". At the end, "carburante" appears as the second candidate of "fuel", and "energía" as the third one, which is more correct.

## 4. Similarity

The standard and context-based methods relies on the same notion of similarity. Similarity between lemmas $l_1$ and $l_2$ is computed using the following non-weighted *Dice* coefficient:

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

where $f(l_1, t_i)$ represents the number of times lemma $l_1$ occurs in a seed template $t_i$. This same measure was also used to extract similar templates from parallel corpora (subsection 3.2). There, we measured similarity between

templates and not between lemmas, whereas the attributes used to measure their similarity were aligned segments and not seed templates.

We compared (1) with a particular weighted version of the same coefficient in which each template was assigned two different weights, as in (Grefenstette, 1994). This weighted version of the measure did not improve the results in a significant way, since unbalanced and sparse templates were filtered out before computing similarity. Moreover, unlike any weighted measure, coefficient (1) does not require a very high-cost algorithm to be implemented. This property is quite useful when we intend to do many experiments with large corpora. So, all the experiments described and evaluated in the next section were performed using equation (1).

## 5. Experiments and Evaluation

Translation equivalents were learned from an English-Spanish comparable, non-parallel corpus selected from the European Parliament proceedings parallel corpus (EuroParl)[1]. The English part consists of 14 million word occurrences while the size of the Spanish part is about 17 million words. The two parts were selected in such a way that no English subparts were translations of Spanish subparts. In the following, we will describe 2 different experiments.

### 5.1. Experment 1: Seed Templates

The aim of the first kind of experiments is to compare the behaviour of different sets of seed templates in our standard extraction method. For this purpose, we made use of 4 small English-Spanish samples of parallel corpora: EuroParl (200,000 word occurrences), the European Constitution (150,000), some localisation files of the Linux system[2] (400,000), and finally some literary texts, namely Don Juan Tenorio and El Quijote, which taken in conjunction come to about 270,000 word occurrences.

In addition, we also generated a set of templates from an external dictionary, IDP[3], with almost 4,000 lemmas. To generate bilingual pairs of templates from the lemmas of the dictionary, we use very basic syntactic rules: e.g., a verb can appear with nouns both to the left and to the right, with groups of prepositions and nouns to the right, etc. Given, for instance, the verb "refuse" and its Spanish translation, "rechazar", we generated potential bilingual pairs of templates such as:

| | |
|---|---|
| *<refuse [NOUN]>* | *<rechazar [NOUN]>* |
| *<[NOUN] refuse>* | *<[NOUN] rechazar>* |
| *<refuse to [NOUN]>* | *<rechazar a [NOUN]>* |
| *<refuse of/from[NOUN]>* | *<rechazar de [NOUN]>* |

The filtering process removes those unbalanced or sparse bilingual pairs overgenerated by the syntactic rules.

To make evaluation easier and faster, we performed the experiments using only nominal seed templates, i.e., lexico-syntactic templates co-occurring with nouns.

Figure 1 shows the 8 sets of seed templates we used in our experiments; each set name is provided with the number of seed templates it contains. Sparse and unbalanced templates were previously filtered out. At the top of the figure, there are three independent sets: *EuroParl, Linux*, and *IDP*. The two first sets contain templates extracted from two parallel corpora, whereas the former consists of those generated from the external bilingual dictionary IDP. Then, there are two sets directly relying on *EuroParl*: *EuroParl-HALF* which is a random selection of almost *50%* of the templates in *EuroParl*, and *EuroParl-Constit*, putting together *EuroParl* with the bilingual templates extracted from the European Constitution. Then, we built other three sets: on the one hand, *EuroParl-Constit-Linux-Lit*, which is the result of putting together the templates extracted from the 4 parallel corpora above mentioned, and on the other hand, two special sets, *Bootstrap-A* and *Bootstrap-B*, generated by means of two different bootstrapping strategies.

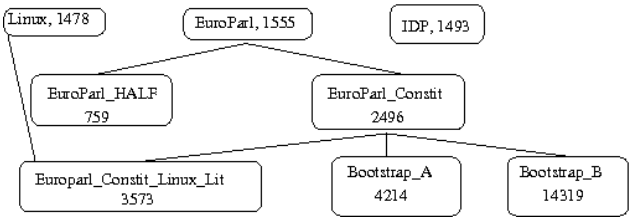


Figure 1: Sets of seed templates used in 8 different experiments.

*Bootstrap-A* put together *EuroParl-Constit* with a new set of bilingual pairs of templates learned as follows. First, we learn a bilingual lexicon of lemmas using *EuroParl-Constit* as seed templates. Then, the lemmas of the lexicon are taken as "seed words" to extract further bilingual templates. This second extraction is just the reverse of the first one: to compute similarity, the compared objects are not lemmas but templates while the attributes are bilingual lemmas instead of templates. Each template of the source language is associated its most similar one if the Dice coefficient is higher than a particular threshold. Finally, *Bootstrap-B* put together *EuroParl-Constit* with another set of bilingual pairs of templates generated in a different way. The bilingual lexicon we have learned using E*uroParl-Constit* is taken as an external dictionary. So, for each bilingual pair of lemmas and a list of very basic rules, we generate the corresponding templates as we did with IDT. *Bootstrap-B* generates more new seed templates as *Bootstrap-A*, but the rate of error is also higher.

No manual correction was made to clean the sets of seed templates. They were used to extract different bilingual lexicons, where each English noun entry was associated a ranked list of translation candidates. This list consists of the top 10 most similar Spanish nouns. We follow the standard strategy described above (subsection 3.3.1).

[1]Available on-line at http://www.statmt.org/europarl/

[2]http://www.iro.umontreal.ca/translation/

[3]http://www.june29.com/IDP/

### 5.1.1. Evaluation

To evaluate the efficiency of the different sets of seed templates in the process of extracting bilingual lexicons, we elaborate an evaluation protocol with the following characteristics. A random sample of 100 test nouns was selected from the English corpus.

Two different precisions were considered: *precision-1* is defined as the number of times a correct translation candidate of the test noun is ranked first, divided by the number of test nouns. *Precision-10* is the the number of correct candidates appearing in the top 10, divided by the number of test nouns.

A candidate translation is considered to be correct if it appears in one of the definitions proposed by the WordReference English-Spanish bilingual dictionary[4]. For acronyms or technical terms not attested in the dictionary, we found the correct translation in the parallel texts. On the other hand, indirect associations are judged to be incorrect. For instance, if "member" is translated by "estado" because of the frequent collocation "member state", the evaluator considers that there is an incorrect association.

As regards the completeness of the extracted lexicons, two different situations are considered. We call *recall* the number of entries in the bilingual lexicon divided by the number of different lemmas (here nouns) occurring more than once in the English part of the comparable corpus. As *hapax legomena* were not taken into account by the algorithm, they are removed from the evaluation protocol. On the other hand, *recall** is defined as the number of times the lexicon entries occur in the English comparable corpus, divided by the total sum of noun tokens in the corpus.

As far as we know, no definition of recall nor coverage has ever been proposed in related work. In most evaluation protocols of previous work, authors only give information on the frequency of the evaluated words. They are sometimes the *N* most frequent expressions in the training corpus (Fung&Yee, 1998), while in other experiments, they are the word types or lemmas with a frequency higher than *N* (where *N* is often >=*100*) (Gamallo, 2005b; Chiao, 2002). In fact, as absolute frequencies are dependent on the corpus size, they are not very useful when we try to compare the precision or accuracy among different approaches. By defining two types of recall, which are independent of the corpus size, we try to overcome such a limitation.

### 5.1.2. Results

Table 4 depicts the precision scores obtained using the 8 different sets of seed templates. In all cases, *recall** was situated at *90%*. It means that the random list of 100 test nouns belongs to the larger list of those nouns that together come to the *90%* of noun tokens in the training corpus. For our English corpus, *recall** of *90%* corresponds to a bilingual lexicon constituted by 1,641 nominal lemmas, each lemma having a token frequency >=*103*. As the English corpus contains 15,881 noun lemmas, *recall* is about *10%*.

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

| | ***Prec-1*** | ***Prec-10*** | ***seed templates*** |
|---|---|---|---|
| **IDP** | .15 | .20 | 1,493 |
| **Linux** | .15 | .18 | 1,478 |
| **Europarl** | .68 | .84 | 1,555 |
| **Europarl-HALF** | .59 | .72 | 759 |
| **Europarl-Constit** | .72 | .84 | 2,496 |
| **Europarl-Constit-Linux-Lit** | .73 | .85 | 3,573 |
| **Bootstrap-A** | .75 | .87 | 4,212 |
| **Bootstrap-B** | .73 | .87 | 14,319 |

Table 4: Precision of 8 bilingual lexicons

Table 4 shows that seed templates extracted from a general-purpose external dictionary (*IDP*) are much less useful as those extracted from a domain-specific parallel text (*EuroParl*), containing documents of the same domain as the comparable corpus used for training. With a similar amount of seed templates (1,555 and 1,493), *EuroParl* reaches *.68/.84* precision, while *IDP* does not get through *.20*. It means that domain-specific templates are much more semantically informative than general-purpose ones for the task described in this paper. Oddly, the results obtained from the general-purpose dictionary are not much better as those obtained using *Linux*, a domain-specific parallel text totally different from the comparable corpus. In fact, the seed templates selected from the computer corpus turned out to be only those that have no domain-specific content, as the templates extracted from the external dictionary. Those domain-specific templates associated to the computer science domain were filtered out because they are either sparse or unbalanced in the EuroParl corpus.

Moreover, Table 4 also allows us to observe that the largest set of templates (*Bootstrap-B*) does not give us the best results. In fact, such a set is very noisy as it contains many odd templates overgenerated by basic syntactic rules. Overgeneration of templates makes noise much higher than 1*0%*, which is the usual rate of odd templates in the other sets.

Furthermore, we can observe that precision improves, even if slightly, when the set of seed templates is getting large and no more noise is introduced. This is true from *EuroParl* to *Bootstrap-A*. *Precision-1* and *precision-10* goes from *.68/.84* to *.75/.87*. The reverse is also true: precision decreases slightly when the set of seed templates is reduced, as in *EuroParl-HALF*.

These results lead us to infer that lexicon extraction from a comparable corpus requires starting with a homogeneous set with domain-specific seed expressions. Later, to improve precision, further significant seed expressions can be obtained from other parallel corpora, external dictionaries, or/and bootstrapping strategies.

### 5.1. Experment 2: Context-Based Strategy

The second type of experiments is focused on the evaluation of bilingual lexicons extracted using the context-based strategy. Here, we only made use of 1 set of templates, namely *Bootstrap-A*, which gave the best results in the previous experiments. Inspired by the evaluation performed in (Melamed, 1997a), 3 lists of 100 nouns were randomly created. Each list belonged to a particular level of *recall**. *Recall** of *50%*, *80%* and *90%* corresponded to bilingual lexicons containing 158, 766 and 1,641 lemmas, respectively (see Figure 1). As the English corpus contains 15,881 lemmas of nouns, the 3 scores of *recall* are quite low: *0.9%*, *5%* and *10%*, respectively. This apparent drawback is inherent to extraction from comparable texts. Indeed, only frequent words can be correctly translated since they need to co-occur with a significant set of seed expressions. However, given that comparable texts are easily available, this is not a serious drawback.

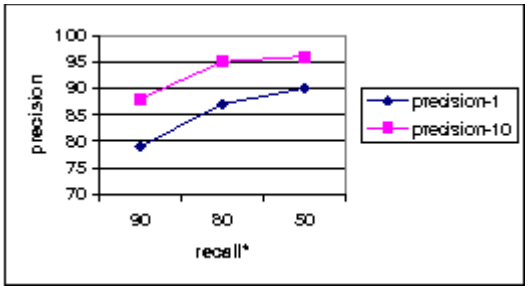


Figure 1: Precision at three levels of *recall**

The upper curve represents *precision-10,* which goes from *.88* to *.96*. The lower one corresponds to *precision-1*, where only the best translation is considered. The scores go from *.79* to *.90*. Note that this context-based strategy improves the results obtained by the standard method. At the *90% recall**, we reached *.88/.79* precision, improving the previous best score: *.87/.77*.

For comparison with other approaches, Rapp (1999) reports a precision of *.72* when only the top candidate is considered (i.e., *precision-1*). This was the best result reported up to now. Note that we achieve *.79* at the *recall** of *90%*. However, such a comparison is not very adequate since Rapp (1999) do not provide any information on recall.

## 6. Conclusion and Discussion

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 templates as seed expressions. This makes a significant improvement to about *79%* of word translations identified correctly if only the best candidate is considered, and almost *90%* if we consider the top 10. These results are not very far from those obtained by approaches based on parallel texts.

However, the main contribution of this paper is to show that there is still a good margin to improve results. Precision scores are getting better when experiences are made using larger sets of seed templates, even if they are taken from unrelated corpora. So, with many small parallel texts, it could be possible to generate big sets of seed templates taken as bilingual anchors to easily pseudo-align many Gigabytes of comparable texts.


### Acknowledgements

This work has been supported by Ministerio de Educación y Ciencia of Spain, within the project GARI-COTERM, ref: HUM2004-05658-D02-02.


### References


Carreras, X. & Padró, L. (2004). An open-source suite of language analyzers. In Proceedings of LREC'04.

Chiao, Y. & Zweigenbaum, P. (2002). Looking for candidate translational equivalents in specialized, comparable corpora. In Proceedings of 19th COLING'02.

Dejean, H., Gaussier, E. &, Sadat, F. (2002). Bilingual terminology extraction: an approach based on multilingual thesaurus applicable to comparable corpora. In Proceedings of COLING'02. Tapei, Taiwan.

Diab, M. & Finch, S. (2001). A Statistical Word-Level Translation Model for Comparable Corpora. In Proceedings of the Conference on Content-Based Multimedia Information Access (RIAO).

Fung, P. (1995a). Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. In Proceedings of 14th Annual Meeting of Very Large Corpora.

Fung, P. (1995b). A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In Proceedings of ACL'95.

Fung, P. & McKeown, K. (1997). An IR Approach for Translating New Words from Non-parallel, Comparable Texts. In Proceedings of 5th Annual Workshop on Very Large Corpora (pp. 192-202). Hong Kong.

Fung, P. & Lo Yuen Yee (1998). Finding terminology translation from non-parallel corpora. In Proceedings of Coling'98 (pp. 414-420). Montreal, Canada.

Gamallo P. (2005). Extraction of Translation Equivalents from Parallel Corpora Using Sense-Sensitive Contexts. In Proceedings of 10th Conference of the European Association for Machine Translation (EAMT'05) (pp. 97-102). Budapest, Hungary.

Gamallo P., Agustini, A. & Lopes, G. (2005). Clustering Syntactic Positions with Similar Semantic Requirements. Computational Linguistics, 31(1), 107-146.

Gamallo P. & Pichel, J.R. (2005). An Approach to Acquire Word Translations from Non-Parallel Texts, Lecture Notes in Computer Science, vol. 3808.Springer-Verlag.

Guinovart, Xavier & Sacau, Elena (2004). Métodos de optimización de la extracción de léxico bilingue a partir de corpus paralelos. Procesamiento del Lenguaje Natural, 33, 133-144.

Grefenstette, Gregory. (1994). Explorations in Automatic Thesaurus Discovery. Kluwer Academic Publishers.

Harris, Z. (1985). Distributional structure. In: J.J. Katz (Eds.). The Philosophy of Linguistics (pp. 26-47). New York: Oxford University Press.

Lin, Dekan. (1998). Automatic retrieval and clustering of similar words. In Proceedings of COLING-ACL'98. Montreal, Canada.

Melamed, Dan (1997a). A Word-to-Word Model of Translational Equivalence. In Proceedings of 35th Conference of the Association of Computational Linguistics (ACL'97). Madrid, Spain.

Melamed, Dan (1997b). A Portable Algorithm for Mapping Bitext Correspondences. In Proceedings of 35th Conference of the Association of Computational Linguistics (ACL'97) (pp. 305--312). Madrid, Spain.

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

Rapp, Reinhard. (1995). Identifying word translations in non-parallel texts. . In Proceedings of ACL'95, (pp. 320-322).

Rapp, Reinhard. (1999). Automatic Identification of Word Translations from Unrelated English and German Corpora. In Proceedings of ACL'99, (pp. 519-526).

Tanala, T. (2002). Measuring the similarity between compound nouns in different languages using non-parallel corpora. In Proceedings of 19th COLING'02, (pp. 981-987).

Tiedemann, Jorg (1998), Extraction of translation equivalents from parallel corpora. In Proceedings of 11th Nordic Conference of Computational Linguistics, Copenhagen, Denmark.

Utsuro, Takehito. (2002). Translation knowledge acqusition from cross-lingually news articles. In Proceedings of 2nd China-Japan Natural Language Processing Joint Research Promotion Conference (pp. 123-134).

Vintar, Ŝ. (2001). Using parallel corpora for translation-oriented term extraction. Babel Journal, 47(2), 121-132.